Detection of Abrupt Changes: Theory and Application[1]

Michèle Basseville
IRISA/CNRS
Rennes, France

Igor V. Nikiforov
Institute of Control Sciences
Moscow, Russia

[1]This book was previously published by Prentice-Hall, Inc.

Contents

Preface

Over the last twenty years, there has been a significant increase in the number of real problems concerned with questions such as

- fault detection and diagnosis (monitoring);
- condition-based maintenance of industrial processes;
- safety of complex systems (aircrafts, boats, rockets, nuclear power plants, chemical technological processes, etc.);
- quality control;
- prediction of natural catastrophic events (earthquakes, tsunami, etc.);
- monitoring in biomedicine.

These problems result from the increasing complexity of most technological processes, the availability of sophisticated sensors in both technological and natural worlds, and the existence of sophisticated information processing systems, which are widely used. Solutions to such problems are of crucial interest for safety, ecological, and economical reasons. And because of the availability of the above-mentioned information processing systems, complex monitoring algorithms can be considered and implemented.

The common feature of the above problems is the fact that the problem of interest is the detection of one or several *abrupt changes* in some characteristic properties of the considered object. The key difficulty is to detect intrinsic changes that are not necessarily directly observed and that are measured together with other types of perturbations. For example, it is of interest to know how and when the modal characteristics of a vibrating structure change, whereas the available measurements (e.g., accelerometers) contain a mix of information related to both the changes in the structure and the perturbations due to the environment.

Many monitoring problems can be stated as the problem of detecting a change in the *parameters* of a static or dynamic stochastic system. The main goal of this book is to describe a unified framework for the design and the performance analysis of the algorithms for solving these change detection problems. We call abrupt change any change in the parameters of the system that occurs either instantaneously or at least very fast with respect to the sampling period of the measurements. Abrupt changes by no means refer to changes with large magnitude; on the contrary, in most applications the main problem is to detect small changes. Moreover, in some applications, the *early warning* of small - and not necessarily fast - changes is of crucial interest in order to avoid the economic or even catastrophic consequences that can result from an accumulation of such small changes. For example, small faults arising in the sensors of a navigation system can result, through the underlying integration, in serious errors in the estimated position of the plane. Another example is the early warning of small deviations from the normal operating conditions of an industrial process. The early detection of slight changes in the state of the process allows to plan in a more adequate manner the periods during which the process should be inspected and possibly repaired, and thus to reduce the exploitation costs.

Our intended readers include engineers and researchers in the following fields :

- signal processing and pattern recognition;
- automatic control and supervision;
- time series analysis;
- applied statistics;
- quality control;
- condition-based maintenance and monitoring of plants.

We first introduce the reader to the basic ideas using a nonformal presentation in the simplest case. Then we have tried to include the key mathematical background necessary for the design and performance evaluation of change detection algorithms. This material is usually spread out in different types of books and journals. The main goal of chapters 3 and 4 is to collect this information in a single place. These two chapters should be considered not as a small textbook, but rather as short notes that can be useful for reading the subsequent developments.

At the end of each chapter, we have added a Notes and References section and a summary of the main results. We apologize for possible missing references.

We would like to acknowledge the readers of earlier versions of the book, for their patient reading and their numerous and useful comments. Thanks are due to Albert Benveniste, who read several successive versions, for numerous criticisms, helpful discussions and suggestions; to Mark Bodson, who reviewed the manuscript; to Fredrik Gustafsson, Eric Moulines, Alexander Novikov, David Siegmund, Shogo Tanaka and Qinghua Zhang, for their numerous comments; and to Alan Willsky for his comments regarding an early version of chapter 7.

Philippe Louarn provided us with extensive and valuable help in using LATEX; his endless patience and kindness in answering our questions undoubtedly helped us in making the manuscript look as it is. Bertrand Decouty helped us in using software systems for drawing pictures.

Our thanks also to Thomas Kailath who accepted the publication of this book in the Information and System Sciences Series which he is editing.

During the research and writing of this book, the authors have been supported by the Centre National de la Recherche Scientifique (CNRS) in France, the Institute of Control Sciences in Moscow, Russia, and the Institut National de la Recherche en Informatique et Automatique (INRIA) in Rennes, France.

The book was typeset by the authors using LATEX. The figures were drawn using MATLAB and XFIG under the UNIX operating system. Part of the simulations were developed using the package AURORA designed at the Institute of Control Sciences in Moscow, Russia.

Michèle Basseville
Igor Nikiforov
Rennes, France

List of Figures

List of Tables

Notation and Symbols

Some symbols may have locally other meanings.

Basic Notation

Notation	Meaning
$\mathbf{1}_{\{x\}}$	Indicator function of event x.
$\mathbb{1}_r$	Vector of size r made of 1.
$\{x : \ldots\}$	Set of x such that.
$\mathbf{R}$	Set of real numbers.
(a, b)	Open interval.
$[a, b]$	Closed interval.
$\max_k, \min_k$	Extrema with respect to an integer value.
$\sup_x, \inf_x$	Extrema with respect to a real value.
$X \sim Y$	Same order of magnitude.
$X \approx Y$	Approximately equal.
$\|v\|_A^2 = v^T A v$	Quadratic form with respect to matrix A.
y	Observation (random variable) of dimension $r = 1$.
Y	Observation (random variable) of dimension $r > 1$.
$\mathcal{Y}_1^k = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{pmatrix}$	or $\mathcal{Y}_1^k = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix}$
$\check{\mathcal{Y}}_1^k = \begin{pmatrix} Y_k \\ Y_{k-1} \\ \vdots \\ Y_1 \end{pmatrix}$	or $\check{\mathcal{Y}}_1^k = \begin{pmatrix} y_k \\ y_{k-1} \\ \vdots \\ y_1 \end{pmatrix}$
k	Current time instant - discrete time.
t	Current time instant - continuous time.
N	Sample size.
γ_i	Weighting coefficients.
$\dot{g}(x)$	First derivative of the function $g(x)$.
$\ddot{g}(x)$	Second derivative of the function $g(x)$.

$\mathrm{tr}\, A$	Trace of matrix A.
$\det A$	Determinant of matrix A.
$\ker A$	Kernel of matrix A.

Notation for Probability

Notation	**Meaning**
Ω	Sample (abstract) space.
$\mathcal{B}$	Event space, sigma algebra of subsets of Ω.
μ	Probability measure.
$\mathbf{P}(B)$	Probability of the event $B \in \mathcal{B}$.
$\mathbf{E}$	Expectation.
Y	Random variable, scalar, or vector.
y	Argument of the distribution functions.
$F(y)$	Cumulative distribution function, cdf.
θ	Vector of parameters with dimension ℓ.
$(\vartheta_i)_{1\leq i\leq \ell}$	Coordinates of the vector parameter θ.
θ	Parameter.
Υ	Change direction.
$\mathcal{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta}$	Parametric family of probability distributions.
$p(y), f(y)$	Probability density, pdf.
$p_\theta(y), f_\theta(y)$	Parameterized probability density, pdf.
$\mathbf{N}(\theta)$	Entropy of the distribution p.
$\mathcal{L}$	Probability law; $\mathcal{L}(Y) = \mathbf{P}_\theta$.
$\mathcal{N}$	Normal (or Gaussian) law.
$\varphi(y)$	Gaussian density, with $\mu = 0, \sigma = 1$.
$\phi(y)$	Corresponding Gaussian cdf.
$\Phi_{\mu,\Sigma}(B)$	$= \int_B f(y)dy$, where $\mathcal{L}(X) = \mathcal{N}(\mu, \Sigma)$.
$\mathcal{K}$	Functional used in stochastic approximation algorithms.
κ	Mean value of $\mathcal{K}$.
ψ	Laplace transform or moment generating function.
ς	Laplace variable.
σ^2	Variance of the scalar random variable Y.
R, Σ	Covariance matrix of the vector random variable Y.
Φ_Y	Power spectrum of a process Y.
$\mathbf{T}_p(Y)$	Toeplitz matrix of order p for a process Y.
$\mathrm{var}(Y)$	Variance of the random variable Y.
$\mathrm{cov}(Y)$	Covariance matrix of the vector random variable Y.
$a_1, ..., a_p$	Scalar AR coefficients.

$A_1, ..., A_p$	Matrix AR coefficients.
$b_1, ..., b_q$	Scalar MA coefficients.
$B_1, ..., B_q$	Matrix MA coefficients.
$T_{-\epsilon,h}$	Exit time from the interval $(-\epsilon, h)$.
$\mathbf{P}_\theta(-\epsilon\|z)$	Abbreviation for $\mathbf{P}_\theta(S_{T_{-\epsilon,h}} \leq -\epsilon \| S_0 = z)$.
$\mathbf{E}_\theta(-\epsilon\|z)$	Abbreviation for $\mathbf{E}_\theta(T_{-\epsilon,h} \| S_0 = z)$.
w.p.1.	Abbreviation for "with probability 1."

Notation for Statistics

Notation	**Meaning**
$\mathbf{H}$	Hypothesis.
α	Error probability.
β	Power.
K_ϵ	Class of tests with level $1 - \epsilon$.
κ_ϵ	Quantile of the normal distribution corresponding to level $1 - \epsilon$.
$\mathbf{I}$	Fisher information.
$\mathbf{K}$	Kullback information.
$\mathbf{J}$	Kullback divergence (symmeterized information).
$b(\theta)$	Bias of an estimate $T(y)$.
$S_i^j(\theta_0, \theta_1)$	Log-likelihood ratio for observations from y_i until y_j.
$S_i^j = S_i^j(\theta_0, \theta_1)$	
$S_j = S_1^j$	
$\tilde{S}_i^j$	Weighted log-likelihood ratio.
s_i	Increment of S_i^j.
Λ_i^j	Likelihood ratio for observations from y_i until y_j.
$\tilde{\Lambda}_i^j$	Weighted likelihood ratio.
$l_\theta(y) = \ln f_\theta(y)$	Log-likelihood function.
$z = \frac{\partial l_\theta(y)}{\partial \theta}$	Efficient score for a random variable with scalar parameter.
$Z = \frac{\partial l_\theta(y)}{\partial \theta}$	Efficient score for a random variable with vector parameter.
$\mathcal{Z}_k^N = \frac{\partial l_\theta(\mathcal{Y}_k^N)}{\partial \theta}$	Efficient score for a sample $Y_k, \ldots, Y_N$ of a random process with scalar or vector parameter.
$\Delta_N = \frac{1}{\sqrt{N}} \mathcal{Z}_N$	
$Z^* = \frac{\partial l_\theta(y)}{\partial \theta}\Big\|_{\theta=\theta^*}$	
g	Test statistics.
T	Stopping time.
$N(z)$	ASN function, as a function of the initial value z of the cumulative sum.
η	Shifted log-likelihood ratio.
$\mathcal{S}$	Cumulative sum of η.

Notation for Change Detection (Except in State-Space Models)

Notation	Meaning
θ_0	Parameter before change.
θ_1	Parameter after change.
ν	Magnitude of change.
Υ	Direction of change in a vector parameter.
ν_m	Minimum magnitude of change.
t_0	Change time.
t_a	Alarm time.
g_k	Decision function.
$-a, \lambda, -\epsilon, h$	Thresholds.
N_k	Number of observations since last vanishing of g_k.
$\bar{T}$	Mean time between false alarms.
$\bar{\tau}$	Mean delay for detection.
$L_z(\theta)$	ARL function.

Notation for State-Space Models

Notation	Meaning
y, Y	Observation, of dimension $r \geq 1$.
X	State, of dimension n.
U	Input (control) of dimension m.
V, W	Noises on observation and state equations, respectively.
R, Q	Covariance matrices for V, W, respectively.
F	State transition matrix.
G, J	Control matrices.
H	Observation matrix.
$\mathcal{H}$	Hankel matrix.
$\mathcal{J}(G, J)$	Toeplitz impulse response.
$\mathcal{O}$	Observability matrix.
$\mathcal{C}$	Controllability matrix.
$\mathcal{T}$	Transfer function.
K	Kalman gain.
P	Covariance matrix of state estimation error.
ε_k	Innovation.
Σ	Covariance matrix of innovation ε.
e_k	Residual.
ζ_k	Parity vector.
$\mathcal{S}$	Parity space.
Υ_x	Direction of the additive change in X.
Υ_y	Direction of the additive change in Y.
Ψ_x	Vector made of successive Υ_x.
Ψ_y	Vector made of successive Υ_y.
Γ	"Gain" matrix for Υ_x.
Ξ	"Gain" matrix for Υ_y.
ρ	Signature of additive change on the innovation.
ϱ	Signature of additive change on the parity vector.
$\mathcal{K}_x$	Transfer function for the contribution of Υ_x to ρ.
$\mathcal{K}_y$	Transfer function for the contribution of Υ_y to ρ.
$\mathcal{H}_x$	Transfer function for the contribution of Υ_x to ϱ.
$\mathcal{H}_y$	Transfer function for the contribution of Υ_y to ϱ.

1

Introduction

In this chapter, we describe the purpose and contents of the book. In section 1.1 we give the theoretical and applied motivations for change detection. The last part of this section consists of three possible statistical problem statements for change detection, together with the intuitive definition of the corresponding criteria to be used for the design and performance analysis of change detection techniques. The formal definition of these criteria is given at the end of chapter 4, after the introduction of the key mathematical tools to be used throughout the book.

In section 1.2, we introduce five typical application examples, which we will use to introduce the main techniques. In section 1.3, we describe the organization of the book, based on a classification of change detection problems according to the types of characteristics that change. We give a short description of each chapter and a general flowchart of the chapters. Finally, in section 1.4, we comment further on several critical issues concerning the design of change detection algorithms and the investigation of their properties.

1.1 Introducing Change Detection

In this section, we introduce abrupt changes for segmentation, fault detection, and monitoring. We describe the main motivations for the investigation of change detection problems. Illustrating examples are described in the next section. Then we classify the topics of change detection methodology into three main classes of problems encountered in signal processing, time series analysis, automatic control, and industrial quality control. Next, we give three statistical problem statements and the intuitive definition of the corresponding criteria. Finally, we describe the purpose of the book.

1.1.1 Motivations

An intensively investigated topic is time series analysis and identification. The main assumptions underlying these investigations are that the properties or parameters describing the data are either constant or slowly time-varying. On the other hand, many practical problems arising in quality control, recognition-oriented signal processing, and fault detection and monitoring in industrial plants, can be modeled with the aid of parametric models in which the parameters are subject to *abrupt changes at unknown time instants*. By abrupt changes, we mean changes in characteristics that occur very fast with respect to the sampling period of the measurements, if not instantaneously. Because a large part of the information contained in the measurements lies in their nonstationarities, and because most of adaptive estimation algorithms basically can follow only slow changes, the detection of abrupt changes is a problem of interest in many applications, as we show in the five examples of section 1.2. The *detection of abrupt changes* refers to tools that help us decide whether such a change occurred in the characteristics of the considered object.

The first meaning of abrupt change thus refers to a time instant at which properties suddenly change, but before and after which properties are constant in some sense, e.g., stationary. This notion serves as a basis to the corresponding formal mathematical problem statement, and to the formal derivation of algorithms for change detection.

It should now be clear that abrupt changes by no means imply changes with large magnitude. Many change detection problems are concerned with the detection of small changes, as we discuss now.

1.1.1.1 Fault Detection in Controlled Dynamic Systems

The problem of fault detection for monitoring an industrial process involves two types of questions. First, of course, the detection of failures or catastrophic events should be achieved. But second, and often of crucial practical interest, the detection of smaller faults - namely of sudden or gradual (incipient) modifications, which affect the process without causing it to stop - is also required to prevent the subsequent occurrence of more catastrophic events. As we explain in examples 1.2.2 and 1.2.5, both faults and failures can be approached in the abrupt change detection framework, with all the aspects of detection, estimation, and diagnosis usually implied in most failure detection and isolation (FDI) systems. Such a detection tool helps to increase the reliability and availability of the industrial process by reducing the number of shutdowns that are necessary for systematic maintenance. Usually, one has to distinguish between instruments and process faults. The detection of these two types of faults do not involve the same degree of difficulty. Instruments faults can often be modeled by an *additive* change in a state-space model, whereas process faults are more often *nonadditive* changes in the state of such models (see section 1.3).

In this situation, fast detection is often of crucial importance, for the reconfiguration of the control law, for example. Two uses of the change detection methodology in this framework are of interest. The first is related to the automatic processing of individual signals, as we discuss in the next paragraph. The second is more involved, from the point of view of the modeling information that is required. If the detection of the process faults is desired, and not only that of the instrument faults, or if isolation information is desired, then a partial knowledge of the physical model of the process is required to achieve the *diagnosis* of the fault in terms of its location in the process and physical cause. Both geometrical tools from system theory and statistical tools for change detection are used in these situations. We refer to example 1.2.5 for further discussion of these issues.

1.1.1.2 Segmentation of Signals

Now we discuss another important practical motivation for change detection. In recognition-oriented signal processing, the *segmentation of signals* refers to the automatic decomposition of a given signal into stationary, or weakly nonstationary, segments, the length of which is adapted to the local properties of the signal. As we show in examples 1.2.3 and 1.2.4, the change detection methodology provides preferential tools for such an automatic segmentation, which can be achieved either on-line or off-line. In this situation, the problems of interest are the detection of the changes in the local characteristics, and the estimation of the places, in time or space, where the changes occur. False alarms are relatively less crucial than in the previous case of fault detection, because they can be dealt with at the next stage of the recognition system. For example, in continuous speech processing, these algorithms can be used for detecting true abrupt changes. However, in practice these algorithms also give relevant results in less simple situations, for more slow transitions between pieces of signal where the properties of the signal are in fact slowly time-varying before and after the abrupt change [André-Obrecht, 1988]. The same is true in biomedical and seismic signal processing, where several segmentation algorithms have been derived and used for detecting onsets of spikes in EEG signals or P-waves in ECG signals or S-waves and P-waves in seismic data. We refer to examples 1.2.3 and 1.2.4

for additional comments on this point. Finally, let us emphasize that this context of signal segmentation is also valid in the framework of monitoring of industrial processes, where the analysis of individual sensors or actuators signals, without using any information about the model of the whole system, can bring key information for monitoring and fault detection.

1.1.1.3 Gain Updating in Adaptive Algorithms

Adaptive identification algorithms basically can track only slow fluctuations of characteristic parameters. For improving their tracking performances when quick fluctuations of the parameters occur, a possible solution consists of detecting abrupt changes in the characteristics of the analyzed system. The estimation of the change time and magnitude basically allows a more accurate updating of the gains of the identification algorithm. Such an approach has proved useful for tracking maneuvering targets, for example, as in [Willsky, 1976, Favier and Smolders, 1984].

1.1.1.4 Summary

The motivations leading to the change detection framework and methodology can be summarized as follows :

- From the theoretical point of view, it allows us to process abrupt changes and thus it is a natural counterpart to the *adaptive* framework and state of the art which basically can deal only with slow changes; and it is *one* way to approach the analysis of *nonstationary* phenomena.
- From the practical point of view, statistical decision tools for detecting and estimating changes are of great potential interest in three types of problems :
  1. quality control and fault detection in measurement systems and industrial processes in view of improved performances and condition-based maintenance;
  2. automatic segmentation of signals as a first step in recognition-oriented signal processing;
  3. gain updating in adaptive identification algorithms for improving their tracking ability.

The implementation of change detection techniques in these three types of situations is generally achieved with the aid of different philosophy and constraints (e.g., by choosing different models and criteria and by tuning of the parameters of the detectors), but basically the same methodology and tools apply in all situations.

1.1.2 Problem Statements and Criteria

We now discuss change detection problems from the point of view of mathematical statistics. We describe several problem statements and give the intuitive definitions of the corresponding criteria. Statistical change detection problems can be classified into three main classes for several reasons. The first lies in the theoretical definition of criteria used for deriving the algorithms; the second motivation comes from practical experience with different types of problems; and the last reason is historical, as sketched below. Recalling that the formal definition of criteria is given at the end of chapter 4, we now describe these three classes of problems.

1.1.2.1 On-line Detection of a Change

A preliminary statement for this question can be formulated as follows. Let $(y_k)_{1\leq k\leq n}$ be a sequence of observed random variables with conditional density $p_\theta(y_k|y_{k-1},\ldots,y_1)$. Before the unknown change time t_0, the conditional density parameter θ is constant and equal to θ_0. After the change, the parameter is equal

to θ_1. The on-line problem is to detect the occurrence of the change as soon as possible, with a fixed rate of false alarms before t_0. The estimation of the change time t_0 is not required. The estimation of the parameters θ_0 and θ_1 is not required, but sometimes can be partly used in the detection algorithm. We implicitly assume that, in case of multiple change times, each change is detected quickly enough, one after the other, such that at each time instant only one change has to be considered. In the on-line framework, the detection is performed by a *stopping rule*, which usually has the form

$$t_a = \inf\{n : g_n(y_1, \ldots, y_n) \geq \lambda\} \tag{1.1.1}$$

where λ is a threshold, and $(g_n)_{n\geq 1}$ is a family of functions of n coordinates. Simple stopping rules are presented in chapter 2. The *alarm time* t_a is the time at which the change is detected. Note that if $t_a = n$, it is sufficient to observe the sample up to time n, which explains the name of "sequential" or on-line point of view.

In the on-line framework, the criteria are the *delay for detection*, which is related to the ability of the algorithm to set an alarm when a change actually occurs, and the *mean time between false alarms* (see section 4.4). Usually the overall criterion consists in *minimizing the delay for detection for a fixed mean time between false alarms*. We explain in section 4.4 that, from the mathematical point of view, two different definitions of the delay can be stated, which give rise to different results and proofs of optimality for various change detection algorithms.

1.1.2.2 Off-line Hypotheses Testing

We now consider the hypotheses "without change" and "with change." This off-line hypotheses testing problem can be formally stated as follows. Given a finite sample $y_1, \ldots, y_N$, test between

$$\begin{aligned} \mathbf{H}_0 &: \text{ for } 1 \leq k \leq N \ : \ p_\theta(y_k|y_{k-1}, \ldots, y_1) = p_{\theta_0}(y_k|y_{k-1}, \ldots, y_1) \\ \mathbf{H}_1 &: \text{ there exists an unknown } 1 \leq t_0 \leq N \text{ such that:} \\ \text{for } 1 \leq k \leq t_0 - 1 &: p_\theta(y_k|y_{k-1}, \ldots, y_1) = p_{\theta_0'}(y_k|y_{k-1}, \ldots, y_1) \\ \text{for } t_0 \leq k \leq N &: p_\theta(y_k|y_{k-1}, \ldots, y_1) = p_{\theta_1}(y_k|y_{k-1}, \ldots, y_1) \end{aligned} \tag{1.1.2}$$

In this problem statement, the estimation of the change time t_0 is not required.

As we explain in section 4.1, the usual criterion used in hypothesis testing is a tradeoff between the ability to detect actual changes when they occur, which requires a great sensitivity to high-frequency effects, and the ability not to detect anything when no change occurs, which requires a low sensitivity to noise effects. These are obviously two contradictory requirements. The standard criterion is usually to *maximize the probability of deciding* $\mathbf{H}_1$ *when* $\mathbf{H}_1$ *is actually true (i.e., the power), subject to the constraint of a fixed probability of deciding* $\mathbf{H}_1$ *when* $\mathbf{H}_0$ *is actually true (i.e., the size or probability of false alarms)*. In section 4.4, we explain why this criterion is especially difficult to use in the statistical change detection framework.

1.1.2.3 Off-line Estimation of the Change Time

In this problem statement, we consider the same hypotheses as before, and we assume that a change does take place in the sample of observations. Let $(y_k)_{1\leq k\leq N}$ be this sequence of random observations with conditional density $p_\theta(y_k|y_{k-1}, \ldots, y_1)$. Before the unknown change time t_0, which is assumed to be such that $1 \leq t_0 \leq N$, the parameter θ of the conditional density is constant and equal to θ_0. After the change, the parameter is equal to θ_1. The unknown change time has to be estimated from the observations $y_1, \ldots, y_N$ $(1 \leq N < \infty)$ with maximum accuracy. The estimation of t_0 can possibly use information about θ_0 and θ_1,

the availability of which depends upon situations. In such an estimation problem, we *intentionally leave out the search for multiple change times* between 1 and N.

The problem is to estimate t_0. This problem is a typical estimation problem for a discrete parameter. Obviously, this estimate has to be as accurate as possible. Usually, this accuracy is estimated by the probability that the estimate belongs to a given confidence interval, or by the first two moments of the probability distribution of the estimation error (bias and standard deviation).

Other types of criteria for deriving change estimation algorithms are discussed in [Bojdecki and Hosza, 1984, Pelkowitz, 1987].

1.1.2.4 Summary

In some practical applications all three types of problems may have to be solved together. We also emphasize here that an off-line point of view may be useful to design a decision and/or estimation function that is finally implemented on-line. We discuss this in section 1.4.

The *five intuitive performance indexes* for designing and evaluating change detection algorithms are the following :

1. *mean time between false alarms*;
2. *probability of false detection*;
3. *mean delay for detection*;
4. *probability of nondetection*;
5. *accuracy of the change* time and magnitude *estimates*.

We use these five indexes throughout the book. Another useful index consists of the Kullback information between the distributions before and after change. This distance does have a strong influence on the above-mentioned performance indexes, and we use it as a weak performance index when discussing *detectability* issues.

Another property of change detection algorithms is of great practical importance, and that is the *robustness*. Algorithms that are robust with respect to noise conditions and to modeling errors, and that are easy to tune on a new signal, are obviously preferred in practice. These robustness features cannot easily be formally stated, but should definitely be kept in mind when designing and experiencing change detection algorithms. This issue is discussed in several places in this book.

1.1.3 Purpose of the Book

This book is basically devoted to the design and investigation of *on-line change detection algorithms*. The off-line problem statement is discussed much more briefly, and mainly with a view to the discussion of some applications.

When designing change detection and estimation algorithms, it may be useful to distinguish two types of tasks :

1. *Generation of "residuals"* : These artificial measurements are designed to reflect possible changes of interest in the analyzed signal or system. They are, for example, ideally close to zero when no change occurs and significantly different from zero after the change. This is the case of the so-called *parity checks*, designed with the aid of the analytical redundancy approach. In other cases, the mean value or the spectral properties of these residuals may change when the analyzed system is changing. From the mathematical statistics point of view, a convenient way for generating these artificial measurements consists of deriving *sufficient statistics*.

2. *Design of decision rules* based upon these residuals : This task consists of designing the convenient decision rule which solves the change detection problem as reflected by the residuals.

In this book, we mainly focus on *parametric statistical tools* for detecting abrupt changes in properties of *discrete time* signals and dynamic systems. We intend to present didactically generalizations of points of view for designing algorithms together with new results, both theoretical and experimental, about their performances. The starting point is elementary well-known detectors used in industrial applications. We then generalize this approach in two tasks to more complex situations in which spectral properties of signals or dynamic properties of systems change. This book is intended to be a bridge between mathematical statistics tools and applied problems. Therefore, we do not derive all mathematical statistics theories, and readers who want complete mathematical results and proofs must consult other books or papers, indicated in references.

Even though great emphasis is placed on task 2, we also address the problem of deriving solutions for task 1. Deterministic solutions, such as in the analytical redundancy approach, are often based on geometrical properties of dynamic systems, as discussed further in section 1.4 and later in chapter 7. Mathematical statistics solutions, such as sufficient statistics or the so-called local approach, are further described in the following chapters, especially chapter 8.

1.2 Application Examples

In this section, we describe five typical application examples of change detection techniques. For each example, we give a short description of the particular problem and its context, including the main references. For some of these models, the detailed information about the possibly complex underlying physical models is given in chapter 11. This selection of examples is not exclusive; it is intended to give only sufficient initial insights into the variety of problems that can be solved within this framework, and to serve as much as possible as a common basis for all the algorithmic equipment presented in the subsequent chapters. In chapter 11, we come back to application problems, showing results of processing real signals with the aid of change detection algorithms, and discussing several potential application domains.

In the present chapter, the five examples are ranged according to the increasing complexity of the underlying change detection problems. We start with quality control and condition monitoring of inertial navigation systems (examples 1.2.1 and 1.2.2). Then we describe seismic signal processing and segmentation of signals (examples 1.2.3 and 1.2.4). Finally, we discuss failure detection in mechanical systems subject to vibrations (example 1.2.5).

1.2.1 Quality Control

One of the earliest applications of change detection is the problem of quality control, or continuous production monitoring. On-line quality control procedures are used when decisions are to be reached sequentially, as when measurements are taken. Consider a production process that can be *in control* and *out of control*. Situations where this process leaves the in control condition and enters the out of control state are called *disorders*. For many reasons, it is necessary to detect the disorder and estimate its time of occurrence. It may be a question of safety of the technological process, quality of the production, or classification of output items of production. For all these problems, the best solution is *quickest detection of the disorder with as few false alarms as possible*. This criterion is used because the delay for detection is a period of time during which the technological process is out of control without action of the monitoring system. From both safety and quality points of view, this situation is obviously highly nondesirable. On the other hand, frequent false

alarms are inconvenient because of the cost of stopping the production and searching for the origin of the defect; nor is this situation desirable from psychological point of view, because the operator will very quickly stop using the monitoring system. Nevertheless, the optimal solution, according to the above-mentioned criterion, is basically a *tradeoff* between quick detection and few false alarms, using a comparison between the losses implied by the two events.

We stress here that we solve this problem using a *statistical approach*. From this point of view, the samples of measurements are a realization of a random process. Because of this random behavior, large fluctuations can occur in the measurements even when the process is in control, and these fluctuations result in false alarms. On the other hand, a given decision rule cannot detect the change instantaneously, again because of the random fluctuations in the measurements. When the technological process is in control, the measurements have a specific probability distribution. When the process is out of control, this distribution changes. If a parametric approach is used, we speak about changes in the parameters of this probability distribution. For example, let us consider a chemical plant where the quality of the output material is characterized by the concentration of some chemical component. We assume that this concentration is normally distributed. Under normal operating conditions, the mean value and standard deviation of this normal distribution are μ_0 and σ_0, respectively. We also assume that under faulty conditions, two basic types of changes can occur in these parameters :

- deviation from the reference mean value μ_0 towards μ_1, with constant standard deviation, as depicted in figure 1.1; in other words, this type of change is a systematic error. This example serves as a common basis for depicting the typical behavior of all the algorithms presented in chapter 2.
- increase in the standard deviation from σ_0 to σ_1, with constant mean, as depicted in figure 1.2. This type of change is a random error.

Composite changes can also occur. The problem is to design a statistical decision function and a decision rule that can detect these disorders. The typical behavior of such a decision function is depicted in figure 1.1.

In the simplest case, all the parameters of each of the two above-mentioned situations are assumed to be known. The tuning of a statistical decision rule is then reduced to the choice of a threshold achieving the requested tradeoff between the false alarm rate and the mean delay for detection. Several types of decision rules are used in industry as standards and are called *control charts*. They are described in detail in section 2.1.

The main references for quality control are [Aroian and Levene, 1950, Goldsmith and Whitfield, 1961, Van Dobben De Bruyn, 1968, Bissell, 1969, Phillips, 1969, Gibra, 1975, Wetherill and Brown, 1991]. Other references can be found in chapter 11.

1.2.2 Navigation System Monitoring

Navigation systems are typical equipments for planes, boats, rockets, and other moving objects. Important examples of such systems are inertial navigation systems, radionavigation systems, and global satellite navigation sets for planes. An inertial navigation system has two types of sensors : gyros and accelerometers. Using this sensor information and the motion equations, the estimation of the coordinates and the velocities of the moving object can be achieved. In view of safety and accuracy requirements, redundant fault-tolerant measurement systems are used. The first task of such a type of system is *detection and isolation of faulty sensors*. This problem can be stated as a particular change detection problem in some convenient modeling framework, as discussed in detail in chapter 11. The criterion to be used is again *quick detection and few false alarms*. Fast detection is necessary because, between the fault onset time and the detection time, we use abnormal measurements in the navigation equations, which is highly undesirable. On the other hand,

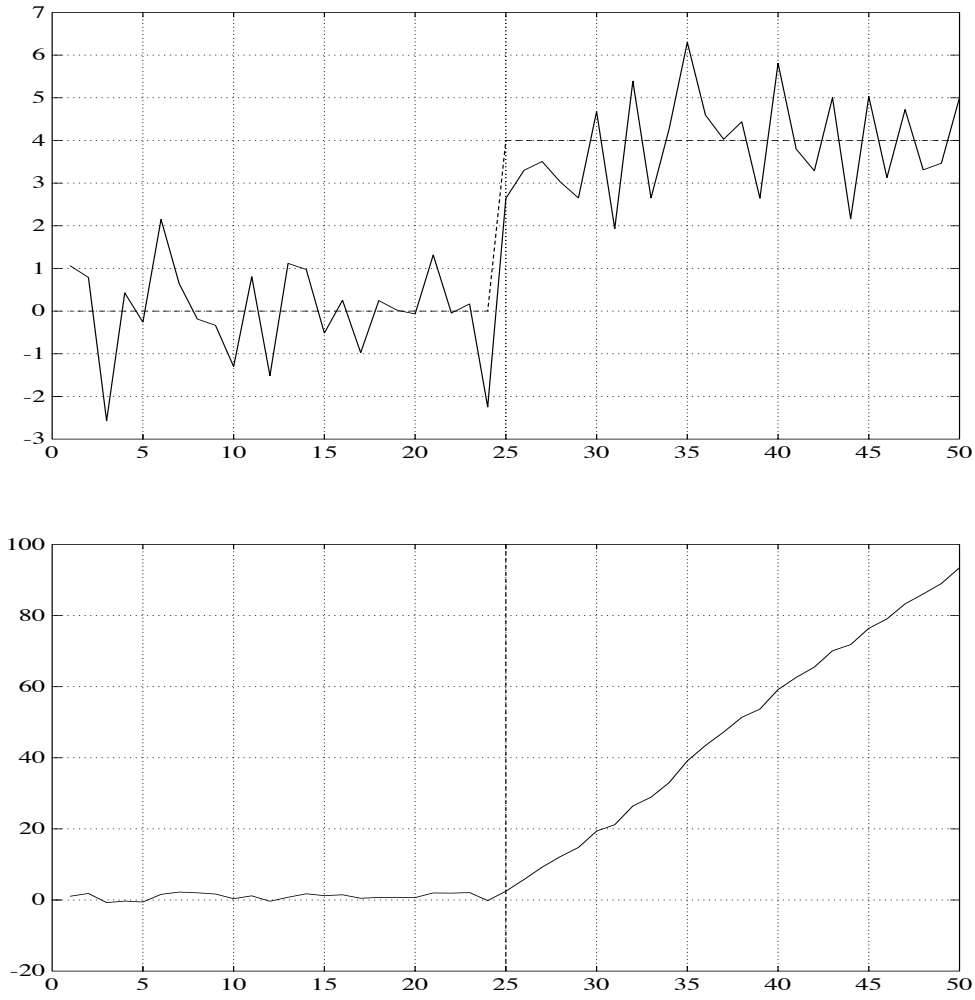

Figure 1.1 Increase in mean with constant variance and the typical behavior of the decision function in quality control.

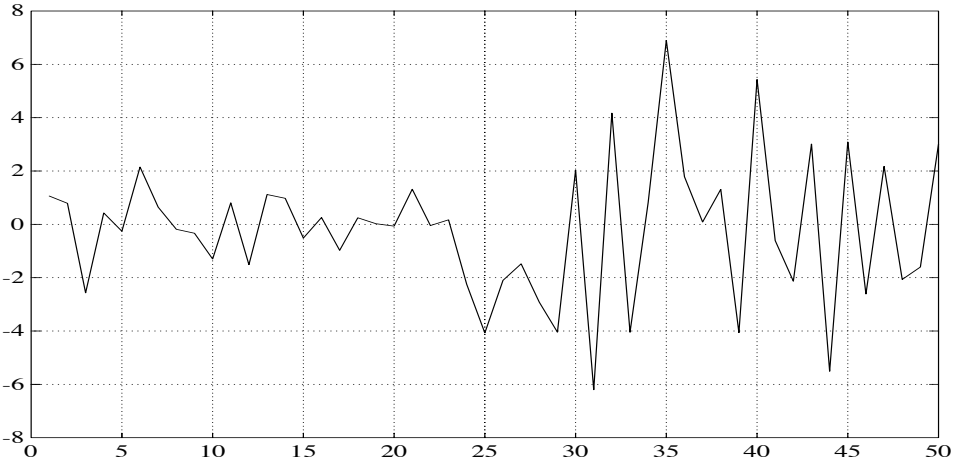

Figure 1.2 Increase in variance with constant mean.

false alarms result in lower accuracy of the estimate because some correct information is not used. The optimal solution is again a tradeoff between these two contradictory requirements. For radionavigation systems, integrity monitoring using redundant measurements is an important problem and is generally solved with the aid of the same criteria.

The main references for the monitoring of inertial navigation systems are [Newbold and Ho, 1968, Willsky *et al*., 1975, Satin and Gates, 1978, Kerr, 1987]. Integrity monitoring of navigation systems is investigated in [Sturza, 1988]. Other references can be found in chapter 11.

1.2.3 Seismic Data Processing

Let us now discuss some typical problems of seismic data processing. In many situations, it is necessary to estimate *in situ* the geographical coordinates and other parameters of earthquakes. Typical three-dimensional signals are shown in figure 1.3, and comprise E-W, Z, and N-S measurements. The two main events to be detected are the P-wave and the S-wave; note that the P-wave can be very "small." From the physical background in seismology, which we explain in chapter 11, it results that the processing of these three-dimensional measurements can be split into three tasks :

1. on-line detection and identification of the seismic waves;
2. off-line estimation of the onset times of these waves;
3. off-line estimation of the azimuth using correlation between components of P-wave segments.

From now on, we consider only the first two questions. Detection of the P-wave has to be achieved *very quickly with a fixed false alarms rate*. The main reason for this is to allow S-wave detection in this on-line processing. P-wave detection is a difficult problem, because the data contain many nuisance signals coming from the environment of the seismic station, and discriminating between these events and a true P-wave is not easy. The same situation holds for the S-wave, where the difficulty is greater, because of low signal-to-noise ratio and numerous nuisance signals between P-wave and S-wave.

After P-wave and S-wave detection, *off-line accurate estimation of onset times* is requested for both types of waves. As we explain in chapter 11, a possible solution consists of using some fixed size samples

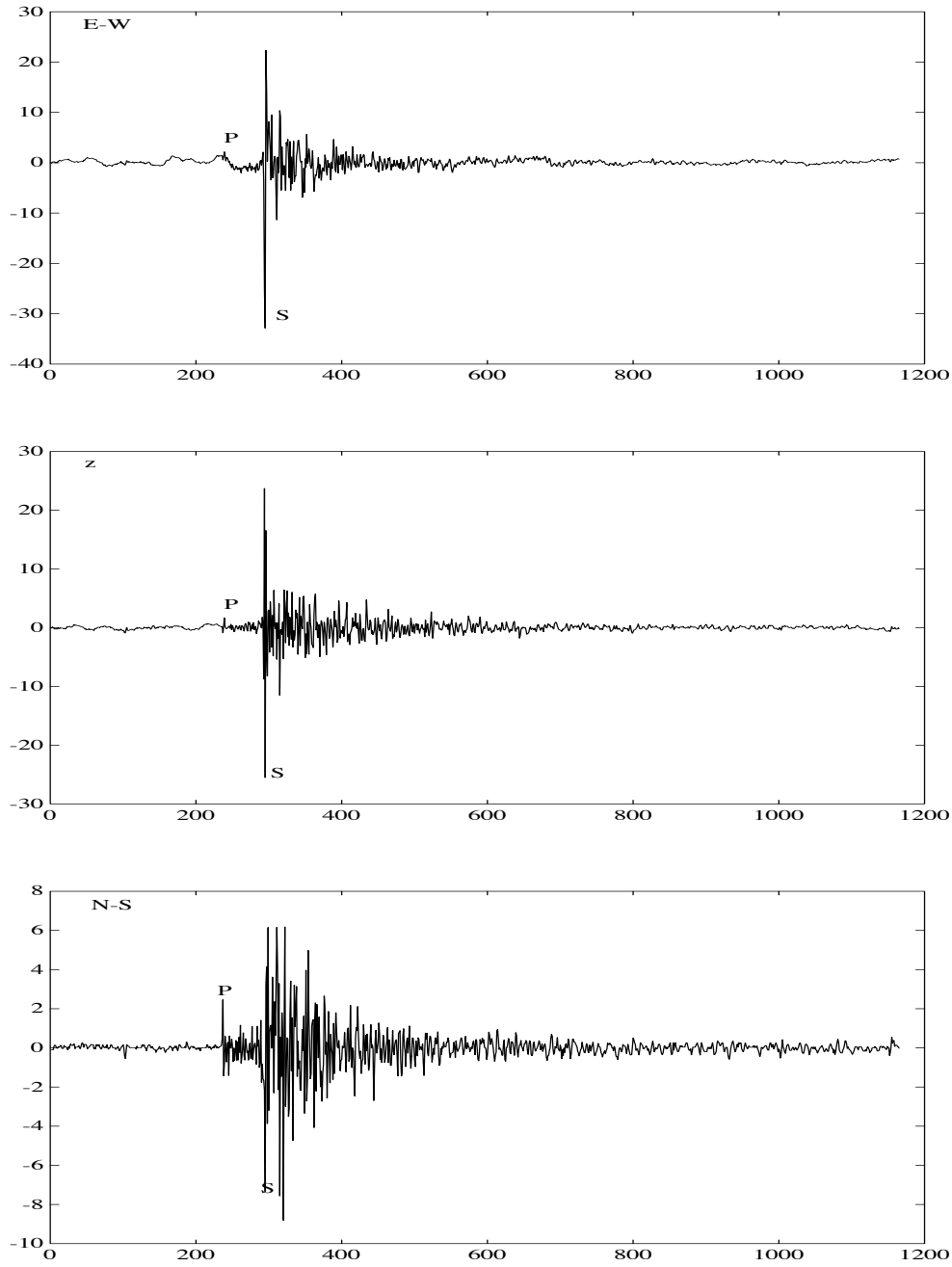

Figure 1.3 The three typical components of a seismogram : E-W, Z and N-S (Courtesy of the Academy of Sciences of USSR, Far Eastern Scientific Center, Institute of Sea Geology) .

of the three-dimensional signals, centered at a rough estimate of the onset time provided by the detection algorithm. This off-line change time estimation is described in section 8.7.

The main references for seismic data processing are [Tjostheim, 1975, Kitagawa and Gersch, 1985, Nikiforov and Tikhonov, 1986, Pisarenko *et al*., 1987, Nikiforov *et al*., 1989]. Other references can be found in chapter 11.

1.2.4 Segmentation of Signals

A possible approach to recognition-oriented signal processing consists of using an automatic segmentation of the signal as the first processing step. A segmentation algorithm splits the signal into homogeneous segments, the lengths of which are adapted to the local characteristics of the analyzed signal. The homogeneity of a segment can be in terms of the mean level or in terms of the spectral characteristics. This is discussed further when we introduce the additive and nonadditive change detection problems. The segmentation approach has proved useful for the automatic analysis of various biomedical signals, for example, electroencephalograms [R.Jones *et al*., 1970, Bodenstein and Praetorius, 1977, Sanderson and Segen, 1980, Borodkin and Mottl', 1976, Ishii *et al*., 1979, Appel and von Brandt, 1983], and electrocardiograms [Gustafson *et al*., 1978, Corge and Puech, 1986]. Segmentation algorithms for recognition-oriented geophysical signal processing are discussed in [Basseville and Benveniste, 1983a]. More recently, a segmentation algorithm has been introduced as a powerful tool for the automatic analysis of continuous speech signals, both for recognition [André-Obrecht, 1988] and for coding [Di Francesco, 1990]. An example of automatic segmentation of a continuous (French) speech signal[1] is shown in figure 1.4. Other examples are discussed in chapter 11.

The main desired properties of a segmentation algorithm are *few false alarms and missed detections, and low detection delay*, as in the previous examples. However, keep in mind the fact that the segmentation of a signal is often nothing more than the first step of a recognition procedure. From this point of view, it is obvious that the properties of a given segmentation algorithm also depend upon the processing of the segments which is performed at the next stage. For example, it is often the case that, for segmentation algorithms, false alarms (sometimes called oversegmentation) are less critical than for onset detection algorithms. A false alarm for the detection of an imminent tsunami obviously has severe and costly practical consequences. On the other hand, in a recognition system, false alarms at the segmentation stage can often be easily recognized and corrected at the next stage. A segmentation algorithm exhibiting the above-mentioned properties is potentially a powerful tool for a recognition system.

It should be clear that a segmentation algorithm allows us to detect several types of events. Examples of events obtained through a spectral segmentation algorithm and concerning recognition-oriented speech processing are discussed in [André-Obrecht, 1988, André-Obrecht and Su, 1988, André-Obrecht, 1990].

1.2.5 Vibration Monitoring of Mechanical Systems

Let us now describe the *vibration monitoring* problem and its connection with change detection. For both complex mechanical structures, such as offshore platforms, bridges, buildings, and dams, and rotating machines, such as turbo-alternators and gearing systems, it is of crucial interest to monitor the vibrating characteristics without using artificial excitation or stop-down, but in the usual functioning mode under natural or usual excitation (swell, road traffic, wind, water pressure, earthquakes, big works in the neighborhood, steam). The vibrating characteristics of a mechanical structure or machine basically reflects its state of health, and any deviation in these characteristics brings information of importance to its functioning mode. The main difficulty in this problem is that the measured signals (accelerometers, gauges) reflect both the

[1] This result is due to Régine André-Obrecht. The help of Bernard Delyon in drawing this figure is also gratefully acknowledged.

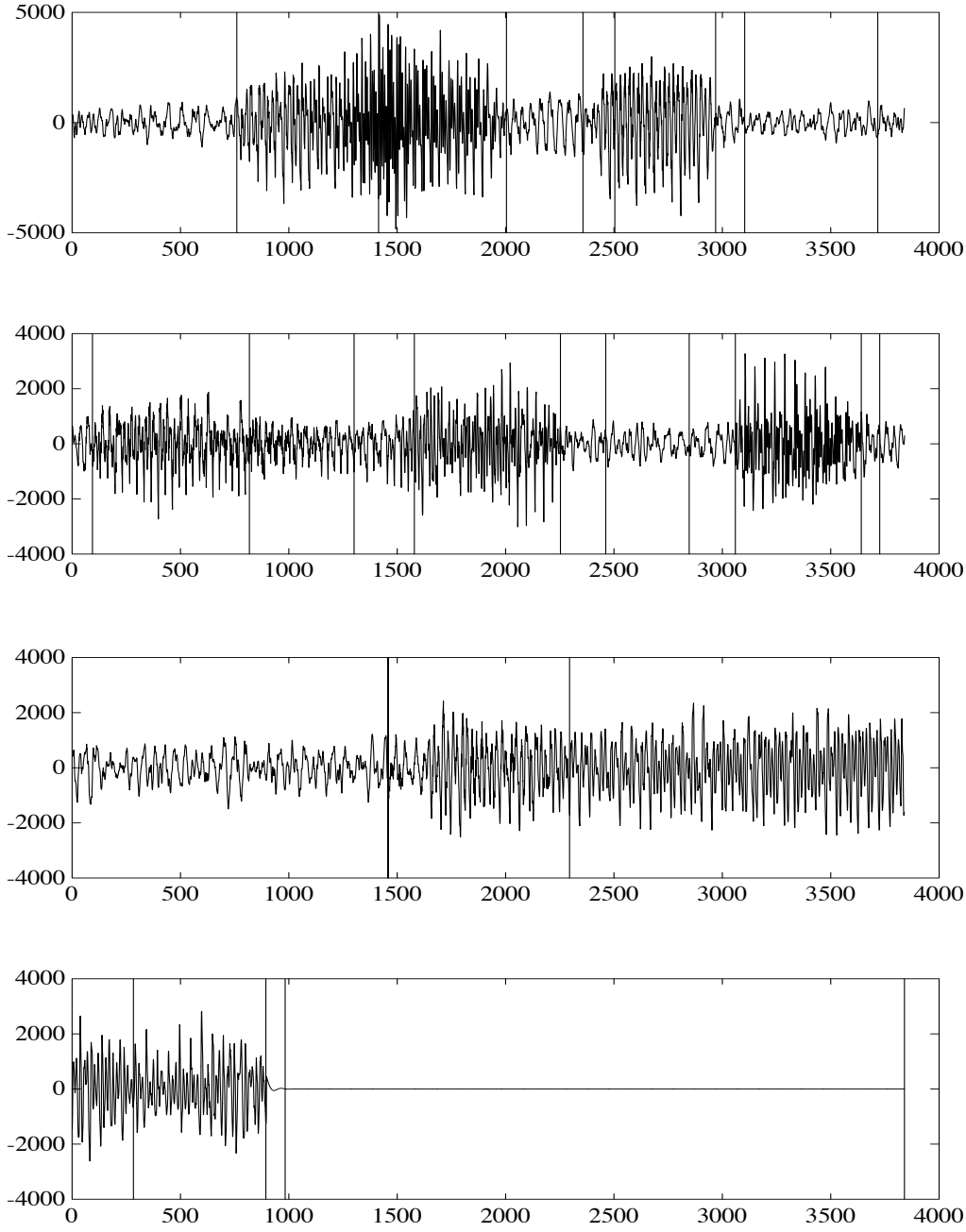

Figure 1.4 An example of speech signal segmentation. The estimated change times (vertical lines) provide us with the boundaries of the segments.

nonstationarities due to the surrounding excitation, which is always highly time-varying, and the nonstationarities due to changes in the eigen characteristics of the mechanical object itself. We show in chapter 11 that this vibration monitoring problem can be stated as the problem of detecting changes in the AR part of a multivariable ARMA model having nonstationary MA coefficients. Typical changes to be detected have a magnitude of about 1% of the eigenfrequencies. The second difficult problem is to diagnose or isolate the detected changes, either in terms of the vibrating characteristics (eigenvalues and eigenvectors), or in terms of the mechanical characteristics (masses, stiffness coefficients, together with an approximate localization in the mechanical object). These questions are investigated in detail in section 9.3. The criteria that are to be used in such a problem are *few false alarms* and ability to *detect small changes* in possibly long samples of data.

Small mechanical systems, such as a small number of masses connected by springs, often serve as laboratory experimental setups for simulating more complex vibrating structures. Thus, they can be used for testing fault detection and diagnostic tools. Examples can be found in [Kumamaru *et al*., 1989, Basseville *et al*., 1987a]. The models of these simulation examples are described in the appendix to chapter 11. The main references concerning signal processing methods for vibration monitoring can be found in [Braun, 1986].

1.3 Content of the Book

In this section, we describe in detail the content of the book. First, let us give some comments referring to the three problem statements described in section 1.1. Even though several chapters address the second and third problems, the main emphasis of this book is on the first problem, namely *on-line change detection using a parametric statistical approach*. We now describe the general organization of the book, then the content of each chapter; finally, we present and discuss the flowchart of the book.

1.3.1 General Organization

The organization of the chapters follows a simple distinction between changes in the *scalar parameter* of an *independent* sequence of observations, and changes in the *multidimensional parameter* of a *dependent* sequence. Thus, we divide the book into two main parts corresponding to these two sets. A third part is devoted to the tuning and application issues. The organization of the second part about multidimensional changes is based on a classification of change detection problems into two categories : *additive* changes and *nonadditive* (or spectral) changes. Basically, we mean that changes can be viewed as either additive or multiplicative on the transfer function of the considered signal or system. Equivalently, changes can be viewed as either changes in the mean value of the law of the observed signals, or changes in the correlations. A more thorough discussion about this classification can be found at the beginning of Part II.

1.3.2 Description of Each Chapter

Before proceeding, let us mention that, at the end of each chapter, the reader can find notes and bibliographical references concerning the problems discussed and a summary of the key results.

Part I is devoted to *changes in the scalar parameter* of an independent sequence. In chapter 2, we introduce the reader to the theory of on-line change detection algorithms in the framework of an independent random sequence parameterized by a scalar parameter. We first consider the case of known parameters before and after change. In section 2.1, we begin with the description of elementary algorithms of common use in industrial applications (quality control, for example) : these are Shewhart control charts, finite or infinite moving average control charts, and filtered derivative algorithms. In section 2.2, we introduce a key detection tool, the CUSUM algorithm, which we derive using both on-line and off-line points of view.

In section 2.3, we describe Bayes-type algorithms. In the case of an unknown parameter after change, we discuss two possible solutions in section 2.4 : the weighted CUSUM and the generalized likelihood ratio (GLR). In section 2.5, we discuss how algorithms used for detecting changes can improve the tracking ability of an adaptive identification scheme. Finally, in section 2.6, we discuss the two off-line problem statements introduced in subsection 1.1.2 : off-line hypotheses testing and estimation of the change time.

Chapters 3 and 4 are an excursion outside the part devoted to changes in a scalar parameter, and are aimed at the presentation of all the *theoretical backgrounds* to be used throughout the book. Chapter 3 is composed of two sections. The first is devoted to the presentation of the main results from probability theory, including conditional probability and expectation, Brownian motion and diffusion processes, martingales, and stopping times. In section 3.2, we summarize some results from the control literature, namely observers, Kalman filter, and connections between state-space and ARMA models. Chapter 4 is composed of four sections. Section 4.1 is concerned with some basic results about estimation and information from a mathematical statistics point of view. Section 4.2 is devoted to statistical hypotheses testing, including expansion of likelihood ratios, and section 4.3 to sequential analysis. Finally, in section 4.4, we formally define the criteria for designing and evaluating change detection algorithms in both the on-line and off-line frameworks.

In chapter 5, we come back to changes in the scalar parameter of an independent random sequence, and present the main analytical and numerical results concerning the algorithms presented in chapter 2. We investigate the properties of the elementary algorithms in section 5.1. Then in section 5.2, we describe in detail the properties of CUSUM-type algorithms, following the key results of Lorden. The properties of the GLR algorithm are discussed in section 5.3, together with the role of *a priori* information. Bayes-type algorithms are briefly investigated in section 5.4. Finally, in section 5.5, we present analytical and numerical comparative results. This concludes the Part I.

Part II is concerned with the *extension* of these algorithms to more complex situations of changes, namely changes in the vector parameter of an independent sequence, additive changes in a possibly dependent sequence, and nonadditive changes in a dependent sequence too. The key ideas of Part II are described in chapter 6.

Chapter 7 is devoted to the extension of the key algorithms developed in the independent case considered in chapter 2, to additive changes in more complex models, namely regression, ARMA, and state-space models. In section 7.1, we introduce general additive changes, and explain transformations from observations to innovations and redundancy relations. Section 7.2 deals with the statistical tools for detecting additive changes. We begin by discussing in subsection 7.2.1 what we call the *basic problem* of detecting a change in the mean vector parameter of an independent Gaussian sequence. Then we discuss the extension of the CUSUM-type and GLR detectors to the more general situations of regression, ARMA, and state-space models in subsections 7.2.2, 7.2.3, and 7.2.4, respectively. Still from a statistical point of view, we then discuss the diagnosis or isolation problem and the detectability issue in subsections 7.2.5 and 7.2.6. The properties of these algorithms are discussed in section 7.3. Section 7.4 is devoted to the presentation of geometrical tools for change detection and diagnosis, known as analytical redundancy techniques. We begin the discussion about redundancy by describing in subsection 7.4.1 the direct redundancy often used in the case of regression models. We extend this notion to the temporal redundancy in subsection 7.4.2. In subsection 7.4.3, we describe another technique for generating analytical redundancy relationships. We conclude this section with a discussion of the detectability issue in subsection 7.4.4, again from a geometrical point of view. This chapter about additive changes concludes with section 7.5, which contains a discussion about some basic links between statistical and geometrical tools. Actually, links exist for the design of detection algorithms as well as for the solutions to the diagnosis problem and the detectability definitions.

Chapter 8 addresses the problem of detecting changes in the spectral properties of a scalar signal by using parametric approaches. We mainly focus on on-line algorithms. In section 8.1, we first introduce

spectral changes and explain their specificities and difficulties with respect to additive changes. We show why the transformation from observations to innovations used for additive changes is not sufficient here, and we introduce to the use of the local approach for change detection. In section 8.2, we investigate the general case of conditional probability distributions, and we describe the main ideas for designing on-line algorithms, namely CUSUM and GLR approaches, and possible simplifications, including the local approach, and leading to either linear or quadratic decision functions. All these algorithms are then described in the cases of AR and ARMA models in section 8.3. In section 8.4, we describe the design of non-likelihood-based algorithms, also using the local approach. This extended design allows a systematic derivation of change detection and diagnosis algorithms associated with any recursive parametric identification method. In section 8.5, we discuss the detectability issue. In section 8.6, we discuss the implementation issues related to the fact that, in practice, the model parameters before and after change are not known. In section 8.7, we consider off-line algorithms, using the likelihood approach, and discuss the connection with on-line algorithms.

Chapter 9 is concerned with spectral changes in the multidimensional case, including the diagnosis problem, and the properties of the detection algorithms in both the scalar and the multidimensional cases. In section 9.1, we introduce the key detection tools, namely the likelihood ratio, the local approach, and the non-likelihood-based algorithms, emphasizing the new multidimensional issues. Then in section 9.2, we extend the likelihood-based algorithms of chapter 8 to multidimensional AR and ARMA models. Section 9.3 is concerned with the application of the non-likelihood-based design of algorithms to the problem of the detection and diagnosis of changes in spectral characteristics of multidimensional signals, or equivalently in the eigenstructure of nonstationary multivariable systems. We describe both on-line and off-line detection algorithms. Then we investigate the diagnosis problem from several points of view. The detectability issue is discussed in section 9.4, from a statistical point of view, as in chapters 7 and 8. The theoretical properties of the various algorithms introduced in this and the previous chapters, are investigated in section 9.5. This concludes the Part II.

We begin the Part III with chapter 10, which is devoted to the problems of implementing and tuning change detection algorithms. This chapter is divided into four sections. In section 10.1, we describe a general methodology for implementing and tuning the algorithms. With respect to the design of the algorithms, this methodology is more philosophical than technical, but it relies on the available theoretical results concerning the properties of the algorithms. Section 10.2 is concerned with the tuning of all the techniques introduced in chapter 2 and investigated in chapter 5, namely the algorithms for detecting changes in the scalar parameter of an independent sequence. In section 10.3, we investigate the case of a vector parameter and a linear decision function, and in section 10.4, the case of a quadratic decision function.

In chapter 11, we come back to the applications. The main goals of this chapter are to show examples of the use of change detection algorithms and examples of potential application of the change detection methodology. Of course, the list of application domains that we investigate there is not exhaustive. The examples of the first type are fault detection in inertial navigation systems, onset detection in seismic signal processing, continuous speech signals segmentation, and vibration monitoring. The examples of the second type are statistical quality control, biomedical signal processing, and fault detection in chemical processes.

1.3.3 Flowchart of the Book

In figure 1.5, we show the general organization of the book and suggestions for using it. Two paths can be used for reading this book. The reader interested mainly in the algorithms themselves can start at beginning with the design of the algorithms, proceed through the properties, and finally reach tuning and applications. The reader interested mainly in the practical design and application of the algorithms can start with the applications at the end in order to select his path through the other chapters.

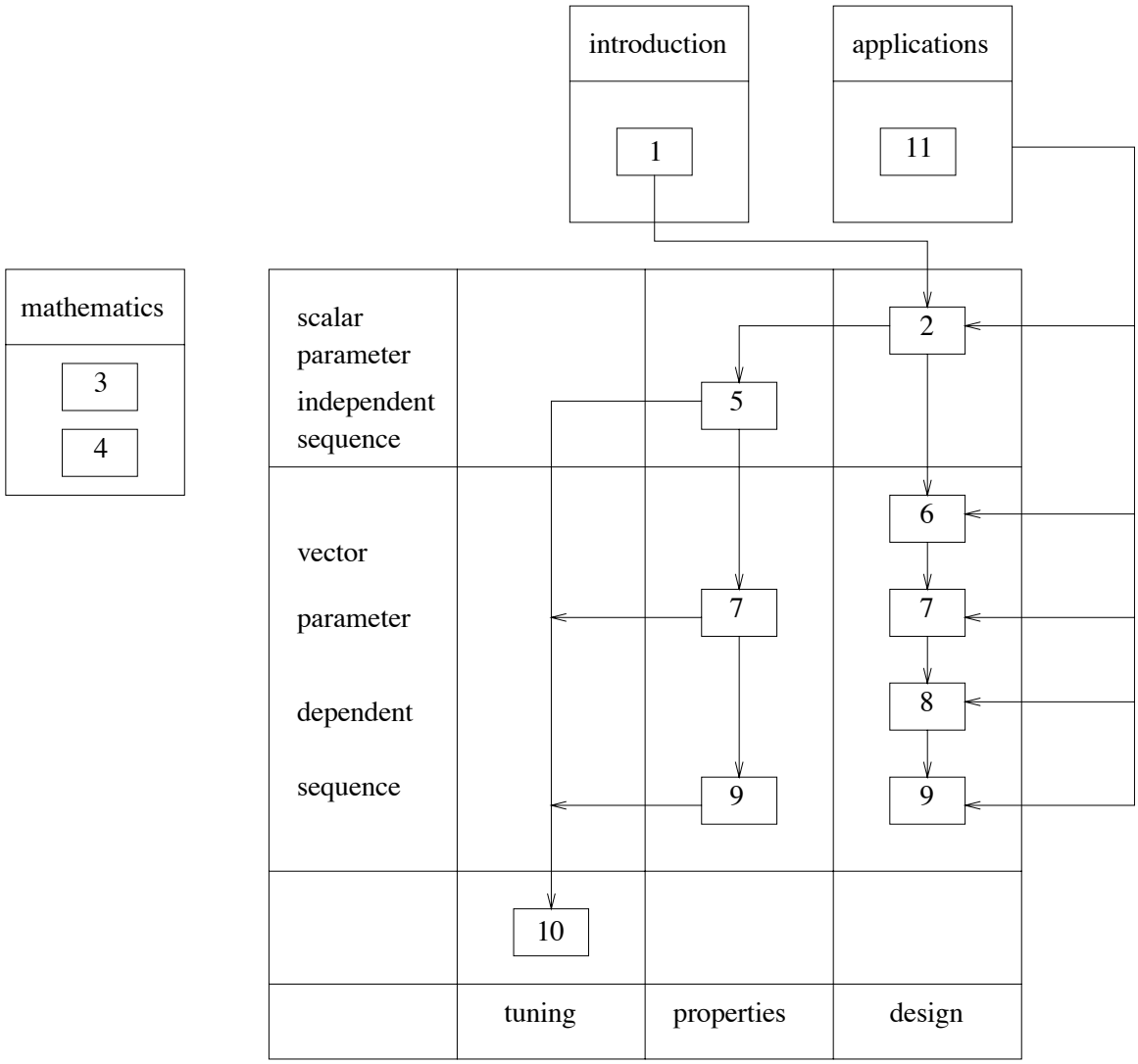

Figure 1.5 Flowchart of the book, showing two paths : one focused on the investigation of the algorithms and the other on the practical design and applications.

1.4 Some Further Critical Issues

In this section, we comment further on some of the issues involved when designing change detection algorithms and investigating their properties. Because we deal mainly with parametric techniques in this book, the key issues of choice of models, use of prior information, redundancy, and nuisance parameters have to be addressed. This is done in the first subsection. We then discuss the properties of the algorithms and detectability.

1.4.1 Designing the Algorithms

We now consider questions related to choice of models and use of prior information, generation of residuals, and nuisance parameters.

1.4.1.1 Choice of Models and Use of Prior Information

When dealing with modeling, and more specifically with parametric techniques, the choice of models is a critical issue. The reader is referred to [Ljung, 1987] for investigation, discussions, and references about linear time-invariant models, or time-varying or nonlinear models. What we would like to stress here is that, in the framework of change detection, the situation is significantly different from the identification point of view. Basically, *models useful for detection and monitoring are usually smaller than physical models and models for identification and recognition* (see the discussion of speech segmentation and recognition in section 11.1).

For example, consider the problem of detecting and diagnosing changes or faults in large structures or industrial processes. Useful results can be obtained with the aid of parametric models of relatively small size with respect to the dimension of the physical model of the process (which is based on partial differential equations, for example) . Even though it is often believed that parametric techniques are useful for diagnosis purposes only when there is a bijection between the parametric model and the physical one (see the survey [Isermann, 1984], for example), diagnosis in terms of the physical model can be inferred from a small black-box parametric model. This has been obtained in the vibration monitoring application introduced in section 1.2, and is discussed in detail in chapter 9. As another example, a relevant segmentation can be obtained with the aid of AR models of order 2, whereas the classification of the resulting segments may very well require AR models of significantly higher orders. This is discussed in section 11.1 for the case of continuous speech signals.

Another important issue when designing change detection algorithms is the use of prior information about the changes. When model structure and parameterization have been chosen, it is useful, if not necessary, to examine what is known about the possible values of the parameters before and after change, and how this prior information should be used.

Referring to the preliminary problem statement, which we formulate in section 1.1, from the on-line point of view, knowing the parameter θ_0 before change is of secondary interest. If θ_0 is unknown, it may be identified with the aid of a convenient identification algorithm. The actual problem lies then in the parameter θ_1 after change. Three cases have to be distinguished :

1. θ_1 is known : this is an easy but unrealistic case. It is often used as a starting point for the design of a detection algorithm, which is then extended to more realistic situations, for example, by replacing unknown parameters by values fixed *a priori* (such as a "minimum" magnitude of jump), or by estimated values. This case is the preferential situation for the derivation of most theoretical optimality results for change detection algorithms, and comparison between these theoretical results and numerical ones

from simulation analysis. It may be also useful to compare empirical estimation of the performances of the change detection algorithm on real data to these theoretical properties. See sections 2.2, 7.2, and 8.2 for examples of such design.

2. Few prior information on values of θ_1 corresponding to interesting changes is available; for example, it is known that there exists a separating hyperplane between the set of values of θ_0 and the set of values of θ_1. How this type of information can be used in the design of a change detection algorithm is explained in sections 7.2 and 8.2.

3. Nothing is known about θ_1. This situation is obviously the most interesting from a practical point of view, but also the most difficult from the point of view of the design and the investigation of the properties of the algorithms. Two main approaches exist for solving this problem, and are described in sections 2.4, 7.2 and 8.2. Because the corresponding algorithms are complex, we also investigate possible simplifications.

1.4.1.2 Redundancy Relationships

We now discuss the use of analytical redundancy relationships for change detection. As we stated at the end of section 1.1, one possible general approach for change detection consists of splitting the task into (1) generation of residuals, which are, for example, ideally close to zero when no change occurs, and significantly different from zero after a change, and (2) design of decision rules based on these (possibly non-statistically optimal) residuals. One way of obtaining such residuals is to use analytical redundancy relations. For example, in chemical processes, static balance equations are helpful for detecting failures in pipes, sensors, and actuators for fluid flows.

For other systems, a complete model may be available and can be used in the formal statistical change detection approach. In this case, the generation of residuals is basically included in the derivation of the algorithm itself and does not have to be considered as a separate task. For example, we use this point of view for discussing fault detection in an inertial navigation system.

There exists a bridge between these two types of solutions, and in chapter 7 we show in which cases they are equivalent.

1.4.1.3 Nuisance Parameters

Assume that a parametric model is characterized by a parameter vector which is divided into two subsets : one subset is useful for detecting changes in the properties of the underlying object; the other subset contains information about the object or its environment, but the changes in this subset are not of interest. It turns out that very often these nuisance parameters are highly involved with the useful parameters, and thus have an influence on the decision function. The use of change detection algorithms that do not take into account this fact leads to additional false alarms and missed detections. A specific design of the change detection algorithm must be used in this case. The so-called min-max approach is introduced in chapter 4 for this purpose. A problem that is very close to the question of nuisance parameters is the problem of isolation or diagnosis. We show in section 7.2 how to use this specific approach to design change detection algorithms to solve the isolation problem. Another example is investigated in section 9.3, where we show that it is possible to design decision functions that decouple as much as possible these two parameter subsets, for example, AR and MA parameters in ARMA models.

1.4.2 Investigating the Algorithms

The investigation of the properties of algorithms is useful for two purposes : First, it helps us understand what can be gained in practice when using such algorithms; and second, it gives answers to the optimality issues. We now discuss these points, distinguishing between the properties that result from a formal definition of criteria and those that result from a weaker but useful performance index.

1.4.2.1 Properties of the Algorithms

The mean delay for detection and the mean time between false alarms are the two key criteria for on-line change detection algorithms. As we discuss in chapter 5, in some cases there exist optimal algorithms that minimize the mean delay for a given mean time between false alarms. From a practical point of view, knowledge of the values of these performance indexes for sets of parameters is useful. A key tool for investigating the properties of on-line change detection algorithms is the so-called *average run length function*, which concentrates the information about both these performances indexes. The computation of this function is difficult for most of the practically relevant change detection problems. For this reason, we introduce numerical algorithms for the evaluation of this function. We also introduce a weaker performance index, which we call *detectability*, that is strongly connected with the two previous criteria and can be computed in more complex cases.

1.4.2.2 Detectability

For defining the detectability of a given change, two levels can be considered. The first investigates which changes are detectable and which are not. In the same way that observability and controllability depend on the observation and control matrices of the system, the detectability depends on the statistical tool that is used for detection. Therefore, the detectability of a change should be defined in terms of the effect or signature that it produces on the "sufficient" statistic that is used in the decision rule. For example, if the statistic reflects possible changes in the system by changes in its own mean value, any change that does not modify the mean value of the statistic is not detectable.

A second level defines the detectability as a performance index of the decision rule. We discuss this detectability issue using both statistical and geometrical points of view, and unify the different definitions into the framework of *information*. More precisely, in the statistical point of view, we define the detectability of a change with the aid of an intrinsic feature of the system, namely the mutual information between the two models before and after change. We show that, surprisingly, the two points of view – "the detectability depends upon the detection tool which is used" and "the detectability is an intrinsic feature of the analyzed system" – basically lead to only one definition of detectability. We discuss these detectability issues in subsections 7.2.6, 7.4.4, and 7.5.4, and again in sections 8.5 and 9.4.

1.5 Notes and References

In this section, we give some historical notes and then references for seminars, survey papers, and books related to change detection. We believe it of interest to put the *on-line* change detection framework, motivations, and methodology in a historical perspective. Because this subject basically grew up at the confluence of several disciplines, a complete historical picture is difficult to draw. Our partial knowledge can be summarized as follows.

1.5.1 Historical Notes

We distinguish two parallel directions of investigations, in the areas of mathematical statistics and automatic control theory, and then summarize investigations concerning the possible merging of these two directions.

1.5.1.1 Mathematical Statistics

Interest in on-line change detection probably arose first in the area of quality control, where *control charts* were introduced in [Shewhart, 1931] and then *cumulative sums charts* in [Page, 1954a]. The two main classes of statistical problem statements are the Bayesian and the non-Bayesian approaches.

Bayesian approach The first Bayesian change detection problem was stated in [Girshick and Rubin, 1952] to solve a typical on-line quality control problem for continuous technological processes. The first optimality results concerning Bayesian change detection algorithms were obtained in [Shiryaev, 1961, Shiryaev, 1963, Shiryaev, 1965]. Since then, the literature in this area has become quite wide. More recent investigations can be found in [Pollak and Siegmund, 1985, Pollak, 1985, Pollak, 1987].

Non-Bayesian approach The first investigation of non-Bayesian change detection algorithms was made in [Page, 1954a]. The asymptotic optimality of cumulative sum algorithms was proved in [Lorden, 1971]. Nonasymptotic optimality results can be found in [Moustakides, 1986, Ritov, 1990]. The extension of such techniques to composite hypotheses testing problems is discussed in [Lorden, 1971, Lorden, 1973, Pollak and Siegmund, 1975]. The generalization of Lorden's results to dependent processes is discussed in [Bansal and Papantoni-Kazakos, 1986].

1.5.1.2 Automatic Control

In the area of automatic control, change detection problems are referred to as model-based fault detection and isolation (FDI). The concept of analytical redundancy for fault detection was investigated approximately independently at the same time in the United States [Beard, 1971] and in the Soviet Union [Britov and Mironovski, 1972]. Further key developments concerning the geometrical aspects can be found in [E.Chow and Willsky, 1984, Lou *et al.*, 1986, Massoumnia, 1986, White and Speyer, 1987, Viswanadham *et al.*, 1987a, Wünnenberg, 1990] and are discussed in the survey papers [Willsky, 1976, Frank, 1990, Patton and Chen, 1991, Gertler, 1991]. Typically, the models used in these investigations are more complex than the models classically used in the mathematical statistics literature.

From a formal point of view, this research direction does not belong to the theory of change detection, because of the lack of statistical problem statements and criteria. Nevertheless, the main ideas underlying fault detection tools, namely the use of innovations or residuals for monitoring purposes, are very close to the concept of sufficient statistics for detection. For this reason, we think it useful to discuss these two types of concepts together. A first attempt to bring together both geometric concepts of analytical redundancy and statistical decision tools is the survey paper [Willsky, 1976].

1.5.1.3 Joint Approach

In the early 1970s, a new research direction arose, involving complex statistical models (much more complex than in classical statistical investigations). The main motivation for these new developments were unsuccessful attempts at using pure mathematical tools for solving concrete problems in the automatization of industrial processes. The starting point of these new investigations was the use of change detection decision rules for the more complex models, and the extension of the available theoretical results existing about them

[Lumel'sky, 1972, Nikiforov, 1975, Bagshaw and R.Johnson, 1977, Nikiforov, 1978, Nikiforov, 1980, Segen and Sanderson, 1980, Basseville and Benveniste, 1983a, Basseville and Benveniste, 1983b, Vorobeichikov and Konev, 1988]. Survey papers reporting these investigations are [Basseville, 1982, Kligiene and Telksnys, 1983, Basseville, 1988].

1.5.1.4 Investigations in Application Domains

The problem of detecting abrupt changes in properties of signals and dynamic systems has received increasing attention in the last twenty years. One key reason for that is its connection to the problem of fault detection, and strong industrial needs in the area of condition-based maintenance and monitoring of plants. Another reason is its usefulness in time-series analysis and signal processing for recognition purposes.

Several books related to quality control exist, such as [Shewhart, 1931, Woodward and Goldsmith, 1964, Van Dobben De Bruyn, 1968, Duncan, 1986]. The analysis of biomedical signals, especially electroencephalograms, is another field where many contributions to the problem of automatic segmentation of signals have been made [R.Jones *et al.*, 1970, Borodkin and Mottl', 1976, Mathieu, 1976, Segen and Sanderson, 1980]. The interest in the change detection methodology in this area is reflected in [Cohen, 1987], where segmentation algorithms are presented as basic signal processing tools. Geophysical signal processing can also be achieved with the aid of segmentation algorithms; for example, diagraphy [Basseville and Benveniste, 1983a] and seismology [Nikiforov and Tikhonov, 1986, Nikiforov *et al.*, 1989]. Automatic segmentation was introduced as a first step toward continuous speech recognition in [André-Obrecht, 1988], and as a first step toward speech coding in [Di Francesco, 1990], both using the algorithm presented in [Basseville and Benveniste, 1983b].

Interest in the change detection methodology also arose in chemical engineering [Himmelblau, 1978]. In the field of econometry, two books are devoted to the problem of structural change detection, i.e., the problem of detection of changes in the parameters of an econometric model. These are [Poirier, 1976, Broemeling and Tsurumi, 1987]. An annotated bibliography can also be found in [Shaban, 1980].

Many other application domains have been investigated, as can be seen from the long list of application studies of innovation-based fault detection/diagnosis methods in [Patton *et al.*, 1989] and [Tzafestas *et al.*, 1987].

1.5.1.5 Related Investigations

The most closely related investigations concern the *off-line* change detection and estimation problems. The historical starting point of these studies is [Page, 1957]. Subsequent investigations are in [Hinkley, 1970, Hinkley, 1971, Kligiene and Telksnys, 1983]. More generally, complete theoretical optimality results about the likelihood approach to change detection are obtained in [Deshayes and Picard, 1979, Deshayes and Picard, 1983].

1.5.2 Seminars, Surveys, and Books

Two national seminars on change detection were organized in 1984 independently in Paris, France, and in Palanga, USSR, emphasizing great interest and activity in this field in both countries. The contents of these seminars are presented in [Basseville and Benveniste, 1986, Telksnys, 1987]. Two subsequent seminars took place in Moscow, USSR, and in Voronej, USSR, in 1988 and 1990, respectively. Many international conferences and workshops in the area of automatic control have had sessions on fault detection and isolation over the last fifteen years.

Many survey papers about this problem have been published over the past twenty years, for example, four survey papers in *Automatica* [Willsky, 1976, Isermann, 1984, Basseville, 1988, Frank, 1990] and two in

Automation and Remote Control [Mironovski, 1980, Kligiene and Telksnys, 1983]; one survey about sensor failure detection in jet engines [Merril, 1985]; and three survey papers in the econometry literature [Zacks, 1983, Krishnaiah and Miao, 1988, Csörgö and Horváth, 1988].

We now list some of the books in this area. The topic of statistical tools for change detection is investigated in [Woodward and Goldsmith, 1964, Van Dobben De Bruyn, 1968, Shiryaev, 1978, Nikiforov, 1983, Basseville and Benveniste, 1986, Siegmund, 1985b, Telksnys, 1987, Zhigljavsky and Kraskovsky, 1988, Brodskiy and Darkhovskiy, 1992]. Books oriented more toward geometrical tools are [Tzafestas *et al*., 1987, Singh *et al*., 1987, Patton *et al*., 1989]. The book [Viswanadham *et al*., 1987b] basically put together, but without integration, reliability theory and fault-tolerant computer systems on one hand and fault detection and diagnosis on the other hand. More specific books are [Himmelblau, 1978, Pau, 1981].

Part I

Changes in the Scalar Parameter of an Independent Sequence

2

Change Detection Algorithms

In this chapter, we describe the simplest change detection algorithms. We consider a sequence of *independent* random variables $(y_k)_k$ with a probability density $p_\theta(y)$ depending upon only one *scalar* parameter. Before the *unknown change time* t_0, the parameter θ is equal to θ_0, and after the change it is equal to $\theta_1 \neq \theta_0$. The problems are then to detect and estimate this change in the parameter.

The main **goal** of this chapter is to introduce the reader to the design of *on-line* change detection algorithms, basically assuming that the parameter θ_0 before change is *known*. We start from elementary algorithms originally derived using an intuitive point of view, and continue with conceptually more involved but practically not more complex algorithms. In some cases, we give several possible derivations of the same algorithm. But the key point is that we introduce these algorithms within a general statistical framework, based upon likelihood techniques, which will be used throughout the book. Our conviction is that the early introduction of such a general approach in a simple case will help the reader to draw up a unified mental picture of change detection algorithms in more complex cases. In the present chapter, using this general approach and for this simplest case, we describe several on-line algorithms of increasing complexity. We also discuss the *off-line* point of view more briefly. The main example, which is carried through this chapter, is concerned with the detection of a change in the mean of an independent Gaussian sequence.

The **tools** for reaching this goal are as follows. First, our description of all the algorithms of this chapter is based on a concept that is very important in mathematical statistics, namely the logarithm of the likelihood ratio, defined by

$$s(y) = \ln \frac{p_{\theta_1}(y)}{p_{\theta_0}(y)} \tag{2.0.1}$$

and referred to as the log-likelihood ratio. The key statistical property of this ratio is as follows : Let $\mathbf{E}_{\theta_0}$ and $\mathbf{E}_{\theta_1}$ denote the expectations of the random variables under the two distributions p_{θ_0} and p_{θ_1}, respectively. Then,

$$\mathbf{E}_{\theta_0}(s) < 0 \ \text{ and } \ \mathbf{E}_{\theta_1}(s) > 0 \tag{2.0.2}$$

In other words, *a change in the parameter θ is reflected as a change in the sign of the mean value of the log-likelihood ratio*. This property can be viewed as a kind of detectability of the change. Because the Kullback information $\mathbf{K}$ is defined by $\mathbf{K}(\theta_1, \theta_0) = \mathbf{E}_{\theta_1}(s)$, we also have that the difference between the two mean values is

$$\mathbf{E}_{\theta_1}(s) - \mathbf{E}_{\theta_0}(s) = \mathbf{K}(\theta_1, \theta_0) + \mathbf{K}(\theta_0, \theta_1) > 0 \tag{2.0.3}$$

From this, we deduce that the detectability of a change can also be defined with the aid of the Kullback information between the two models before and after change. These concepts are used throughout the book.

Second, even for this simple case, it is of interest to classify all possible practical problem statements with respect to two different issues :

- The first possible classification is with respect to assumptions about the unknown change time t_0. In some applications, it is useful to consider t_0 as a nonrandom unknown value, or a random unknown value with unknown distribution. In other words, we deal with a nonparametric approach as far as this change time t_0 is concerned. This assumption is useful because very often in practice, either it is very difficult to have *a priori* information about the distribution of the change times, or this distribution is nonstationary. This point of view is taken in sections 2.1, 2.2, and 2.4 for on-line algorithms and in section 2.6 for off-line algorithms. In some applications, it is possible to use *a priori* information about the distribution of the change time, taking a Bayesian point of view. Such *a priori* information can be available from life-time estimations made in reliability investigations. This point of view is used in section 2.3.
- The second possible classification of algorithms is with respect to the available information about the value θ_1 of the parameter after change, as we discussed in section 1.4. We first consider that this value is known : This is the case of sections 2.1, 2.2, and 2.3. The case of unknown value for θ_1 is investigated in section 2.4 for on-line algorithms and in section 2.6 for off-line algorithms.

Before proceeding, let us add one comment concerning the performances of these algorithms and the detectability of a given change. The criteria for the performance evaluation of these algorithms were introduced in section 1.4 from an intuitive point of view. The performances of the *on-line* algorithms presented in the present chapter are investigated in detail in chapter 5 with the aid of the formal definition of these criteria, given in section 4.4. These performance evaluations can be computationally complex, even in the present simple case. For this reason, it is also of interest to consider a kind of weak performance index, the positivity of which simply states the detectability of a change (with no more indication on the properties of the detection). The Kullback information is a good candidate for such a weak index, both because of the above-mentioned inequalities and because, as shown in chapter 4, it is an adequate index of separability between two probability measures. This mutual information is zero only when the parameters are equal, and can be shown to be an increasing function of the Euclidean distance between the parameters θ_0 and θ_1 when this distance is small. This detectability definition is investigated in detail in more complex cases in chapters 7, 8, and 9.

2.1 Elementary Algorithms

In this section, we describe several simple and well-known algorithms. Most of the algorithms presented here work on samples of data with *fixed* size; only one uses a growing memory. In the next section, we deal basically with a random-size sliding window algorithm. In quality control, these elementary algorithms are usually called *Shewhart control charts* and finite or infinite *moving average control charts*. We also introduce another elementary algorithm, called a *filtered derivative* algorithm, which is often used in image edge detection. We place these algorithms in our general likelihood framework, and consider the case in which the only unknown value is the change time t_0. Recall that all the key mathematical concepts are described in chapters 3 and 4.

2.1.1 Limit Checking Detectors and Shewhart Control Charts

Let us first introduce the initial idea used in quality control under the name of continuous inspection. Samples with fixed size N are taken, and at the end of each sample a decision rule is computed to test between

the two following hypotheses about the parameter θ :

$$\begin{aligned} \mathbf{H}_0 &: \theta = \theta_0 \\ \mathbf{H}_1 &: \theta = \theta_1 \end{aligned} \tag{2.1.1}$$

As long as the decision is taken in favour of $\mathbf{H}_0$, the sampling and test continue. Sampling is stopped after the first sample of observations for which the decision is taken in favor of $\mathbf{H}_1$.

We introduce the following notation, which is used throughout this and the subsequent chapters. Let

$$\begin{aligned} S_j^k &= \sum_{i=j}^{k} s_i \\ s_i &= \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \end{aligned} \tag{2.1.2}$$

be the log-likelihood ratio for the observations from y_j to y_k. We refer to s_i as the *sufficient statistic* for reasons that are explained in section 4.1.

The following statement is a direct consequence of the Neyman-Pearson lemma, which we recall in chapter 4. For a fixed sample size N, the optimal decision rule d is given by

$$d = \begin{cases} 0 & \text{if} \quad S_1^N < h; \quad \mathbf{H}_0 \text{ is chosen} \\ 1 & \text{if} \quad S_1^N \geq h; \quad \mathbf{H}_1 \text{ is chosen} \end{cases} \tag{2.1.3}$$

where h is a conveniently chosen threshold. The sum S_1^N is said to be the *decision function*. The decision is taken with the aid of what is called a stopping rule, which in this case is defined by

$$t_a = N \cdot \min\{K : d_K = 1\} \tag{2.1.4}$$

where d_K is the decision rule for the sample number K (of size N) and t_a is the *alarm time*. In other words, the observation is stopped after the first sample of size N for which the decision is in favor of $\mathbf{H}_1$.

Example 2.1.1 (Change in mean). *Let us now consider the particular case where the distribution is Gaussian with mean value μ and constant variance σ^2. In this case, the changing parameter θ is μ. The probability density is*

$$p_\theta(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \tag{2.1.5}$$

and the sufficient statistic s_i is

$$s_i = \frac{\mu_1 - \mu_0}{\sigma^2} \left(y_i - \frac{\mu_0 + \mu_1}{2} \right) \tag{2.1.6}$$

which we shall write as

$$\begin{aligned} s_i &= \frac{b}{\sigma} \left(y_i - \frac{\mu_0 + \mu_1}{2} \right) \\ &= \frac{b}{\sigma} \left(y_i - \mu_0 - \frac{\nu}{2} \right) \end{aligned} \tag{2.1.7}$$

where

$$\nu = \mu_1 - \mu_0 \tag{2.1.8}$$

is the change magnitude *and*

$$b = \frac{\mu_1 - \mu_0}{\sigma} \tag{2.1.9}$$

is the signal-to-noise ratio. *Therefore, the decision function (2.1.2) is*

$$S_1^N = \frac{b}{\sigma} \sum_{i=1}^{N} \left(y_i - \mu_0 - \frac{\nu}{2} \right) \tag{2.1.10}$$

The stopping rule for the change detection algorithm is as in (2.1.4), with the decision rule defined by

$$d = \begin{cases} 0 & \text{if} \quad S_1^N(K) < h \\ 1 & \text{if} \quad S_1^N(K) \geq h \end{cases} \tag{2.1.11}$$

where

$$S_1^N(K) = S_{N(K-1)+1}^{NK} \tag{2.1.12}$$

with S_i^j defined in (2.1.2). This change detection algorithm is one of the oldest and most well-known algorithms for continuous inspection, and is called Shewhart control chart *[Shewhart, 1931]. For this control chart, when $\mu_1 > \mu_0$, the alarm is set the first time at which*

$$\bar{y}(K) \geq \mu_0 + \kappa \frac{\sigma}{\sqrt{N}} \tag{2.1.13}$$

where

$$\bar{y}(K) = \frac{1}{N} \sum_{i=N(K-1)+1}^{NK} y_i \tag{2.1.14}$$

Note that the threshold is related to the standard deviation of the left side of this inequality. This stopping rule is standard in quality control, where the name for the right side of this inequality is the upper control limit. *The tuning parameters of this Shewhart control chart are κ and N. The behavior of this chart, when applied to the signal of figure 1.1, is depicted in figure 2.1.*

It is often more useful to detect deviations from μ_0 in both directions, namely increases and decreases. In this case, assume that the mean value after the change is either $\mu_1^+ = \mu_0 + \nu$ or $\mu_1^- = \mu_0 - \nu$. Then the alarm is set the first time at which

$$|\bar{y}(K) - \mu_0| \geq \kappa \frac{\sigma}{\sqrt{N}} \tag{2.1.15}$$

where $\mu_0 - \kappa \frac{\sigma}{\sqrt{N}}$ is the lower control limit. *This is depicted in the figure 2.2. The tuning parameters of this algorithm are κ and N again. The optimal tuning of these parameters can be obtained with the aid of an a priori information concerning the change magnitude ν.*

Let us add one comment about a slightly different use of control charts [S.Roberts, 1966]. To prevent false alarms and to obtain more reliable detection results, the intuitive idea consists of deciding a change when a preassigned number of crossings in (2.1.15) occur among several successive data samples of size N. This idea is known as a *run test* in quality control, and sometimes as a *counter* in the engineering literature. Various types of run tests have been used to supplement Shewhart control charts, as explained in [S.Roberts, 1966]. A similar idea is also used for another change detection algorithm in subsection 2.1.4.

2.1.2 Geometric Moving Average Control Charts

Two key ideas underlie the geometric moving average (GMA) algorithm. The first idea is related to the above-mentioned behavior of the log-likelihood ratio (2.0.1). The second deals with the widespread intuitive idea of exponential weighting of observations. As usual in nonstationary situations, because of the unknown

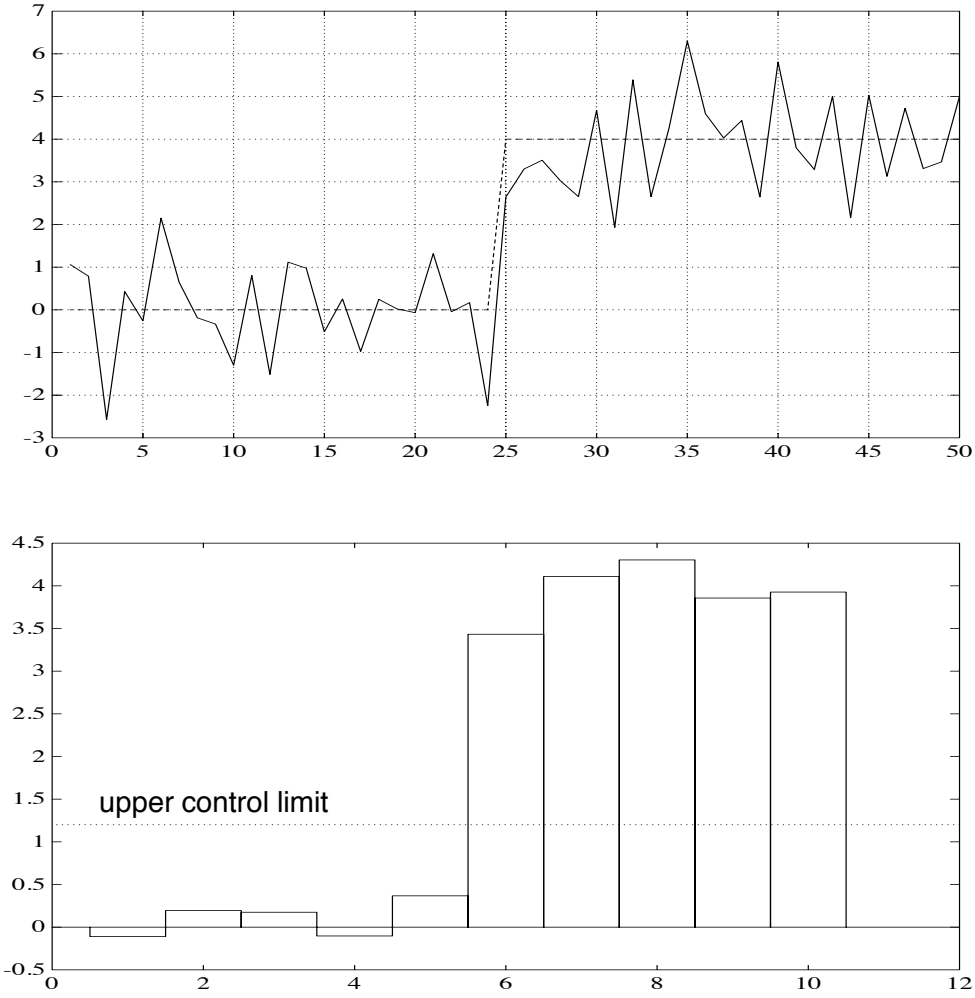

Figure 2.1 A Shewhart control chart corresponding to a change in the mean of a Gaussian sequence with constant variance.

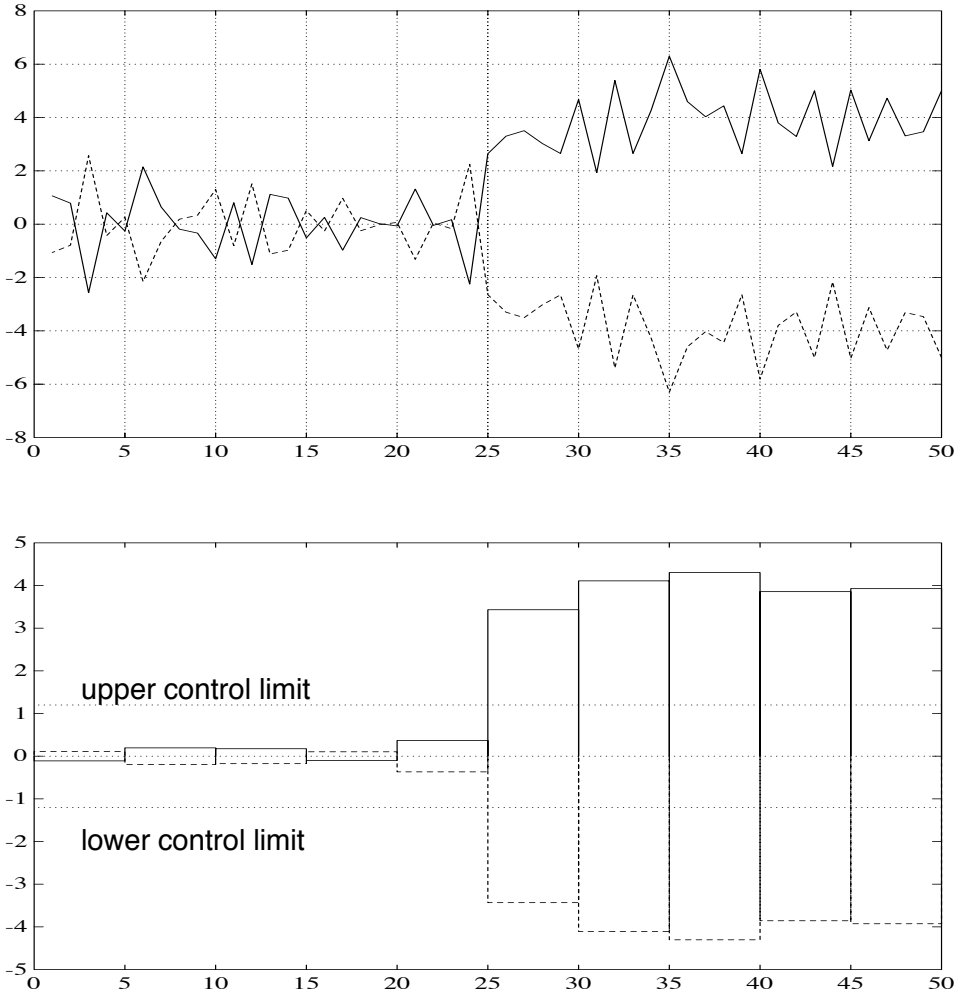

Figure 2.2 A two-sided Shewhart control chart.

change time t_0, it is of interest to use higher weights on recent observations and lower weights on past ones. Therefore, the following decision function is relevant [S.Roberts, 1959, Hines, 1976a, Hines, 1976b] :

$$\begin{aligned} g_k &= \sum_{i=0}^{\infty} \gamma_i \ln \frac{p_{\theta_1}(y_{k-i})}{p_{\theta_0}(y_{k-i})} \\ &= \sum_{i=0}^{\infty} \gamma_i s_{k-i} \end{aligned} \tag{2.1.16}$$

where the weights γ_i are exponential, namely

$$\gamma_i = \alpha(1-\alpha)^i, 0 < \alpha \leq 1 \tag{2.1.17}$$

The coefficient α acts as a forgetting factor. This decision function can be rewritten in a recursive manner as

$$g_k = (1-\alpha)\ g_{k-1} + \alpha\ s_k, \ \text{ with: } \ g_0 = 0 \tag{2.1.18}$$

The alarm time is defined by the following stopping rule :

$$t_a = \min\{k : g_k \geq h\} \tag{2.1.19}$$

where h is a conveniently chosen threshold.

Example 2.1.2 (Change in mean - contd.). *In the case of a change in the mean of an independent Gaussian sequence, s_k is given by (2.1.6), and the GMA decision function is*

$$\tilde{g}_k = (1-\alpha)\ \tilde{g}_{k-1} + \alpha\ (y_k - \mu_0), \ \textit{ with: } \ \tilde{g}_0 = 0 \tag{2.1.20}$$

where $\tilde{g}$ and g are related through

$$\tilde{g}_k = \frac{\sigma^2}{\mu_1 - \mu_0}\ g_k - \frac{\mu_1 - \mu_0}{2} \tag{2.1.21}$$

The behavior of this decision function, when applied to the signal of figure 1.1, is depicted in figure 2.3. In the corresponding two-sided situation, the stopping rule is

$$t_a = \min\{k : |\tilde{g}_k| \geq h\} \tag{2.1.22}$$

Example 2.1.3 (Change in variance). *In the case of a change in the variance σ^2, which is relevant in quality control, as explained in example 1.2.1, we have*

$$s_k = \ln \frac{\sigma_0}{\sigma_1} + \left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right) \frac{(y_k - \mu)^2}{2} \tag{2.1.23}$$

Therefore, the relevant decision function can be written as

$$\tilde{g}_k = \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2}\ g_k - \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2} \ln \frac{\sigma_0}{\sigma_1} \tag{2.1.24}$$

where g_k is defined in (2.1.18). In a recursive form, this becomes

$$\tilde{g}_k = (1-\alpha)\ \tilde{g}_{k-1} + \alpha\ (y_k - \mu)^2, \ \textit{ with: } \ \tilde{g}_0 = 0 \tag{2.1.25}$$

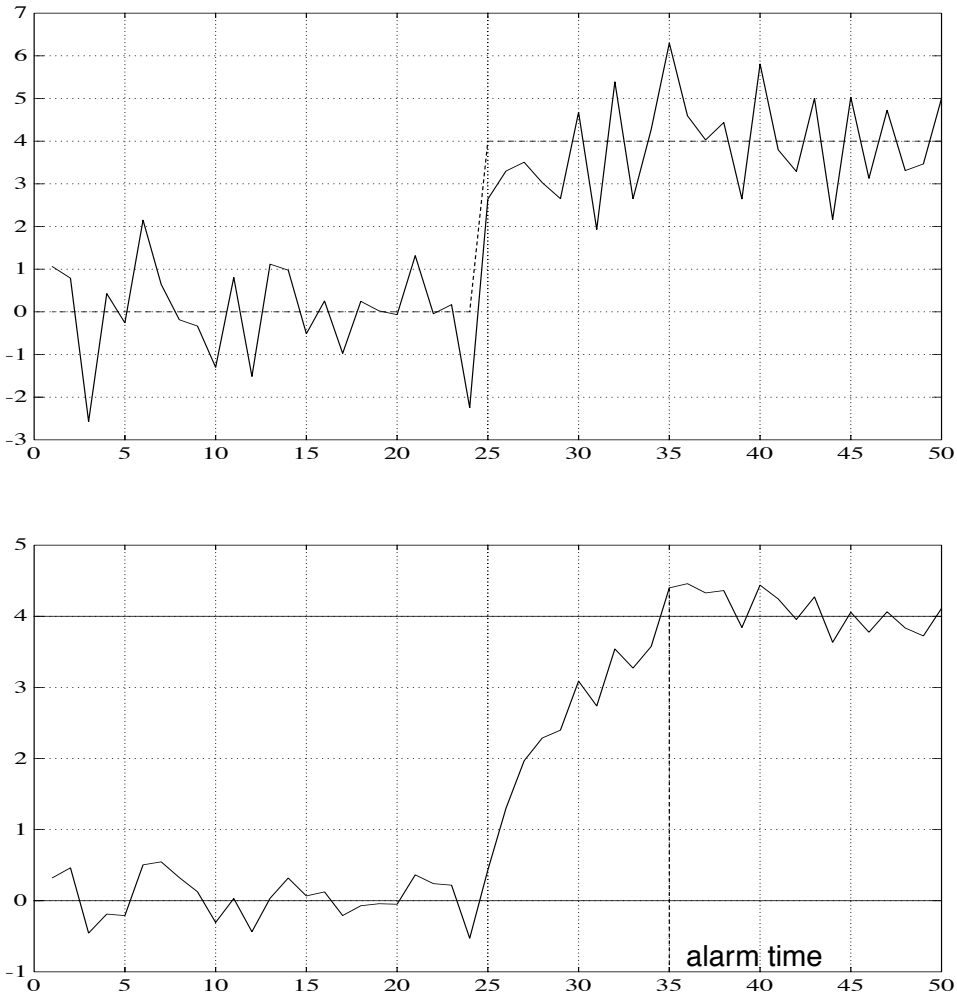

Figure 2.3 A geometric moving average algorithm corresponding to a change in the mean of a Gaussian sequence with constant variance.

2.1.3 Finite Moving Average Control Charts

A similar idea to the previous control charts consists in replacing the exponential forgetting operation by a finite memory one, and thus in using a finite set of weights, which are no longer assumed to form a geometric sequence. For defining this new detector, which is called finite moving average (FMA) algorithm, let us follow the derivation of the geometric moving average control charts. First, consider the following variant of the causal filtering (2.1.16) used in these charts :

$$g_k = \sum_{i=0}^{N-1} \gamma_i \ln \frac{p_{\theta_1}(y_{k-i})}{p_{\theta_0}(y_{k-i})} \tag{2.1.26}$$

where the weights γ_i are any weights for causal filters. The stopping rule is as in the previous control chart :

$$t_a = \min\{k : g_k \geq h\} \tag{2.1.27}$$

Example 2.1.4 (Change in mean - contd.). *In the case of an* increase *in the mean, this stopping rule can be computed as follows. Using (2.1.6), the decision function g_k in (2.1.26) can be expressed as*

$$g_k = \sum_{i=0}^{N-1} \gamma_i (y_{k-i} - \mu_0) \tag{2.1.28}$$

In the two-sided case, g_k is the same, and the stopping rule is

$$t_a = \min\{k : |g_k| \geq h\} \tag{2.1.29}$$

2.1.4 Filtered Derivative Algorithms

In the case of a change in the mean of a Gaussian sequence, the filtered derivative algorithms are based on the following very intuitive idea. Ideally, that is, in a no noise situation, a change in the mean level of a sequence of observations is locally characterized by a great absolute value of the (discrete) derivative of the sample observations. Because the derivative operator acts in a very poor manner as soon as noise is present in observations, a more realistic detector should use a filtering operation before derivation. This explains the title of this subsection. The typical behavior of this algorithm is depicted in figure 2.4 for the ideal and realistic situations. Now, because of the smoothing operation on the jump, several alarms are to occur in the neighborhood of t_0. An elementary way to increase the robustness of this detector is to count the number of threshold crossings during a fixed time interval before deciding the change actually occurred.

Let us now put this intuition-based detector into our more formal framework for change detection algorithms. We use again the derivation of the finite moving average control charts :

$$g_k = \sum_{i=0}^{N-1} \gamma_i \ln \frac{p_{\theta_1}(y_{k-i})}{p_{\theta_0}(y_{k-i})} \tag{2.1.30}$$

where the weights γ_i are again any weights for causal filters, and we consider the discrete derivative of g_k :

$$\nabla g_k = g_k - g_{k-1} \tag{2.1.31}$$

and the following stopping rule :

$$t_a = \min\{k : \sum_{i=0}^{N-1} \mathbf{1}_{\{\nabla g_{k-i} \geq h\}} \geq \eta\} \tag{2.1.32}$$

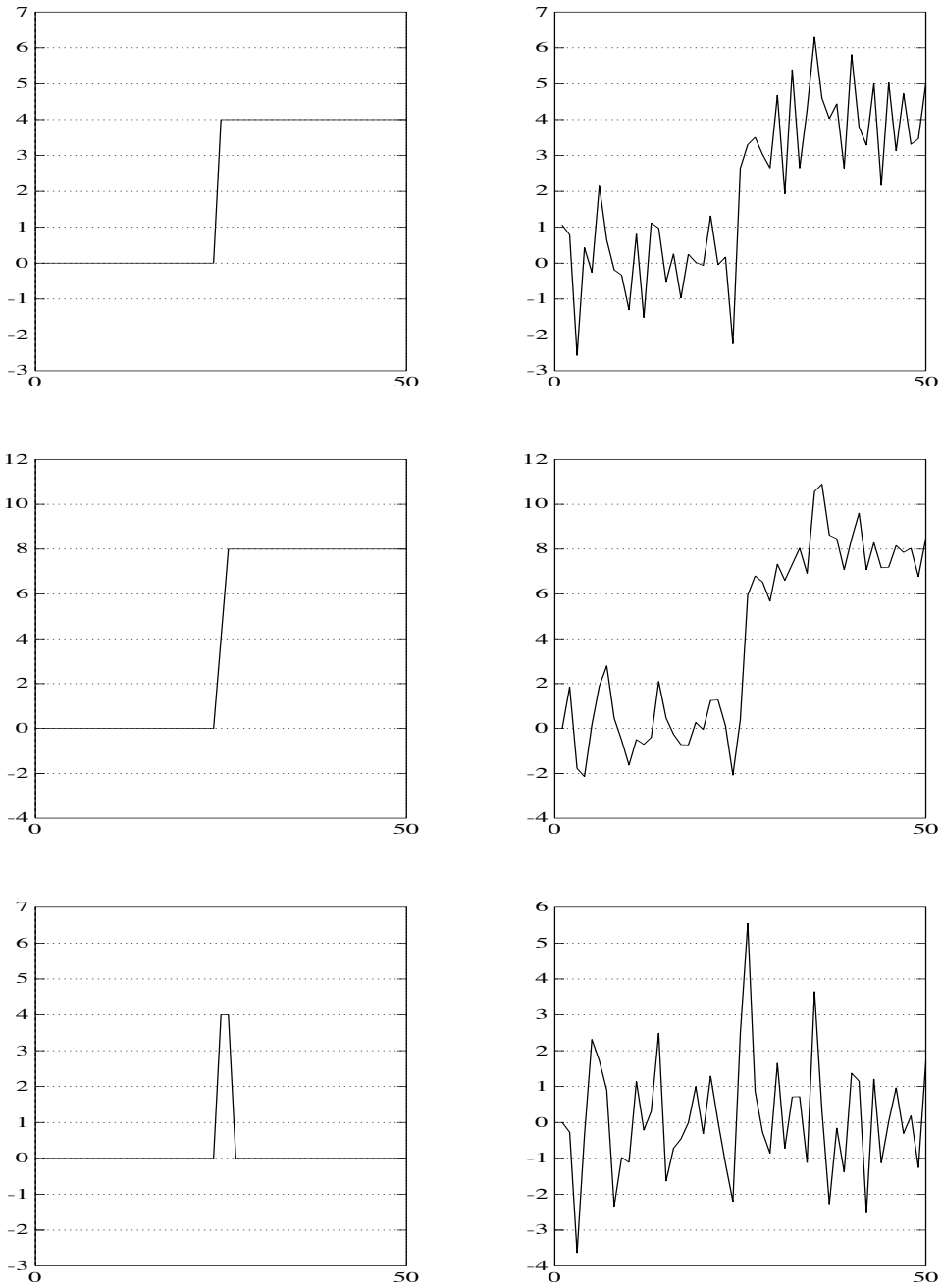

Figure 2.4 Ideal (left) and realistic (right) behaviors of a filtered derivative algorithm corresponding to a change in the mean of a Gaussian sequence with constant variance : signal (first row), filtered signal (second row), and filtered and derivate signal (third row).

where $\mathbf{1}_{\{x\}}$ is the indicator of event $\{x\}$. In this formula, h is the threshold for the derivative, and η is a threshold for the number of crossings of h. This threshold η is used for decreasing the number of alarms in the neighborhood of the change due to the smoothing operation. It turns out that, in practice, $\eta = 2$ is often a convenient value for achieving this goal.

Example 2.1.5 (Change in mean - contd.). *In the case of an* increase *in the mean, the decision function g_k corresponding to (2.1.30) can again be taken as*

$$g_k = \sum_{i=0}^{N-1} \gamma_i (y_{k-i} - \mu_0) \tag{2.1.33}$$

The stopping rule is as in (2.1.32). In the two-sided case of jump in mean in an unknown direction, the stopping rule is

$$t_a = \min\{k : \sum_{i=0}^{N-1} \mathbf{1}_{\{|\nabla g_{k-i}| \geq h\}} \geq \eta\} \tag{2.1.34}$$

Two elementary choices of smoothing filters in (2.1.30) are as follows :

- *An integrating filter with N constant unit weights γ_i, which results in*

$$\nabla g_k = y_k - y_{k-N}$$

- *A triangular filter with impulse response of triangular form, namely $\gamma_{p+i} = \gamma_{p-i} = i$ for $0 \leq i \leq p$, where $N - 1 = 2p$, which results in*

$$\nabla g_k = \sum_{i=0}^{p-1} y_{k-i} - \sum_{i=p}^{2p-1} y_{k-i}$$

In other words, the corresponding stopping rules are based upon the difference between either sample values or local averages of sample values.

2.2 CUSUM Algorithm

We now introduce the cumulative sum (CUSUM) algorithm, which was first proposed in [Page, 1954a]. We describe four different derivations. The first is more intuition-based, and uses ideas connected to a simple integration of signals with *adaptive threshold*. The second derivation is based on a more formal on-line statistical approach, similar to the approach used before for introducing control charts, and based upon a *repeated use of the sequential probability ratio test*. The third derivation comes from the use of the off-line point of view for a *multiple hypotheses testing* approach. This derivation is useful for the introduction of the geometrical interpretation of the CUSUM algorithm with the aid of a V-mask. The fourth derivation is based upon the concept of open-ended tests.

2.2.1 Intuitive Derivation

As we mentioned in the previous section, the typical behavior of the log-likelihood ratio S_k shows a negative drift before change, and a positive drift after change, as depicted in figure 2.5, again for the signal of figure 1.1. Therefore, the relevant information, as far as the change is concerned, lies in the difference

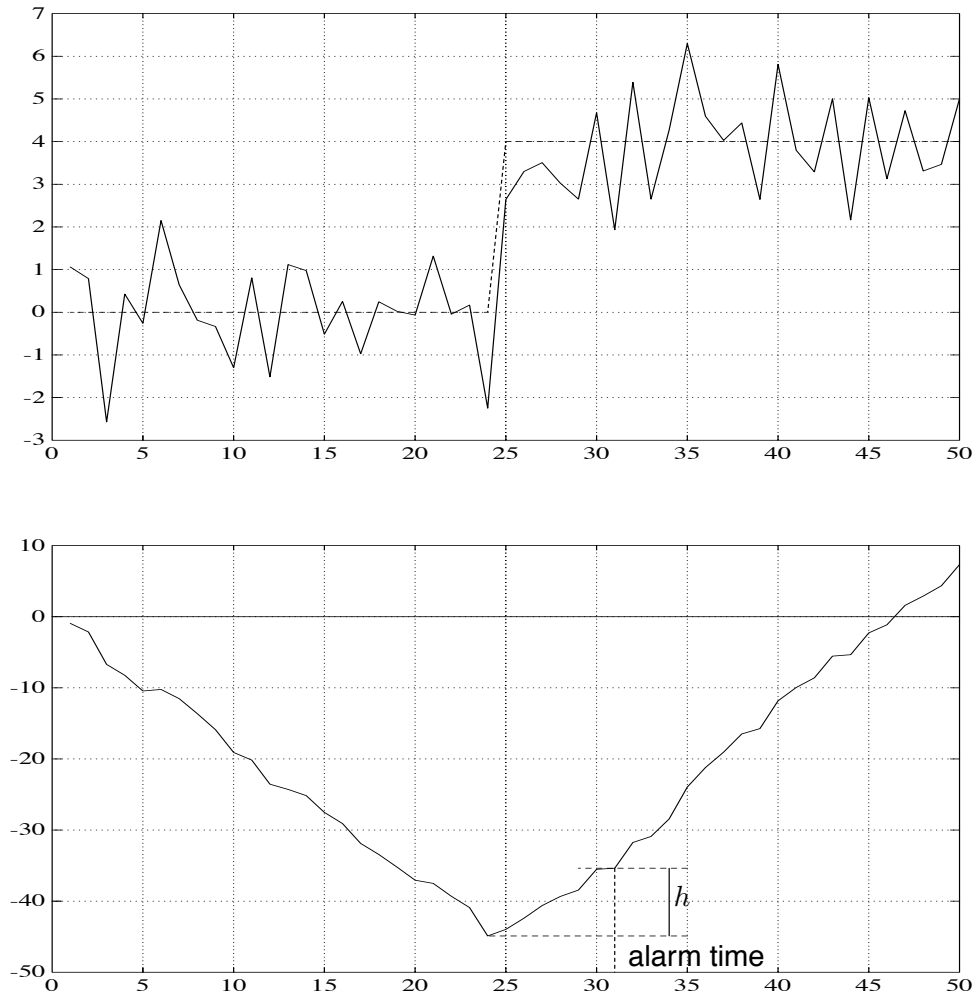

Figure 2.5 Typical behavior of the log-likelihood ratio S_k corresponding to a change in the mean of a Gaussian sequence with constant variance : negative drift before and positive drift after the change.

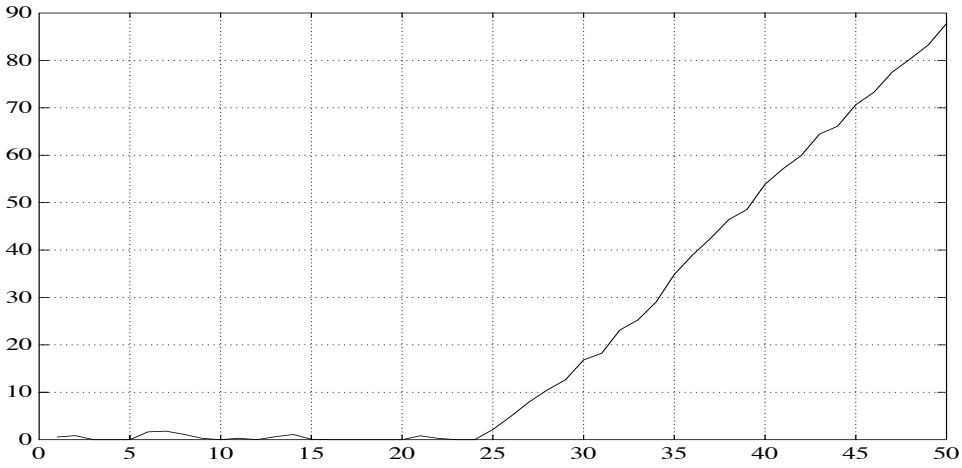

Figure 2.6 Typical behavior of the CUSUM decision function g_k.

between the value of the log-likelihood ratio and its current minimum value; and the corresponding decision rule is then, at each time instant, to compare this difference to a threshold as follows :

$$g_k = S_k - m_k \geq h \tag{2.2.1}$$

where

$$\begin{aligned} S_k &= \sum_{i=1}^{k} s_i \\ s_i &= \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \\ m_k &= \min_{1 \leq j \leq k} S_j \end{aligned} \tag{2.2.2}$$

The typical behavior of g_k is depicted in figure 2.6. The stopping time is

$$t_a = \min\{k : g_k \geq h\} \tag{2.2.3}$$

which can be obviously rewritten as

$$t_a = \min\{k : S_k \geq m_k + h\} \tag{2.2.4}$$

Now it becomes clear that this detection rule is nothing but a comparison between the cumulative sum S_k and an *adaptive threshold* $m_k + h$. Because of m_k, this threshold not only is modified on-line, but also keeps *complete* memory of the entire information contained in the past observations. Moreover, it is obvious from (2.1.6) that, in the case of change in the mean of a Gaussian sequence, S_k is a standard *integration* of the observations.

2.2.2 CUSUM Algorithm as a Repeated Sequential Probability Ratio Test

Page suggested the use of repeated testing of the two simple hypotheses :

$$\mathbf{H}_0 \quad : \quad \theta = \theta_0 \tag{2.2.5}$$

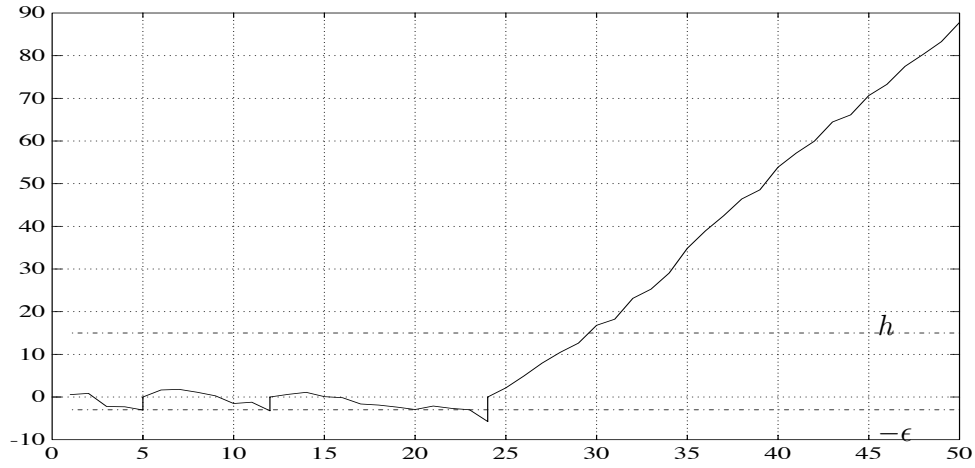

Figure 2.7 Repeated use of SPRT. $T_i = 5, 12, 24$, and 30 are the stopping times in each successive cycle, and $d_i = 0, 0, 0$, and 1 are the corresponding decision rules.

$$\mathbf{H}_1 \quad : \quad \theta = \theta_1$$

with the aid of the *sequential probability ratio test (SPRT)*. Let us first define a single use of the SPRT algorithm. The SPRT is defined with the aid of the pair (d, T) where d is the decision rule and T is a stopping time, exactly as the Neyman-Pearson rule is defined with the aid of the decision rule d. The stopping time T is the time at which the final decision is taken and thus at which observation is stopped. The definition of the SPRT is thus

$$d = \begin{cases} 0 & \text{if} \quad S_1^T \leq -\epsilon \\ 1 & \text{if} \quad S_1^T \geq h \end{cases} \tag{2.2.6}$$

where T is the exit time :

$$T = T_{-\epsilon,h} = \min\{k : (S_1^k \geq h) \cup (S_1^k \leq -\epsilon)\} \tag{2.2.7}$$

where $\epsilon \geq 0$ and $h > 0$ are conveniently chosen thresholds. Now, as in section 2.1, we use repeated SPRT until the decision $d = 1$ is taken. The typical behavior of this repeated use of the SPRT is depicted in figure 2.7, where $T_i = 5, 12, 24$, and 30 are the stopping times in each successive cycle, and $d_i = 0, 0, 0$, and 1 are the corresponding decision rules. The key idea of Page was to *restart the SPRT algorithm as long as the previously taken decision is* $d = 0$. The first time at which $d = 1$, we stop observation and do not restart a new cycle of the SPRT. This time is then the *alarm time* at which the change is detected.

Using an intuitive motivation, Page suggested that the optimal value of the lower threshold ϵ should be zero. This statement was formally proven later [Shiryaev, 1961, Lorden, 1971, Moustakides, 1986, Ritov, 1990] and is discussed in section 5.2. Starting from the repeated SPRT with this value of lower threshold, the resulting decision rule can be rewritten in a recursive manner as

$$g_k = \begin{cases} g_{k-1} + \ln \frac{p_{\theta_1}(y_k)}{p_{\theta_0}(y_k)} & \text{if} \quad g_{k-1} + \ln \frac{p_{\theta_1}(y_k)}{p_{\theta_0}(y_k)} > 0 \\ 0 & \text{if} \quad g_{k-1} + \ln \frac{p_{\theta_1}(y_k)}{p_{\theta_0}(y_k)} \leq 0 \end{cases} \tag{2.2.8}$$

where $g_0 = 0$. Remembering the definition of s_k in (2.1.2), this can be compacted into

$$g_k = (g_{k-1} + s_k)^+ \tag{2.2.9}$$

where $(x)^+ = \sup(0, x)$. Finally, the stopping rule and alarm time are defined by

$$t_a = \min\{k : g_k \geq h\} \tag{2.2.10}$$

where g_k is given in (2.2.9). The typical behavior of this decision function is depicted in figure 2.6. It is easy to prove that this form of decision rule is equivalent to the other form that we presented in (2.2.4). On the other hand, it can also be written as

$$g_k = \left(S^k_{k-N_k+1}\right)^+ \tag{2.2.11}$$

where

$$N_k = N_{k-1} \cdot \mathbf{1}_{\{g_{k-1}>0\}} + 1 \tag{2.2.12}$$

$\mathbf{1}_{\{x\}}$ is the indicator of event x, and t_a is defined in (2.2.10). In this formula, N_k is the number of observations after re-start of the SPRT. The formula (2.2.11) can be interpreted as an integration of the observations over a *sliding window with random size*. This size is chosen according to the behavior of the entire past observations.

2.2.3 Off-line Statistical Derivation

As we discussed in chapter 1, when taking an off-line point of view, it is convenient to introduce the following hypotheses about the observations $y_1, ..., y_k$:

$$\begin{array}{rrlll} & \mathbf{H}_0 : & \theta = \theta_0 & \text{for} & 1 \leq i \leq k \\ \text{for } 1 \leq j \leq k, & \mathbf{H}_j : & \theta = \theta_0 & \text{for} & 1 \leq i \leq j-1 \\ & & \theta = \theta_1 & \text{for} & j \leq i \leq k \end{array} \tag{2.2.13}$$

The likelihood ratio between the hypotheses $\mathbf{H}_0$ and $\mathbf{H}_j$ is

$$\Lambda^k_1(j) = \frac{\prod_{i=1}^{j-1} p_{\theta_0}(y_i) \cdot \prod_{i=j}^{k} p_{\theta_1}(y_i)}{\prod_{i=1}^{k} p_{\theta_0}(y_i)} \tag{2.2.14}$$

(where $\prod_{i=1}^{0} = 1$). Thus, the log-likelihood ratio is

$$S^k_j = \sum_{i=j}^{k} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \tag{2.2.15}$$

When the change time j is unknown, the standard statistical approach consists of estimating it by using the maximum likelihood principle, which leads to the following decision function :

$$g_k = \max_{1 \leq j \leq k} S^k_j \tag{2.2.16}$$

This decision function is the same as those obtained in formulas (2.2.4) and (2.2.9). It can also be written as

$$t_a = \min\{k : \max_{1 \leq j \leq k} S^k_j \geq h\} \tag{2.2.17}$$

Up to now, we have discussed only the *detection* issue in change detection problems. Let us now consider the *estimation of the change time* t_0. It follows from equation (2.2.16) that the maximum likelihood estimate of t_0 *after detection* is equal to the time j at which the maximum in (2.2.16) is reached. This estimate can be computed using the following formula :

$$\hat{t}_0 = t_a - N_{t_a} + 1 \tag{2.2.18}$$

We discuss this formula in section 2.6.

Example 2.2.1 (Change in mean - contd.). *We now continue the discussion about the simple example of a change in the mean value μ of an independent Gaussian random sequence, with known variance σ^2. We first consider the one-sided case of an increase in the mean, namely $\mu_1 > \mu_0$. In this case, (2.1.6) holds, and the decision function g_k introduced in (2.2.1), (2.2.9), and (2.2.16) becomes in the first formulation,*

$$\begin{aligned} g_k &= S_1^k - \min_{1\leq j\leq k} S_1^j \\ S_1^j &= \frac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^{j}\left(y_i - \frac{\mu_1+\mu_0}{2}\right) \end{aligned} \tag{2.2.19}$$

and in the second formulation,

$$g_k = \left[g_{k-1} + \frac{\mu_1-\mu_0}{\sigma^2}\left(y_k - \frac{\mu_1+\mu_0}{2}\right)\right]^+ \tag{2.2.20}$$

and finally

$$g_k = \max_{1\leq j\leq k} S_j^k \tag{2.2.21}$$

in the third formulation. It is obvious from the formula for S_1^j that the observations are first processed through an ordinary integration; and then, as stated before, an adaptive threshold is used.

2.2.4 Parallel Open-ended Tests

Now let us emphasize the connection between formulas (2.2.15)-(2.2.17) and an idea due to [Lorden, 1971] which turns out to be very useful for the design and the analysis of change detection algorithms. The CUSUM stopping time t_a can be interpreted using a set of *parallel* so-called open-ended SPRT, which are activated at each possible change time $j = 1, \ldots, k$, and with upper threshold h and lower threshold $-\epsilon = -\infty$. Each of these SPRT stops at time k if, for some $j \leq k$, the observations $y_j, \ldots, y_k$ are significant for accepting the hypothesis about change. Let us formalize this in the following way. Let T_j be the stopping time for the open-ended SPRT activated at time j :

$$T_j = \min\{k \geq j \ : S_j^k \geq h\} \tag{2.2.22}$$

where we use the convention that $T_j = \infty$ when this minimum is never reached. Lorden defined the following *extended stopping time* as the minimum of the T_j :

$$T^* = \min_{j=1,2,\ldots}\{T_j\} \tag{2.2.23}$$

The comparison between (2.2.17) and (2.2.22)-(2.2.23) shows that $t_a = T^*$. We continue this discussion when describing the geometrical interpretation after.

2.2.5 Two-sided CUSUM Algorithm

Let us now investigate further the situation discussed in section 2.1 where the mean value after change is either $\mu_1^+ = \mu_0 + \nu$ or $\mu_1^- = \mu_0 - \nu$, with ν known. In this case, it is relevant [Page, 1954a] to use two CUSUM algorithms together; the first for detecting an increase in the mean, and the second for detecting a decrease in the mean. The resulting alarm time is

$$\begin{aligned} t_a &= \min\{k : (g_k^+ \geq \bar{h}) \cup (g_k^- \geq \bar{h})\} \\ g_k^+ &= \left(g_{k-1}^+ + y_k - \mu_0 - \frac{\nu}{2}\right)^+ \\ g_k^- &= \left(g_{k-1}^- - y_k + \mu_0 - \frac{\nu}{2}\right)^+ \end{aligned} \tag{2.2.24}$$

In these formulas, we canceled the multiplicative term $\frac{\mu_1-\mu_0}{\sigma^2}$, which can be incorporated in the threshold $\bar{h}$ in an obvious manner. Formula (2.2.24) corresponds to the well-known *cumulative sum control chart* widely used in continuous inspection for quality control.

Let us add some comments about ν. When introducing this chapter, we discussed the availability of information about θ_1, or, equivalently from an on-line point of view, about the change magnitude ν. In most practical cases, little is known about this parameter. However, three possible *a priori* choices can be made for using the CUSUM algorithm in this case. The first consists of choosing ν as a minimum possible magnitude of jump. In the second, we choose *a priori* the most likely magnitude of jump. The third choice for ν is a kind of worst-case value from the point of view of the cost of a nondetected change. In these three cases, the resulting change detection algorithm is optimal for only *one* possible jump magnitude equal to ν. Notice that an *a posteriori* choice of the most likely magnitude leads to the GLR algorithm, which is introduced in subsection 2.4.3, and leads to the almost optimal algorithm in such a case.

From the point of view of minimum magnitude of change, the limit case is $\nu = 0$. In other words, this situation occurs when all possible jumps are to be detected, whatever their magnitude. It is useful to note [Nadler and Robbins, 1971] that, for this situation, the double CUSUM algorithm presented before in formula (2.2.24) is equivalent to

$$t_a = \min\{k : R_k \geq \bar{h}\} \tag{2.2.25}$$

where

$$R_k = \max_{j \leq k} \sum_{i=1}^{j} (y_i - \mu_0) - \min_{j \leq k} \sum_{i=1}^{j} (y_i - \mu_0) \tag{2.2.26}$$

2.2.6 Geometrical Interpretation in the Gaussian Case

If we rewrite the decision function (2.2.21), we obtain

$$g_k = \max_{1 \leq j \leq k} \sum_{i=j}^{k} \left(y_i - \mu_0 - \frac{\nu}{2} \right) \tag{2.2.27}$$

In the corresponding decision rule, the alarm is set the first time k at which there exists a time instant j_0 such that

$$\sum_{i=j_0}^{k} \left(y_i - \mu_0 - \frac{\nu}{2} \right) \geq \bar{h} \tag{2.2.28}$$

At each time k, this can be seen as a SPRT with reverse time and only one (upper) threshold $\bar{h}$ [Lorden, 1971, Page, 1954a]. For this purpose, look at figure 2.8 upside down. This can be geometrically interpreted, as depicted in figure 2.9. In this figure the cumulative sum

$$\tilde{S}_1^k = \frac{1}{\sigma} \sum_{i=1}^{k} (y_i - \mu_0) \tag{2.2.29}$$

is plotted in the case $\mu_0 = 0$. Because this cumulative sum does not contain the term $-\frac{\nu}{2}$, the corresponding threshold is no longer a constant value, but a straight line with slope $\omega \tan(\alpha)$, where ω is the horizontal distance between successive points in terms of a unit distance on the vertical scale, and α is the angle between this line and the horizontal one. It is obvious that

$$\tan(\alpha) = \frac{\nu}{2\omega} \tag{2.2.30}$$

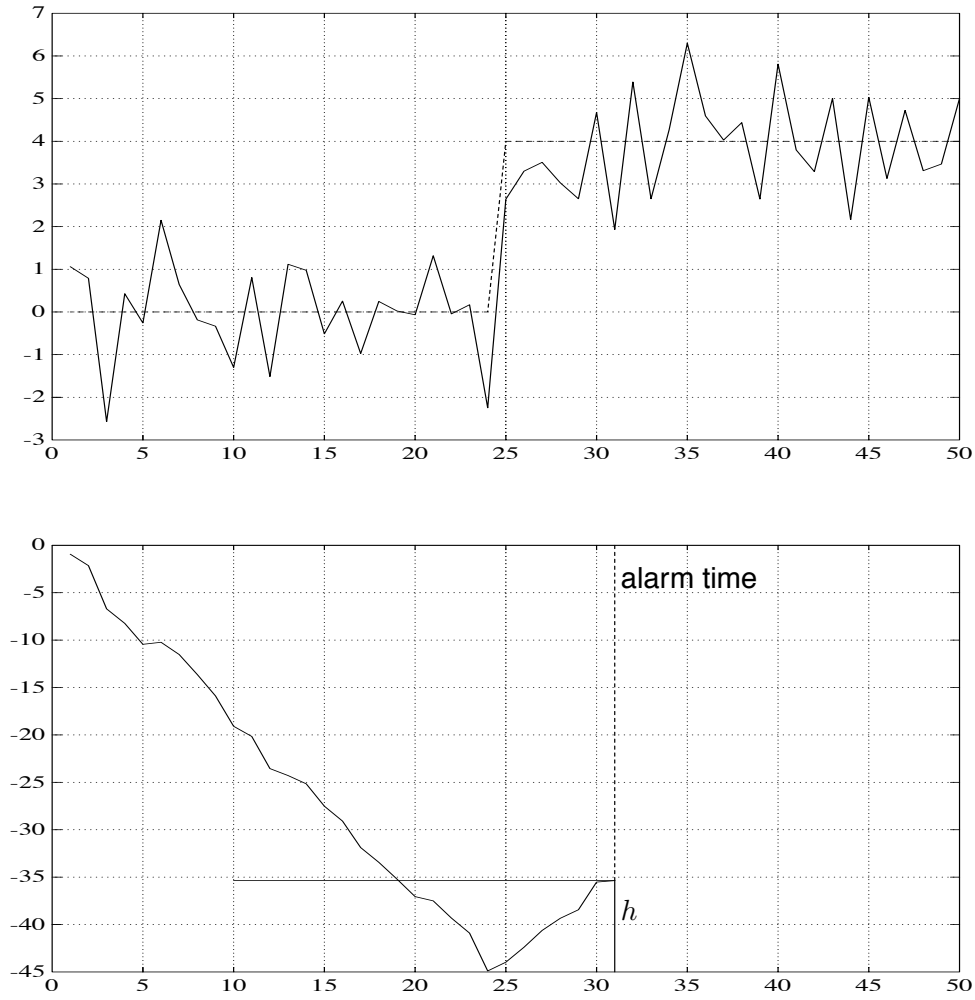

Figure 2.8 Behavior of S_j^k as a SPRT with reverse time (look upside down).

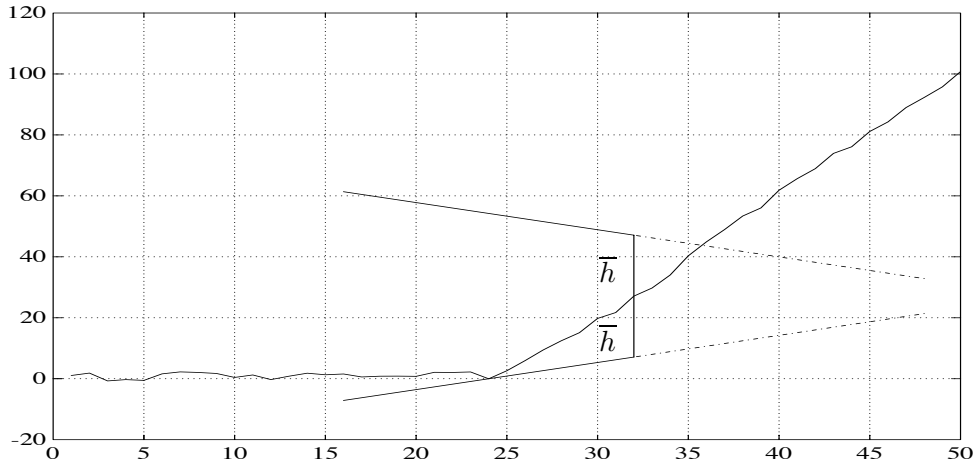

Figure 2.9 The cumulative sum $\tilde{S}_1^k$ intersected by a V-mask, in the case $\mu_0 = 0, \sigma = 1$.

This defines half a V-mask, as depicted in figure 2.9. Let $d = \bar{h}/\tan(\alpha)$ be the distance between the current sample point y_k and the vertex of the V-mask plotted forward. Then equation (2.2.28) can be rewritten in terms of these parameters :

$$\sum_{i=j_0}^{k} [y_i - \mu_0 - \omega \tan(\alpha)] \geq d \ \tan(\alpha) \tag{2.2.31}$$

Notice that, because of (2.2.30), the size of the angle α of the V-mask decreases with the magnitude ν of the jump. This concludes the geometrical interpretation for one-sided CUSUM algorithms. The geometrical interpretation of two-sided CUSUM algorithms is obtained with the aid of a symmetry of the previous picture with respect to the horizontal line, which gives rise to the so-called V-mask. The decision rule is then simply to stop when the boundaries of this mask cover any point already plotted.

The geometrical interpretation of the CUSUM algorithm when viewed as a set of open-ended SPRT is based on figure 2.10, again for the signal of figure 1.1. In this figure are depicted the cumulative sum $\tilde{S}_1^k$, several upper thresholds for the open-ended SPRT, and a standard V-mask. Note that the center of local coordinates for the SPRT beginning at time k is placed at $(k-1, y_{k-1})$. It is obvious that the slope of the upper thresholds of the parallel one-sided SPRT is the same as the slope $\omega \tan(\alpha)$ of the V-mask. This figure shows that the stopping time t_a in (2.2.17) or T^* in (2.2.23) is attained when the decision function of the one-sided SPRT reaches the upper threshold or when the cumulative sum in reverse time reaches the V-mask.

2.3 Bayes-type Algorithms

In this section, we continue to investigate the problem of detecting a change in the scalar parameter of an independent random sequence. As stated in the introduction, we discuss the Bayesian approach in which *a priori* information about the distribution of the change time is available. We assume that this information is in the form of an *a priori* probability distribution for the change time t_0. This approach was first investigated in [Girshick and Rubin, 1952] for continuous inspection of a technological process with known transition probabilities between the two (normal and abnormal) functioning modes. The theoretical derivation of opti-

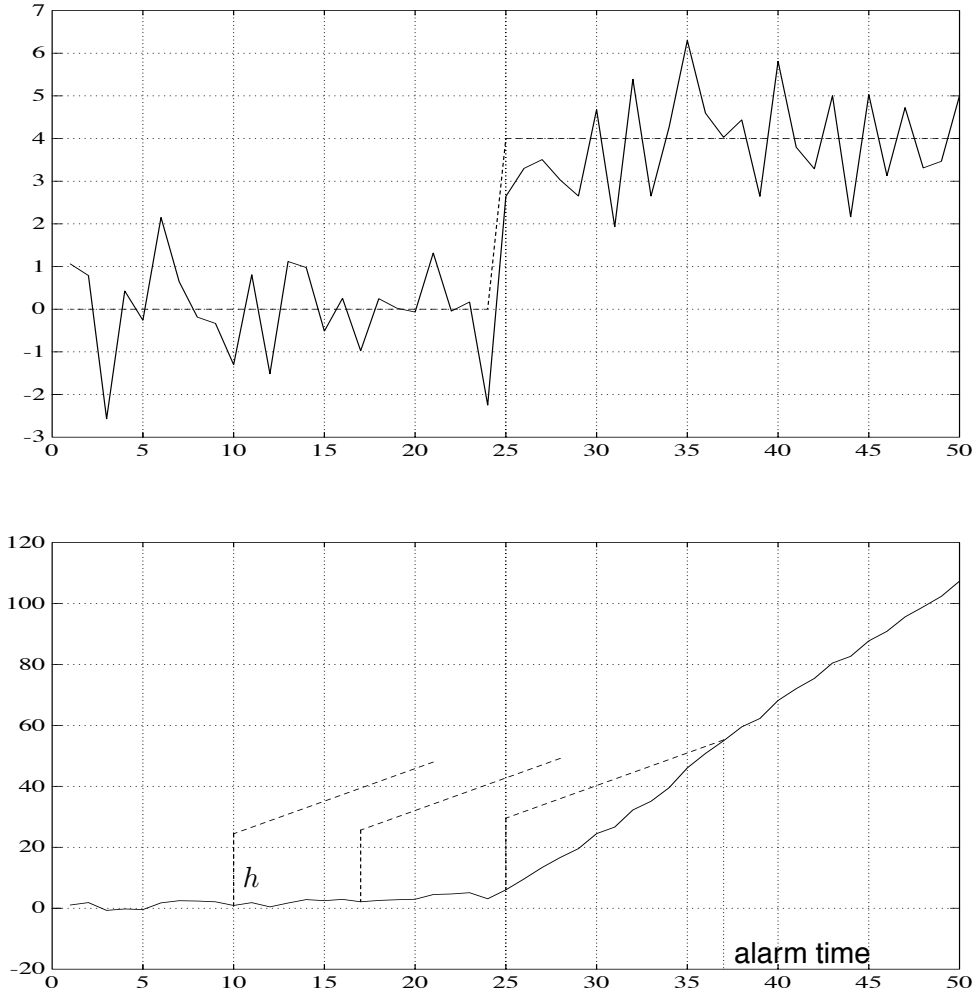

Figure 2.10 The CUSUM algorithm as a set of open-ended SPRT.

mal Bayesian algorithms for change detection was obtained in [Shiryaev, 1961]. This pioneering work was the starting point and theoretical background of a great number of other papers about Bayes-type algorithms.

The main (classical Bayesian) idea consists of deciding that a change has occurred when the *a posteriori* probability of a change exceeds a conveniently chosen threshold. We assume here that the *a priori* distribution of the change time t_0 is geometric :

$$\mathbf{P}(t_0 = k) = \varrho\,(1-\varrho)^{k-1} \quad \text{, for } k > 0$$

We assume that the change from θ_0 to θ_1 in the probability density $p_\theta(y_k)$ of our independent sequence can be modeled by a Markov chain with two states, 0 and 1. The transition matrix of this Markov chain is

$$P = \begin{pmatrix} p(0|0) & p(0|1) \\ p(1|0) & p(1|1) \end{pmatrix} = \begin{pmatrix} 1-\varrho & 0 \\ \varrho & 1 \end{pmatrix} \tag{2.3.1}$$

where $p(i|j)$ is the probability of a transition from state j to state i. The probability of the initial state is given by $p(0) = 1 - \pi$ and $p(1) = \pi$. Note that the expectation of the change time is $\mathbf{E}(t_0|t_0 > 0) = \frac{1}{\varrho}$.

Let π_k be the *a posteriori probability* of state 1 of this Markov chain. It results from Bayes' rule that

$$\pi_k = \frac{\pi_{k-1}\; p_{\theta_1}(y_k) + (1-\pi_{k-1})\; \varrho\; p_{\theta_1}(y_k)}{\pi_{k-1}\; p_{\theta_1}(y_k) + (1-\pi_{k-1})\; \varrho\; p_{\theta_1}(y_k) + (1-\pi_{k-1})(1-\varrho)\; p_{\theta_0}(y_k)} \tag{2.3.2}$$

For simplicity, we will deal with a monotonic function of π_k instead of π_k alone, because it will be more convenient for recursive computations. This function is

$$\varpi_k = \frac{\pi_k}{1-\pi_k} \tag{2.3.3}$$

The recursive formula for ϖ_k is

$$\varpi_k = \frac{1}{1-\varrho}(\varpi_{k-1} + \varrho)\; \frac{p_{\theta_1}(y_k)}{p_{\theta_0}(y_k)} \tag{2.3.4}$$

To deal with the log-likelihood ratio as in the previous sections, we rewrite this formula as follows :

$$g_k = \ln(\varrho + e^{g_{k-1}}) - \ln(1-\varrho) + \ln \frac{p_{\theta_1}(y_k)}{p_{\theta_0}(y_k)} \tag{2.3.5}$$

where

$$g_k = \ln \varpi_k \tag{2.3.6}$$

The last term is the log-likelihood ratio, which basically contains the updating information available at time k. Because g_k is an increasing function of π_k, the Bayesian stopping rule becomes :

$$t_a = \min\{k \;:\; g_k \geq h\} \tag{2.3.7}$$

exactly as in the previous sections (remember (2.2.10)).

Example 2.3.1 (Change in mean - contd.). *Let us return to our basic example. We assume here that the mean values μ_0, μ_1, and the constant variance σ^2 are known. In this case, the log-likelihood ratio is given in (2.1.6), and consequently the decision function g_k is*

$$g_k = \ln(\varrho + e^{g_{k-1}}) - \ln(1-\varrho) + \frac{\mu_1 - \mu_0}{\sigma^2}\left(y_k - \frac{\mu_0 + \mu_1}{2}\right) \tag{2.3.8}$$

The behavior of this decision function is depicted in figure 2.11, again for the signal of figure 1.1. In this figure, the influence of the choice of the parameter ϱ of the geometric distribution is emphasized. The solid line corresponds to the ideal case where we know the true value 0.05 of this parameter. The two other lines correspond to cases where the tuning value of ϱ is different from this true value.

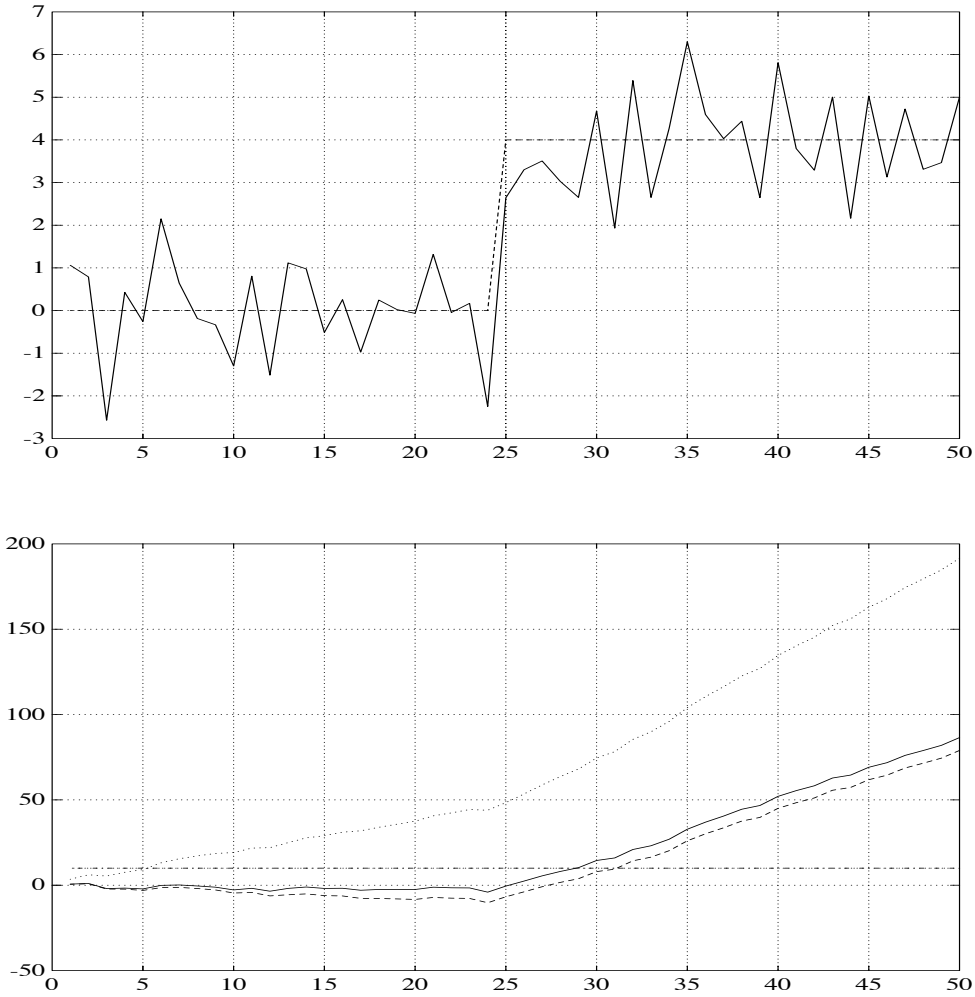

Figure 2.11 Typical behavior of a Bayesian decision function : ϱ chosen to be the true value $\varrho = 0.05$ (solid line); noncorrect but acceptable choice of $\varrho = 0.001$ (dashed line); nonacceptable choice of $\varrho = 0.9$ (dotted line).

Notice that, in some sense, the Bayesian decision rule is not of the same type as the other ones before, because it assumes the availability of the parameter ϱ of the geometric *a priori* distribution of the change time t_0, and of the initial probability π which is implicit in g_0. For this reason, the practical implementation of this decision rule is not so simple and requires a preliminary investigation of this question of *a priori* information. The effect of the choice of the parameter ϱ on the behavior of g_k is depicted in figure 2.11.

2.4 Unknown Parameter After Change

We now discuss the case where the parameter θ_1 after change is unknown. Without loss of generality in our on-line framework, the parameter θ_0 before change is assumed to be known.

2.4.1 Introduction

It follows from the previous discussion that a sequential change detection algorithm can be interpreted as a set of "parallel" open-ended tests. We begin the present discussion with these tests.

As explained in [Wald, 1947], two possible solutions exist in the present case. The first one consists of weighting the likelihood ratio with respect to all possible values of the parameter θ_1, using a weighting function $dF(\theta_1)$, where $F(\theta_1)$ may be interpreted as the cumulative distribution function of a probability measure. In the second solution, the unknown parameter θ_1 is replaced by its maximum likelihood estimate, which results in the generalized likelihood ratio (GLR) algorithm. In other words, for known θ_1, change detection algorithms are based on the likelihood ratio :

$$\Lambda_n = \frac{p_{\theta_1}(y_1, \ldots, y_n)}{p_{\theta_0}(y_1, \ldots, y_n)} \tag{2.4.1}$$

and for unknown θ_1 we must replace Λ_n by other statistic. More precisely, the first solution is based upon the weighted likelihood ratio :

$$\tilde{\Lambda}_n = \int_{-\infty}^{\infty} \frac{p_{\theta_1}(y_1, \ldots, y_n)}{p_{\theta_0}(y_1, \ldots, y_n)}\, dF(\theta_1) \tag{2.4.2}$$

and the second one uses the GLR :

$$\hat{\Lambda}_n = \frac{\sup_{\theta_1} p_{\theta_1}(y_1, \ldots, y_n)}{p_{\theta_0}(y_1, \ldots, y_n)} \tag{2.4.3}$$

We investigate these two solutions in subsections 2.4.2 and 2.4.3, respectively.

2.4.2 Weighted CUSUM Algorithm

Let us now explain in detail the algorithm resulting from the idea of weighting the unknown parameter.

2.4.2.1 Derivation of the Algorithm

We follow Lorden's idea introduced before, which explains the CUSUM algorithm as an extended stopping time associated with a family of open-ended SPRT. The weighted-CUSUM algorithm was derived for change detection in [Pollak and Siegmund, 1975], and is a direct extension of the CUSUM stopping time. It is defined as follows. Let

$$\tilde{\Lambda}_j^k = \int_{-\infty}^{\infty} \frac{p_{\theta_1}(y_j, \ldots, y_k)}{p_{\theta_0}(y_j, \ldots, y_k)} dF(\theta_1) \tag{2.4.4}$$

be the weighted likelihood ratio for the observations from time j up to time k. Then the stopping time is

$$t_a = \min\{k : \max_{1\leq j\leq k} \ln \tilde{\Lambda}_j^k \geq h\} \tag{2.4.5}$$

Typical choices of the weighting function $F(\theta)$ are the following. The most simple choices involve using the uniform distribution over a specified interval that contains all possible values of the parameter θ_1, or Dirac masses on some specified values. Another useful choice is the Gaussian distribution. Note that this type of algorithm *cannot* be written in a recursive manner as the simple CUSUM algorithm (2.2.9) that we describe in section 2.2.

Example 2.4.1 (χ^2-CUSUM algorithm). *Let us now discuss the problem of detecting a change in the mean of a Gaussian sequence with known variance σ^2, in the special case where the distribution $F(\theta) = F(\mu)$ is concentrated on two points, $\mu_0 - \nu$ and $\mu_0 + \nu$. In this case, the weighted likelihood ratio is easily shown to be*

$$\tilde{\Lambda}_j^k = \int_{-\infty}^{\infty} \exp\left[b\tilde{S}_j^k - \frac{b^2}{2}(k-j+1)\right] dF(\nu) \tag{2.4.6}$$

where

$$b = \frac{\nu}{\sigma} \tag{2.4.7}$$

is the signal-to-noise ratio, and

$$\tilde{S}_j^k = \frac{1}{\sigma}\sum_{i=j}^{k}(y_i - \mu_0) \tag{2.4.8}$$

This reduces to

$$\begin{aligned} \tilde{\Lambda}_j^k &= \cosh\left(b\tilde{S}_j^k\right) e^{-\frac{b^2}{2}(k-j+1)} \\ &= \cosh\left[b(k-j+1)\chi_j^k\right] e^{-\frac{b^2}{2}(k-j+1)} \end{aligned} \tag{2.4.9}$$

where

$$\chi_j^k = \frac{1}{k-j+1}|\tilde{S}_j^k| \tag{2.4.10}$$

Note that $\tilde{\Lambda}_j^k$ in (2.4.9) is the likelihood ratio for testing the noncentrality parameter of a χ^2 distribution with one degree of freedom, between the values 0 and $(k-j+1)\ b^2$. This fact explains the name of the χ^2-CUSUM algorithm.

The stopping time is thus

$$t_a = \min\{k : g_k \geq h\} \tag{2.4.11}$$

where

$$g_k = \max_{1\leq j\leq k}\left[\ln\cosh\left(b\tilde{S}_j^k\right) - \frac{b^2}{2}(k-j+1)\right] \tag{2.4.12}$$

As we said before, this algorithm cannot be written in a recursive manner because it is derived from Lorden's open-ended test. However, using Page's and Shiryaev's interpretation of the CUSUM algorithm as a repeated SPRT with lower threshold equal to 0 and upper threshold equal to h as discussed in subsection 2.2.2, it is possible to design a slightly modified decision rule which is written in a recursive manner.

This results in

$$
\begin{aligned}
g_k &= (\check{S}^k_{k-N_k+1})^+ && (2.4.13)\\
\check{S}^k_{k-N_k+1} &= -\frac{1}{2}N_k\, b^2 + \ln\cosh(b\tilde{S}^k_{k-N_k+1}) && (2.4.14)\\
\bar{S}_k &= \tilde{S}^k_{k-N_k+1} && (2.4.15)\\
\bar{S}_k &= \bar{S}_{k-1}\mathbf{1}_{\{g_{k-1}>0\}} + \frac{y_k-\mu_0}{\sigma} && (2.4.16)
\end{aligned}
$$

where $N_k = N_{k-1}\mathbf{1}_{\{g_{k-1}>0\}} + 1$.

This CUSUM algorithm can be used in the same situations as the two-sided CUSUM algorithm. The multidimensional parameter counterpart of this algorithm is investigated in section 7.2, case 3.

2.4.2.2 Geometrical Interpretation in the Gaussian Case

We continue to investigate the detection of a change in the mean of a Gaussian sequence, and give now the geometrical interpretation of the weighted CUSUM (2.4.4) and χ^2-CUSUM (2.4.9) algorithms in this case. We discuss first a one-sided weighted CUSUM algorithm, and then a two-sided one. We finish with the geometrical interpretation of the χ^2-CUSUM algorithm.

Let us assume that the probability measure $F(\mu)$ is confined to the interval $[\mu_0, \infty)$. The weighted CUSUM algorithm is based upon the stopping time :

$$t_a = \min\{k : g_k = \max_{1\le j\le k} \ln\tilde{\Lambda}^k_j \ge h\} \tag{2.4.17}$$

where the weighted likelihood ratio is

$$\tilde{\Lambda}^k_j = \int_0^\infty \exp\left[\frac{\nu}{\sigma}\tilde{S}^k_j - \frac{\nu^2}{2\sigma^2}(k-j+1)\right] dF(\nu) \tag{2.4.18}$$

Let us define the following function :

$$f(x,l) = \ln\int_0^\infty \exp\left(\frac{\nu}{\sigma}x - \frac{\nu^2}{2\sigma^2}l\right) dF(\nu) \tag{2.4.19}$$

Because F defines a probability measure on $(\mathbf{R}, \mathcal{R})$, the function $f(x,l)$ is an increasing function of x. It is obvious that the decision rule involves stopping the first time k at which the cumulative sum $\tilde{S}^k_j$ reaches the curve line threshold $\tilde{c}_{k-j+1}$, where $\tilde{c}_l$ is the unique positive solution of the equation $f(x,l) = h$ [Robbins, 1970]. This threshold $\tilde{c}_l$ is the half lower part of the curve in figure 2.12 and is called a U-mask. The geometrical interpretation is now the same as for the CUSUM algorithm.

If we now assume that F is a symmetric distribution over $(-\infty, \infty)$, then

$$f(x,l) \ge h \text{ if and only if } |x| \ge \tilde{c}_l \tag{2.4.20}$$

Therefore, the geometrical interpretation of the two-sided weighted CUSUM algorithm is obtained from the one-sided one, with the aid of a symmetry with respect to the horizontal line drawn at the last observation point, as depicted in the figure 2.12, and as for the ordinary CUSUM algorithm before.

Finally, let us assume that F is concentrated on two points, which corresponds to the χ^2-CUSUM algorithm. In this case, the function f can be written as

$$f(x,l) = \ln\cosh(bx) - \frac{b^2}{2}l \tag{2.4.21}$$

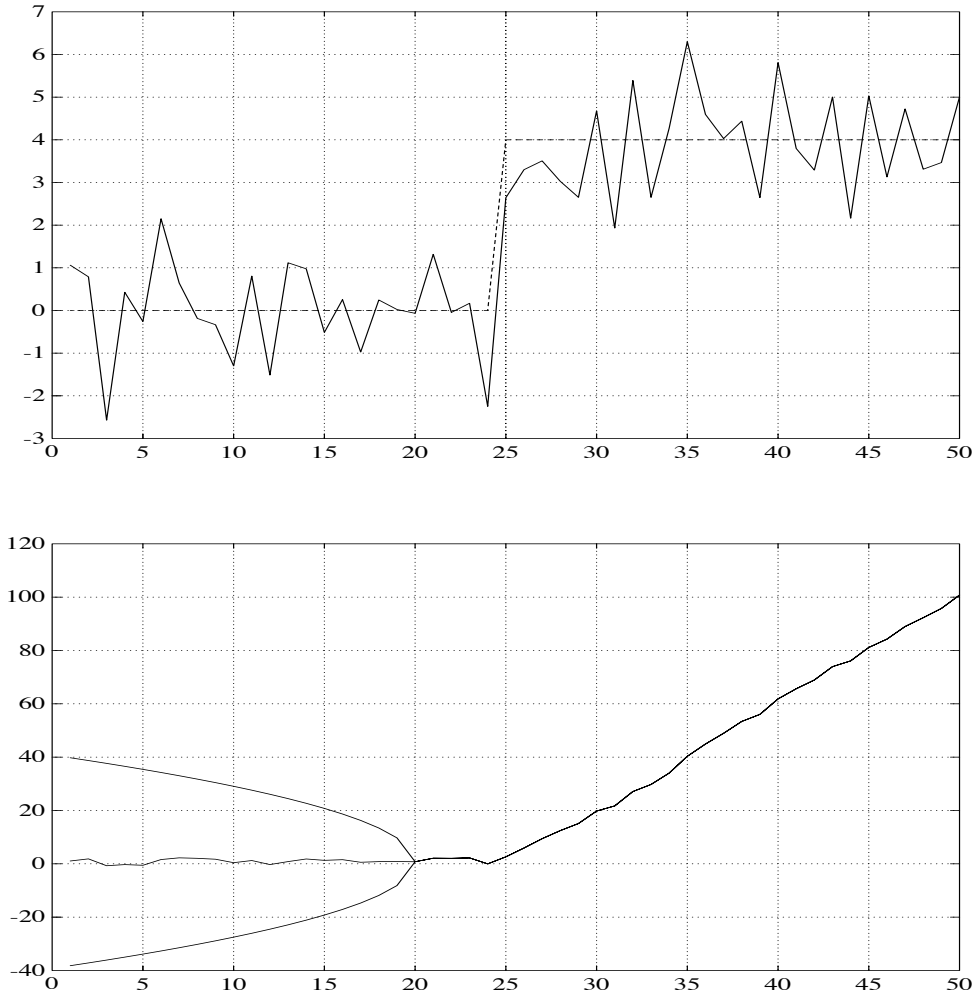

Figure 2.12 U-mask for the weighted CUSUM algorithm.

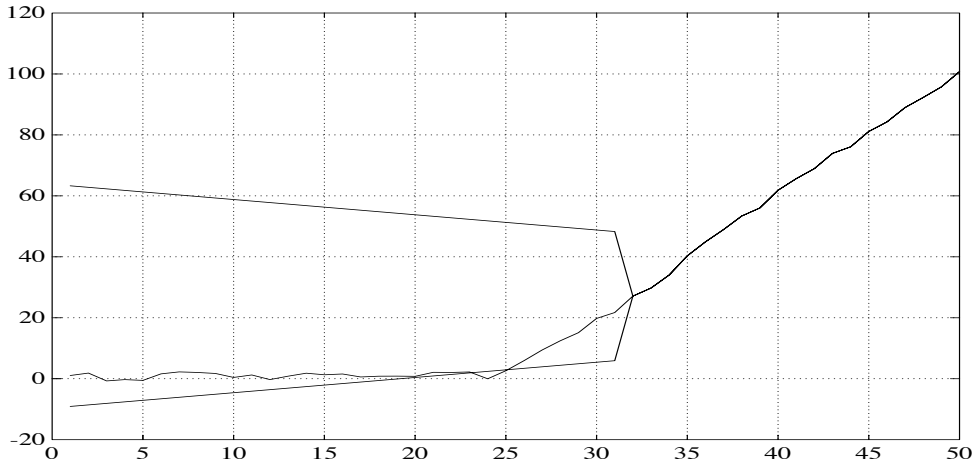

Figure 2.13 Mask for the χ^2-CUSUM algorithm.

and we wish to find $\tilde{c}_l$ such that

$$f(\tilde{c}_l, l) = h \tag{2.4.22}$$

For $v \geq 0$, the equation $\ln \cosh |u| = v$ has a unique positive solution, which is given by

$$|u| = \ln(e^v + \sqrt{e^{2v} - 1}) = v + \ln(1 + \sqrt{1 - e^{-2v}}) \tag{2.4.23}$$

From this solution the boundary $\tilde{c}_l$ is

$$|\tilde{c}_l| = \frac{1}{b}\left(h + \ln\left\{1 + \sqrt{1 - \exp\left[-2\left(h + \frac{b^2 l}{2}\right)\right]}\right\}\right) + \frac{b}{2}\, l \tag{2.4.24}$$

When l goes to infinity, the two asymptotes of this boundary have the equation

$$c_l = \pm\left(\frac{h + \ln 2}{b} + \frac{b}{2} l\right) \tag{2.4.25}$$

This fact is depicted in figure 2.13. From these formulas the difference between the boundary and its asymptotes decreases very quickly when h increases for all l. In other words,

$$\tilde{c}_l - c_l = O(e^{-2h}) \tag{2.4.26}$$

when h goes to infinity. Therefore, the stopping boundary for the χ^2-CUSUM algorithm is made nearly of straight lines, and thus is very close to the stopping boundary of the two-sided CUSUM algorithm. We continue this discussion in section 11.1.

Example 2.4.2 (Change in mean - contd.). *Let us again discuss the problem of detecting a change in the mean of a Gaussian sequence with* unit *variance, in another special case where the distribution $F(\theta) = F(\mu)$ is Gaussian with mean μ_0 and known variance σ^2. In this case, the weighted likelihood ratio can be written as*

$$\tilde{\Lambda}_j^k = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[\nu \tilde{S}_j^k - \frac{\nu^2}{2}(k - j + 1)\right] \exp\left[-\frac{\nu^2}{2\sigma^2}\right] d\nu \tag{2.4.27}$$

or

$$\ln \tilde{\Lambda}_j^k = \frac{\sigma^2}{2[\sigma^2(k-j+1)+1]} \left(\tilde{S}_j^k\right)^2 - \frac{1}{2}\ln[\sigma^2(k-j+1)+1] \tag{2.4.28}$$

where $\tilde{S}_j^k$ is defined in (2.4.8). The function $f(x,l)$ can be written as

$$f(x,l) = \frac{\sigma^2}{2(\sigma^2 l+1)}\, x^2 - \frac{1}{2}\ln(\sigma^2 l+1) \tag{2.4.29}$$

and satisfies (2.4.20). The equation $f(|x|,l) = h$ has a unique positive solution from which we deduce that the boundary $\tilde{c}_l$ is

$$\tilde{c}_l = \pm\sqrt{2(l+\sigma^{-2})\left[h + \frac{1}{2}\ln(\sigma^2 l+1)\right]} \tag{2.4.30}$$

2.4.3 GLR Algorithm

We continue to discuss the case where the parameter θ_1 after change is unknown. The parameter θ_0 before change is again assumed to be known. The derivation of the GLR algorithm proceeds in the same way as the third derivation of the CUSUM algorithm. Actually we follow [Lorden, 1971], except that we use the widely accepted term "generalized likelihood ratio" (GLR) instead of "maximum likelihood."

2.4.3.1 Derivation of the Algorithm

We now describe Wald's second solution for the case of unknown parameter after change. Let us start from the generalized likelihood ratio given in equation (2.4.3). As before, the log-likelihood ratio for the observations from time j up to time k is

$$S_j^k(\theta_1) = \sum_{i=j}^{k} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \tag{2.4.31}$$

In the present case, θ_1 is unknown; therefore, this ratio is a function of two unknown independent parameters : the change time and the value of the parameter after change. The standard statistical approach is to use the maximum likelihood estimates of these two parameters, and thus the *double* maximization :

$$g_k = \max_{1\le j\le k} \ln \hat{\Lambda}_j^k = \max_{1\le j\le k} \sup_{\theta_1} S_j^k(\theta_1) \tag{2.4.32}$$

The precise statement of the conditions on the probability densities p_{θ_i} under which this double maximization can be performed is found in [Lorden, 1971]. Actually, the densities should belong to the so-called Koopman-Darmois family of probability densities :

$$p_\theta(y) = e^{\theta T(y) - d(\theta)} h(y) \tag{2.4.33}$$

where d is strictly concave upward and infinitely differentiable over an interval of the real line. This family is discussed in detail in chapter 4. The corresponding stopping rule is the same as in (2.2.10). As we said before, this algorithm cannot be written in a recursive manner.

Now let us discuss further the issue of level of available *a priori* information about the parameter after change. In many applications, it is possible to know a minimum magnitude ν_m of the changes of interest

in the parameter θ. In this case, the second maximization in the GLR algorithm can be achieved using this minimum magnitude of change as follows :

$$g_k = \max_{1\leq j\leq k} \sup_{\theta_1:|\theta_1-\theta_0|\geq\nu_m>0} S_j^k(\theta_1) \tag{2.4.34}$$

If information about a maximum possible magnitude of change is also available, the decision function is modified accordingly in an obvious manner.

Let us now discuss the *estimation issue*. In the present case, two unknown values have to be estimated after a change has been detected : the change time t_0 and the magnitude of the jump $(\theta_1 - \theta_0)$. As far as t_0 is concerned, the estimation is the same as before in the third derivation of the CUSUM algorithm, namely the maximum likelihood estimate which is given by (2.2.18). The conditional maximum likelihood estimates of the change magnitude and time are given by

$$(\tilde{\jmath}, \tilde{\theta}_1) = \arg \max_{1\leq j\leq t_a} \sup_{\theta_1:|\theta_1-\theta_0|\geq\nu_m>0} \sum_{i=j}^{t_a} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \tag{2.4.35}$$

and $\hat{t}_0 = \tilde{\jmath}$.

Example 2.4.3 (Change in mean - contd.). *Let us return to the example of change in the mean of an independent Gaussian sequence. In this case, the mean μ_0 before change is known, and the mean μ_1 after change is unknown. The constant variance σ^2 is also known. The corresponding cumulative sum can be rewritten as*

$$S_j^k = \frac{\mu_1 - \mu_0}{\sigma^2} \sum_{i=j}^{k} \left(y_i - \frac{\mu_1 + \mu_0}{2} \right) \tag{2.4.36}$$

Let us introduce $\nu = \mu_1 - \mu_0$. Then equation (2.4.34) can be rewritten as

$$g_k = \max_{1\leq j\leq k} \sup_{\nu:|\nu|\geq\nu_m>0} \sum_{i=j}^{k} \left[\frac{\nu(y_i - \mu_0)}{\sigma^2} - \frac{\nu^2}{2\sigma^2} \right] \tag{2.4.37}$$

In the present independent Gaussian case, the constrained maximization over ν is explicit :

$$g_k = \max_{1\leq j\leq k} \sum_{i=j}^{k} \left[\frac{\hat{\nu}_j(y_i - \mu_0)}{\sigma^2} - \frac{\hat{\nu}_j^2}{2\sigma^2} \right] \tag{2.4.38}$$

where the absolute value of the constrained *change magnitude estimate is*

$$|\hat{\nu}_j| = \left(\frac{1}{k-j+1} \sum_{i=j}^{k} |y_i - \mu_0| - \nu_m \right)^+ + \nu_m \tag{2.4.39}$$

and its sign is the same as the sign of the mean value $\frac{1}{k-j+1}\sum_{i=j}^{k}(y_i - \mu_0)$ of the last centered observations or "innovations." Note that the second term $\frac{\nu^2}{2\sigma^2}$ on the right side of (2.4.37) is nothing but the Kullback information between the two laws before and after the change.

Note also that, when $\nu_m = 0$, the decision function is

$$g_k = \frac{1}{2\sigma^2} \max_{1\leq j\leq k} \frac{1}{k-j+1} \left[\sum_{i=j}^{k} (y_i - \mu_0) \right]^2 \tag{2.4.40}$$

The above property of explicit maximization over the unknown parameter θ_1 after change can be exploited in more complex situations, as explained in section 7.2.4. Furthermore, (2.4.38) can be viewed as a correlation *between the innovation $(y_i - \mu_0)$ and the "signature" of the change $\hat{\nu}_k$. This correlation property, which is typical for matched-filtering operations, is recovered in (7.2.118) for the more general situation of additive changes in state-space models.*

Finally, let us comment further on the asymptotic equivalence, in the Gaussian case again, between the three algorithms, which we describe for the case of unknown parameter after change. As we explain in the previous subsection, the χ^2-CUSUM algorithm is asymptotically equivalent to the two-sided CUSUM algorithm when the threshold goes to infinity. But it should be clear that the two-sided CUSUM algorithm is nothing but the GLR algorithm corresponding to the degenerate situation where $\mu_1 = \mu_0 \pm \nu$.

2.4.3.2 Geometrical Interpretation in the Gaussian Case

We describe the geometrical interpretation of the GLR algorithm in the same way we described the CUSUM algorithm, namely starting from the reverse time interpretation of the decision function. We begin with a one-sided GLR algorithm, and we use a symmetry with respect to the horizontal line for the two-sided case as before. From the decision function (2.4.32), it follows that the stopping rule can be rewritten in reverse time as follows. There exists a time instant l such that the following inequality holds :

$$\sup_{\nu:\nu\geq\nu_m>0} \sum_{i=1}^{l} \left[\nu(y_i - \mu_0) - \frac{\nu^2}{2}\right] \geq h\sigma^2 \tag{2.4.41}$$

This can be rewritten as

$$\tilde{S}_1^l = \frac{1}{\sigma}\sum_{i=1}^{l}(y_i - \mu_0) \geq \inf_{\nu:\nu\geq\nu_m>0} \left(\frac{h\sigma}{\nu} + \frac{\nu}{2\sigma}l\right) \tag{2.4.42}$$

Let us now introduce the lower boundary $\hat{c}_l$ for the cumulative sum $\tilde{S}_1^l$:

$$\hat{c}_l = \inf_{\nu:\nu\geq\nu_m>0} \left(\frac{h\sigma}{\nu} + \frac{\nu}{2\sigma}l\right) \tag{2.4.43}$$

and discuss this minimization. We distinguish two situations for the parameter ν: $\nu = \nu_m$ and $\nu > \nu_m$. For the situation $\nu = \nu_m$, and from the discussion in section 2.2 about the geometrical interpretation of the stopping rule in terms of the V-mask, we find that, for large l, the boundary in (2.4.43) is the straight line with minimal angle with respect to the horizontal line, as depicted in figure 2.14. For $\nu > \nu_m$, the boundary is a curve, as we explain now. Let us consider again the reverse time SPRT with one threshold h. Because of the Wald's identity (which we explain in detail in chapter 4), for a SPRT with threshold h, the average number of samples until the threshold is reached is asymptotically

$$\mathbf{E}(l) \approx \frac{h}{\mathbf{K}(\nu)} \tag{2.4.44}$$

where $\mathbf{K}$ is the Kullback information. In the Gaussian case, it is well known that $\mathbf{K}(\nu) = \frac{\nu^2}{2\sigma^2}$. It follows that, for $l \geq \frac{h}{\mathbf{K}(\nu_m)}$, the minimum in equation (2.4.43) is then reached for $\nu = \nu_m$. On the other hand, for small values of l, the minimum in equation (2.4.43) is then reached for ν such that $l\ \mathbf{K}(\nu) = h$. Inserting this value in equation (2.4.43), we obtain

$$\hat{c}_l = \sqrt{2hl} \tag{2.4.45}$$

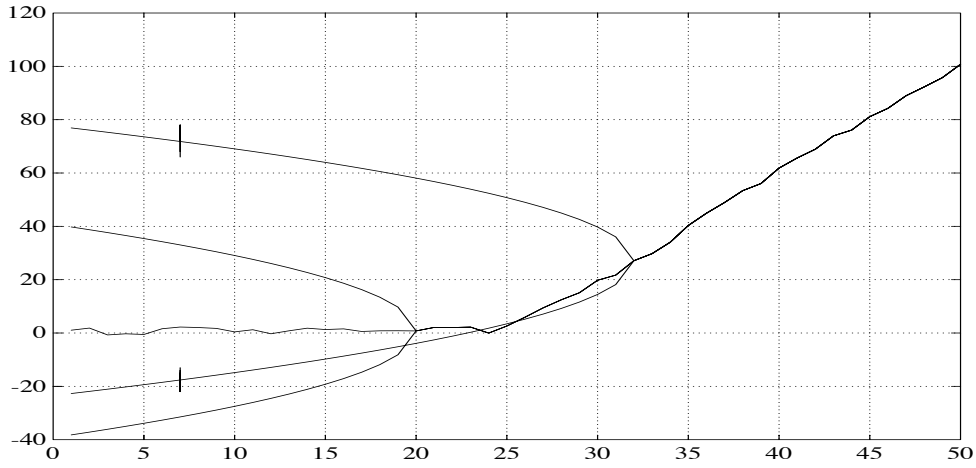

Figure 2.14 U-mask for the GLR algorithm : boundary with equation (2.4.46).

which is the equation of a parabola, leading to the so-called U-mask depicted in figure 2.14. This parabola is inscribed in the V-mask discussed before, because the points of tangency between the straight line and the parabola have the abscissa $l = \frac{2h\sigma^2}{\nu_m^2}$ as depicted by vertical segments in this figure. In summary, the equation of the boundary is

$$\hat{c}_l = \begin{cases} \sqrt{2hl} & \text{if } l \leq \frac{2h\sigma^2}{\nu_m^2} \\ \frac{h\sigma}{\nu_m} + \frac{\nu_m l}{2\sigma} & \text{otherwise} \end{cases} \tag{2.4.46}$$

The explanation for the upper boundary is the same.

As we explained before, the GLR algorithm is computationally complex. Approximations of this algorithm, with lower computational cost, are thus of interest. In [Lorden and Eisenberger, 1973], a possible approximation of the GLR algorithm dealing with the joint use of two CUSUM algorithms is proposed. These two algorithms are designed to detect changes with large and small magnitudes, respectively. The geometrical interpretation of this approximation is that a U-mask can be approximated by the intersection of two V-masks, as depicted in figure 2.15. This point is further discussed in chapter 11.

2.5 Change Detection and Tracking

In this section, we do not introduce any other derivations of change detection algorithms. Instead we explain an example of the use of one of the previously described algorithms in the framework of adaptive identification, for improving the tracking capability of adaptive identification algorithms.

Let us consider the simple example of a piecewise constant sequence perturbed by a white Gaussian noise ε. In other words, we consider the multiple change times counterpart of the above widely discussed example, modeled as

$$y_k = \varepsilon_k + \mu(k) \tag{2.5.1}$$

where $\mu(k)$ is an unknown piecewise constant function of time, as depicted in figure 2.16. The standard recursive estimation of the mean value can be written as

$$\bar{y}_k = \frac{k-1}{k}\, \bar{y}_{k-1} + \frac{1}{k}\, y_k \tag{2.5.2}$$

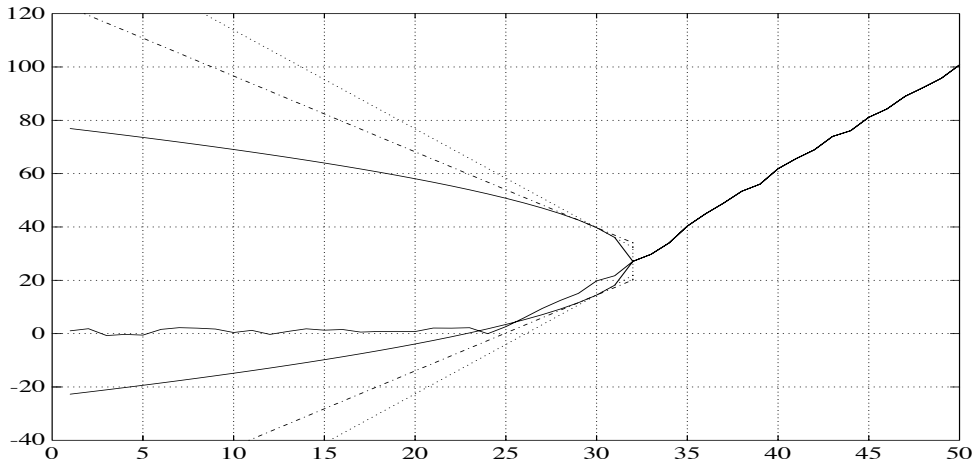

Figure 2.15 Two V-masks (dotted lines) approximating one U-mask (solid curve) : how a GLR algorithm can be approximated by two CUSUM algorithms for detecting changes with small and large magnitudes, respectively.

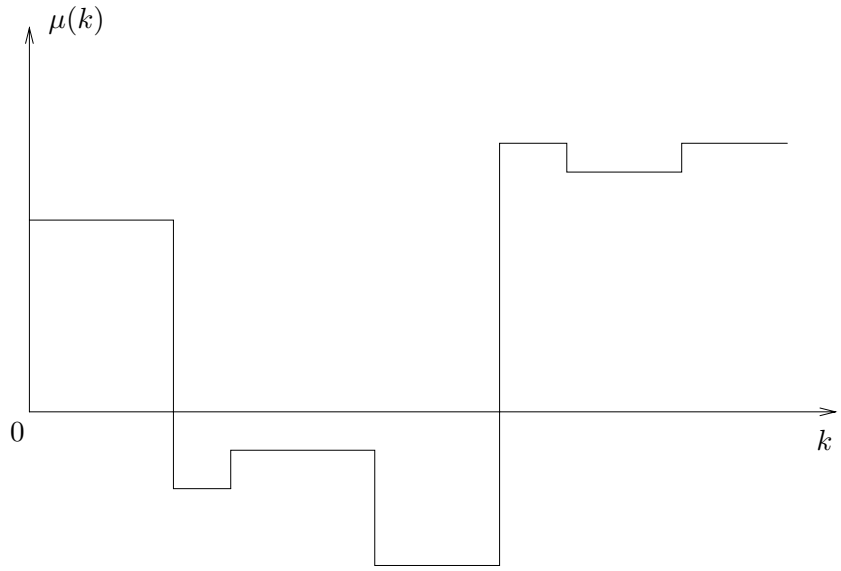

Figure 2.16 Piecewise constant signal.

This estimation is known to be efficient provided that the underlying unknown mean value is constant. Our suggestion is to use change detection algorithms for checking this assumption. We assume that the time duration between successive jumps is bounded from below. This assumption is necessary for the initial estimation of the mean to be used in the subsequent detection of change. The joint use of the estimation and detection algorithms results in cycles made of the following steps :

1. Initial estimation of the mean, during a fixed size time interval during which the detection algorithm is switched off; let $\bar{y}_N$ be this estimated mean value.
2. Carrying on the estimation and activation of the change detection algorithm using $\mu_0 = \bar{y}_k$ for $k \geq N$.
3. Updating the initial estimation after a change has been detected. This updating can take place either at the alarm time if no other information is provided by the change detection algorithm, or at the estimated change time $\hat{t}_0$ if this information is available. Similarly, the updating can include the possible estimate $\hat{\nu}$ of the magnitude of the jump. If both values $\hat{t}_0$ and $\hat{\nu}$ are available, returning to step 1 after a change has been detected is not necessary; the cycle restarts from step 2.

The two main types of relevant change detection algorithms to be used in such a cycle are the CUSUM and GLR algorithms introduced before. The main reason is that these are the only algorithms that can provide us with an estimate of the change time t_0 in addition to an alarm time t_a.

Let us add some comments about the tuning of change detection algorithms in such a framework. Minimum values ν_m of jump magnitudes (for the CUSUM and GLR algorithms) and thresholds are required. Minimum values of jumps must be close to the precision of the estimation algorithm, for example, of the order of magnitude of the corresponding standard deviation of the estimate. On the other hand, the threshold has to be chosen in such a way that the mean time between false alarms should not be too much less than the mean time between successive jumps in the piecewise function.

2.6 Off-line Change Detection

In this section, we introduce two new tasks, which were mentioned in subsection 1.1.2 :

1. *Off-line hypotheses testing* between the hypotheses "without change" and "with change."
2. *Off-line estimation of the unknown change time*.

The main difference between this section and the previous ones is that now the complete sample of observations is available before beginning the investigation for a change.

This task was first investigated in [Page, 1957], using basically the same type of ideas that he used for the CUSUM algorithm, which are described in subsection 2.2.3. The problem of off-line estimation of the change time was investigated in [Hinkley, 1970, Hinkley, 1971], including precision issues and the distribution of the estimation error.

2.6.1 Off-line Hypotheses Testing

Let $(y_k)_{1 \leq k \leq N}$ be a sequence of independent random observations with density $p_\theta(y)$. Two situations are possible. Either all the observations in this sample have the same density, characterized by $\tilde{\theta}_0$, or there exists an *unknown change time* $1 < t_0 \leq N$ such that, before t_0, the parameter θ is equal to θ_0, and after the change it is equal to $\theta_1 \neq \theta_0$. Let us first assume that $\tilde{\theta}_0$, θ_0, and θ_1 are known. As discussed in subsection 2.2.3, it is convenient to introduce the following hypotheses about this sequence of observations :

$$\begin{array}{rrll} & \mathbf{H}_0 : & \theta = \tilde{\theta}_0 & \text{for} \quad 1 \leq k \leq N \\ \text{for } 1 \leq j \leq N, & \mathbf{H}_j : & \theta = \theta_0 & \text{for} \quad 1 \leq k \leq j-1 \\ & & \theta = \theta_1 & \text{for} \quad j \leq k \leq N \end{array} \tag{2.6.1}$$

The problem is to test between the hypothesis $\mathbf{H}_0$ and the composite hypothesis :

$$\mathcal{H}_1 = \cup_{j\geq 1}\mathbf{H}_j \tag{2.6.2}$$

Note that the estimation of the change time is *not* included in this problem statement, and that the unknown change time may be interpreted here as a *nuisance* parameter. The estimation of the change time is discussed in the next subsection.

The likelihood ratio corresponding to the hypotheses $\mathbf{H}_0$ and $\mathbf{H}_j$ is

$$\Lambda_1^N(j) = \frac{\prod_{i=1}^{j-1} p_{\theta_0}(y_i) \cdot \prod_{i=j}^{N} p_{\theta_1}(y_i)}{\prod_{i=1}^{N} p_{\tilde{\theta}_0}(y_i)} \tag{2.6.3}$$

(where $\prod_{i=1}^{0} = 1$). The standard statistical approach in this situation consists of replacing the unknown parameter t_0 by its *maximum likelihood estimate* (M.L.E.). Therefore, we consider the following statistic :

$$\Lambda_N = \max_{1\leq j\leq N} \Lambda_1^N(j) \tag{2.6.4}$$

and the decision rule d such that $d = 0\ (1)$, according to which hypothesis $\mathbf{H}_0\ (\mathcal{H}_1)$ is chosen, is given by

$$d = \begin{cases} 0 & \text{if} \quad \ln \Lambda_N < h \\ 1 & \text{if} \quad \ln \Lambda_N \geq h \end{cases} \tag{2.6.5}$$

When the parameters $\tilde{\theta}_0, \theta_0$ and θ_1 are unknown, they are also replaced by their M.L.E. This results in the following decision function :

$$\tilde{\Lambda}_N = \max_{1\leq j\leq N} \sup_{\tilde{\theta}_0} \sup_{\theta_0} \sup_{\theta_1} \Lambda_1^N(j, \tilde{\theta}_0, \theta_0, \theta_1) \tag{2.6.6}$$

2.6.2 Off-line Estimation of the Change Time

We consider the same hypotheses as in the previous subsection. We assume the existence of a change point (typically this assumption is the result of the previous hypotheses testing) and the problem is now to estimate the change time. In the present case, all the parameters θ_0, θ_1, and t_0 are assumed to be unknown. Therefore, the corresponding M.L.E. algorithm is

$$(\hat{t}_0, \hat{\theta}_0, \hat{\theta}_1) = \arg \max_{1\leq k\leq N} \sup_{\theta_0} \sup_{\theta_1} \ln \left[\prod_{i=1}^{k-1} p_{\theta_0}(y_i) \prod_{i=k}^{N} p_{\theta_1}(y_i) \right] \tag{2.6.7}$$

which can be condensed into

$$\hat{t}_0 = \arg \max_{1\leq k\leq N} \ln \left[\prod_{i=1}^{k-1} p_{\hat{\theta}_0}(y_i) \prod_{i=k}^{N} p_{\hat{\theta}_1}(y_i) \right] \tag{2.6.8}$$

where $\hat{\theta}_0$ is the M.L.E. estimate of θ_0 based on the observations $y_1, ..., y_{k-1}$, and $\hat{\theta}_1$ is the M.L.E. estimate of θ_1 based upon the observations $y_k, ..., y_N$. When θ_0 and θ_1 are assumed to be known, this can be simplified to

$$\hat{t}_0 = \arg \max_{1\leq k\leq N} \ln \left[\prod_{i=1}^{k-1} p_{\theta_0}(y_i) \prod_{i=k}^{N} p_{\theta_1}(y_i) \right] \tag{2.6.9}$$

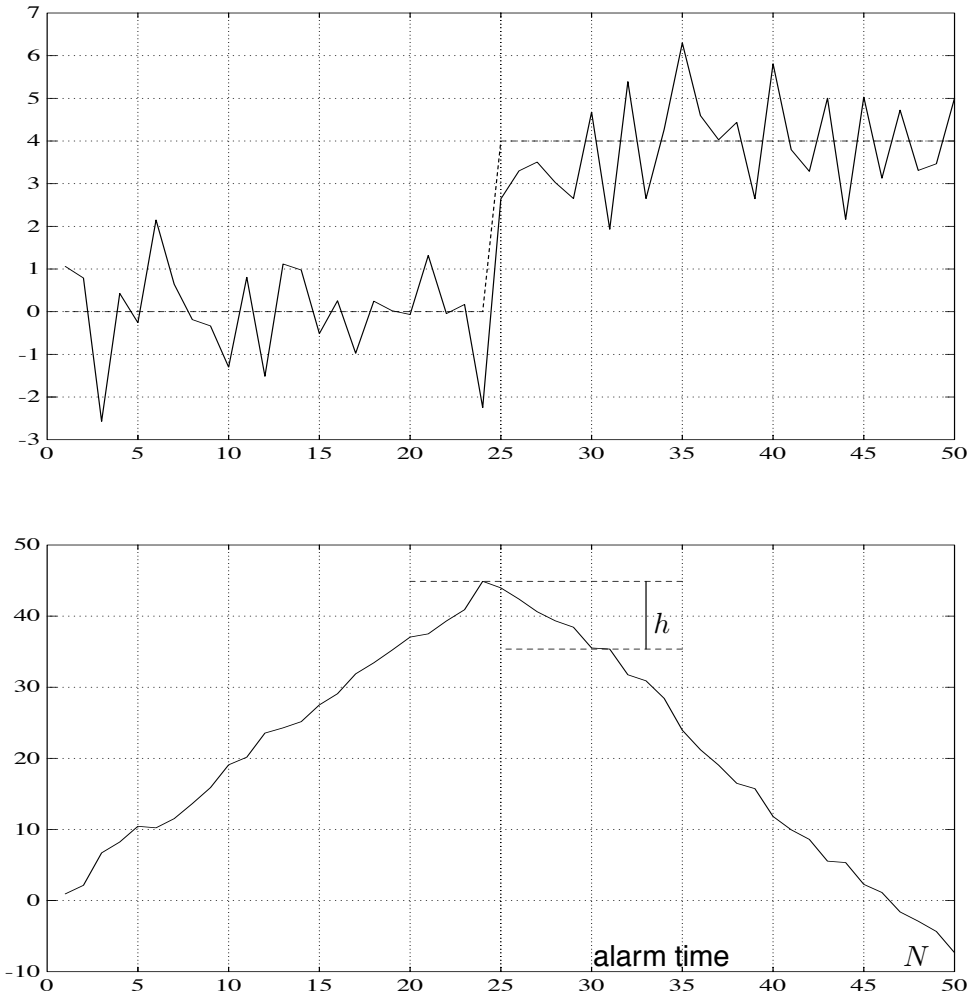

Figure 2.17 Estimation of the change time. The MLE of the change time is the abscissa of the maximum value of the cumulative sum S_k^N.

and rewritten as

$$\hat{t}_0 = \arg \max_{1 \leq k \leq N} \left[\ln \frac{\prod_{i=k}^{N} p_{\theta_1}(y_i)}{\prod_{i=k}^{N} p_{\theta_0}(y_i)} + \ln \prod_{i=1}^{N} p_{\theta_0}(y_i) \right] \tag{2.6.10}$$

The second term on the right of this equation is constant for a given sample. Therefore, the estimate of the change time is

$$\hat{t}_0 = \arg \max_{1 \leq k \leq N} \sum_{i=k}^{N} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \tag{2.6.11}$$

The geometrical interpretation of this estimation method is depicted in figure 2.17, in which we plot the cumulative sum :

$$S_k^N = \sum_{i=k}^{N} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \tag{2.6.12}$$

The figure shows that the M.L.E. of t_0 is the abscissa of the maximum value of this sum. Let us add some further comments about the relationship between this algorithm and the CUSUM algorithm described in subsection 2.2.3. Formula (2.6.11) can be rewritten as

$$\hat{t}_0 = \arg \min_{1 \leq k \leq N} \sum_{i=1}^{k-1} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \tag{2.6.13}$$

which has the following geometrical interpretation. Let us return once more to figure 2.5. From the previous formula, it is obvious that the estimate $\hat{t}_0$ is one plus the abscissa of the minimum value of the cumulative sum plotted in this figure. On the other hand, the on-line CUSUM algorithm can be geometrically interpreted with the aid of figure 2.17 in the following manner. The alarm of this on-line algorithm is set when the deviation of the cumulative sum S_k^N with respect to its current maximum value is greater than the threshold h. If you look at figure 2.17 *both upside down and from the back*, you see that you exactly recover the picture of figure 2.5. From this explanation, it is obvious that estimate (2.6.13) can be rewritten as in (2.2.18).

Example 2.6.1 (Change in mean - contd.). *We continue the investigation of the Gaussian independent case, and we assume that the variance σ^2 is known, but that the two mean values μ_0 before and μ_1 after the change are unknown. In this case, the M.L.E. formula (2.6.8) can be written as*

$$\hat{t}_0 = \arg \max_{1 \leq k \leq N} \left\{ - \left[\sum_{i=1}^{k-1} (y_i - \hat{\mu}_0)^2 + \sum_{i=k}^{N} (y_i - \hat{\mu}_1)^2 \right] \right\} \tag{2.6.14}$$

where we canceled the terms that do not modify the argument of the maximization. By replacing the estimates by their values, which are the relevant empirical means of the observations,

$$\hat{\mu}_0 = \frac{1}{k-1} \sum_{i=1}^{k-1} y_i \tag{2.6.15}$$

and

$$\hat{\mu}_1 = \frac{1}{N-k+1} \sum_{i=k}^{N} y_i \tag{2.6.16}$$

we obtain, after straightforward manipulations,

$$\hat{t}_0 = \arg \max_{1 \leq k \leq N} \left[-(k-1)(N-k+1)(\hat{\mu}_0 - \hat{\mu}_1)^2 \right] \tag{2.6.17}$$

The geometrical interpretation is the same as before in figure 2.17.

Let us give a further interpretation of (2.6.14) in terms of least-squares estimation. This equation can be rewritten as

$$\hat{t}_0 = \arg \min_{1 \leq k \leq N} \inf_{\mu_0, \mu_1} \left[\sum_{i=1}^{k-1} (y_i - \mu_0)^2 + \sum_{i=k}^{N} (y_i - \mu_1)^2 \right] \tag{2.6.18}$$

In other words, we use a least-squares estimation algorithm for the following piecewise regression problem :

$$y_k = \mu(k) + \varepsilon_k \tag{2.6.19}$$

where $\mathcal{L}(\varepsilon_k) = \mathcal{N}(0, \sigma^2)$ and

$$\mu(k) = \begin{cases} \mu_0 & \text{if} \quad k < t_0 \\ \mu_1 & \text{if} \quad k \geq t_0 \end{cases} \tag{2.6.20}$$

as depicted in figure 2.18. This problem is the simplest case of the more complex problem of choice of segments for piecewise approximation, which is also called two-phase regression. More details can be found in [Quandt, 1958, Quandt, 1960, Hinkley, 1969, Hinkley, 1971, Seber, 1977].

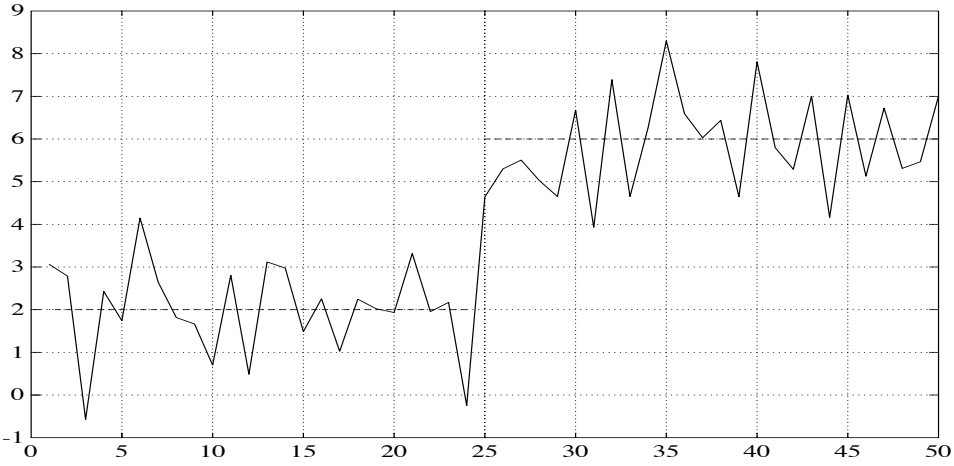

Figure 2.18 Least-squares regression : piecewise constant mean (dotted line), and corresponding Gaussian signal (solid line).

2.7 Notes and References

Section 2.1

All these algorithms were introduced for solving problems in quality control [Duncan, 1986], which is the origin of the word "chart", as used in this context. The first proposed algorithm was Shewhart's control chart [Shewhart, 1931], which was investigated further in [Page, 1954c]. The geometric moving average algorithm was introduced in [S.Roberts, 1959] as a more efficient alternative to Shewhart's chart in many cases. Another alternative, finite moving average chart, was introduced in [Page, 1954a, Laï, 1974]. A close although essentially different algorithm, the filtered derivative algorithm, was introduced in [Basseville *et al*., 1981]; this algorithm is similar to the gradient techniques used for edge detection in image processing [L.Roberts, 1965].

Section 2.2

The CUSUM algorithm was introduced in [Page, 1954a]. The literature concerning this algorithm is quite extensive [Phillips, 1969, Woodward and Goldsmith, 1964, Van Dobben De Bruyn, 1968, Hinkley, 1969, Hinkley, 1970, Hinkley, 1971]. One reason for this situation is the optimal property of this algorithm, which was proved in [Lorden, 1971]. This algorithm is also often referred to as Shiryaev's SPRT [Shiryaev, 1961].

Section 2.3

Bayesian techniques for change detection were introduced in [Girshick and Rubin, 1952], further developed and investigated in [Shiryaev, 1961, Shiryaev, 1963, Shiryaev, 1965, S.Roberts, 1966], and more recently in [Shiryaev, 1978, Pollak, 1985, Pollak, 1987]. They were initially the result of the first attempt to solve change detection problems in quality control with the aid of a formal mathematical problem statement. The optimal properties of these algorithms were obtained *before* the proof of optimality of CUSUM techniques, and with the aid of slightly different criteria.

Section 2.4

In the case of an unknown parameter after change, the GLR algorithm was derived in [Lorden, 1971] as a generalization of the CUSUM algorithm for this situation. The interest in this algorithm is justified by its "uniformly optimal properties" [Lorden, 1971, Lorden, 1973]. This algorithm is less efficient than the CUSUM algorithm because it does not require the precise knowledge of the parameter after change. Furthermore, the possibility of adapting it to more complex situations makes this algorithm quite attractive. Another less sensitive algorithm is the weighted CUSUM algorithm introduced in [Pollak and Siegmund, 1975]. The χ^2-CUSUM algorithm was introduced in [Nikiforov, 1980, Nikiforov, 1986].

Section 2.5

To our knowledge, the idea of using a change detection algorithm to improve the performance of an adaptive identification algorithm was introduced in [Willsky and Jones, 1976], which is an extension of the work in [MacAulay and Denlinger, 1973]. For earlier investigations concerning the joint use of detection and identification, the reader is referred to [Lainiotis, 1971]. In the present framework of a change in a scalar parameter, the CUSUM algorithm was used in [Perriot-Mathonna, 1984, Favier and Smolders, 1984, Bivaikov, 1991]. Similar attempts, although not based on the same detection algorithms, can be found in [Hägglund, 1983, Chen and Norton, 1987, Mariton *et al*., 1988].

Section 2.6

The off-line hypotheses testing problem was first addressed in [Page, 1957]. Other investigations can be found in [Deshayes and Picard, 1986, Siegmund, 1985b]. The off-line estimation of a change time was originally obtained in [Page, 1957]. The literature on this issue is extensive [Hinkley, 1969, Hinkley, 1970, Hinkley, 1971, Kligiene and Telksnys, 1983, Picard, 1985, Deshayes and Picard, 1986].

2.8 Summary

Main notation :

$$\begin{aligned} s_i &= \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \\ S_j^k &= \sum_{i=j}^{k} s_i; \qquad S_k = S_1^k \\ t_a &= \min\{k : g_k \geq h\} \end{aligned}$$

For the basic example of a change in the mean μ of a Gaussian distribution with constant variance σ^2, we also use the notation :

$$\begin{aligned} b &= \frac{\mu_1 - \mu_0}{\sigma} \\ s_i &= \frac{b}{\sigma}\left(y_i - \frac{\mu_0 + \mu_1}{2}\right) \\ \tilde{S}_j^k &= \frac{1}{\sigma}\sum_{i=j}^{k}(y_i - \mu_0) \end{aligned}$$

Elementary Algorithms

Shewhart control chart

$$g_{KN} = S_1^N(K) = S_{N(K-1)+1}^{NK}$$

where K is the sample number. The tuning parameters are the size N of the sample of observations tested and the threshold h.

GMA algorithm

$$g_k = (1-\alpha)g_{k-1} + \alpha s_k, \text{ with: } g_0 = 0$$

The tuning parameters are the weight $0 < \alpha \leq 1$ and the threshold h.

FMA algorithm

$$g_k = \sum_{i=0}^{N} \gamma_i \ln \frac{p_{\theta_1}(y_{k-i})}{p_{\theta_0}(y_{k-i})}$$

The tuning parameters are the size N of the sliding window, the weights γ_i, which are any weights for causal filters, and the threshold h.

Filtered derivative algorithm

$$\begin{aligned} \nabla g_k &= g_k - g_{k-1} \\ t_a &= \min\{k : \sum_{i=0}^{N} \mathbf{1}_{\{\nabla g_{k-i} \geq h\}} \geq \eta\} \end{aligned}$$

The tuning parameters are again the size N of the sliding window, the weights γ_i, which are any weights for causal filters, the threshold h, and the counter of alarms η. For the basic example, two useful choices are

$$\begin{aligned} \nabla g_k &= y_k - y_{k-N} \\ \nabla g_k &= \sum_{i=0}^{N-1} y_{k-i} - \sum_{i=N}^{2N-1} y_{k-i} \end{aligned}$$

CUSUM Algorithm

Intuitive derivation of the CUSUM algorithm

$$\begin{aligned} g_k &= S_k - m_k \\ m_k &= \min_{1 \leq j \leq k} S_j \end{aligned}$$

The stopping rule can thus be rewritten as

$$t_a = \min\{k : S_k \geq m_k + h\}$$

or equivalently as an integrator compared to an adaptive threshold.

CUSUM as a repeated SPRT The CUSUM algorithm can be recursively written as

$$g_k = (g_{k-1} + s_k)^+$$

or equivalently as

$$\begin{aligned} g_k &= \left(S_{k-N_k+1}^k\right)^+ \\ N_k &= N_{k-1} \cdot \mathbf{1}_{\{g_{k-1}>0\}} + 1 \end{aligned}$$

The CUSUM algorithm can thus be seen as a random size sliding window algorithm.

Off-line derivation

$$g_k = \max_{1 \le j \le k} S_j^k$$

The estimate of the change time is

$$\hat{t}_0 = t_a - N_{t_a} + 1$$

Two-sided CUSUM algorithm For the basic example,

$$\begin{aligned} t_a &= \min\{k : (g_k^+ \ge \bar{h}) \cup (g_k^- \ge \bar{h})\} \\ g_k^+ &= \left(g_{k-1}^+ + y_k - \mu_0 - \frac{\nu}{2}\right)^+ \\ g_k^- &= \left(g_{k-1}^- - y_k + \mu_0 - \frac{\nu}{2}\right)^+ \end{aligned}$$

Bayes-type Algorithms

$$g_k = \ln(\varrho + e^{g_{k-1}}) - \ln(1-\varrho) + \ln \frac{p_{\theta_1}(y_k)}{p_{\theta_0}(y_k)}$$

The tuning parameters of this Bayes-type algorithm are the *a priori* probability ϱ of a change, the initial probability π implicit in g_0, and the threshold h.

Unknown Parameter After Change

χ^2-CUSUM algorithm For the basic example,

$$g_k = \max_{1 \le j \le k} \left[\ln \cosh\left(b\tilde{S}_j^k\right) - \frac{b^2}{2}(k-j+1) \right]$$

GLR algorithm

$$g_k = \max_{1 \le j \le k} \sup_{\theta_1} S_j^k(\theta_1)$$

For the basic example, the second maximization is explicit :

$$\begin{aligned} g_k &= \max_{1 \le j \le k} \sum_{i=j}^{k} \left[\frac{\hat{\nu}_j (y_i - \mu_0)}{\sigma^2} - \frac{\hat{\nu}_j^2}{2\sigma^2} \right] \\ \hat{\nu}_j &= \frac{1}{k-j+1} \sum_{i=j}^{k} (y_i - \mu_0) \end{aligned}$$

Off-line Change Detection

Off-line hypotheses testing

$$\Lambda_N = \max_{1 \le j \le N} \Lambda_1^N(j)$$

$$\tilde{\Lambda}_N = \max_{1 \le j \le N} \sup_{\tilde{\theta}_0} \sup_{\theta_0} \sup_{\theta_1} \Lambda_1^N(j, \tilde{\theta}_0, \theta_0, \theta_1)$$

Off-line estimation

$$\hat{t}_0 = \arg \max_{1 \le k \le N} \sum_{i=k}^{N} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)}$$

3

Background on Probability and System Theory

In this chapter and the next, we provide the reader with the theoretical background of this book. The present chapter discusses basic results from probability and system theories, while chapter 4 is devoted to the statistical background necessary for the design and performance evaluation of change detection algorithms.

Results from probability theory are presented in section 3.1. We recall the definition and main properties of conditional probability and expectation, Markov chains, and martingales. An optional stopping theorem for martingales is reported and some properties of Brownian motion and diffusion processes between boundaries are described. These results are useful for estimating properties of stopping times, and thus of change detection algorithms. In section 3.2, we report some key results from system theory, which are used mainly in chapters 7 and 9. We investigate observers, the Kalman filter, and the connection between state-space and ARMA models. We give further notes and bibliographical references on all these topics in section 3.3.

3.1 Some Results from Probability Theory

This section presents results from probability theory which we use throughout the book. We first recall the definition and basic properties of conditional probability and expectation, and Markov chains. Then we report a stopping formula for martingales which is useful for deriving properties of some stopping times. After that, we investigate different types of boundaries and first crossing problems for both Brownian motion and diffusion processes between boundaries. These results are necessary for deriving both optimal stopping times and approximations for evaluation of their performance.

3.1.1 Notation and Main Definitions

Let us first introduce some notation and definitions. Let $(\Omega, \mathcal{B}, \mathbf{P})$ be a *probability space*, where Ω is an abstract space (the sample space), $\mathcal{B}$ is a sigma algebra of subsets of Ω (the event space), and $\mathbf{P}$ is a probability measure defined over all members of $\mathcal{B}$. Two events, B_1 and B_2 in $\mathcal{B}$, are *independent* if

$$\mathbf{P}(B_1 \cap B_2) = \mathbf{P}(B_1)\mathbf{P}(B_2) \tag{3.1.1}$$

3.1.1.1 Scalar Random Variables and Distributions

A real-valued *random variable* Y is a measurable function $\rho : \Omega \to \mathbf{R}$, i.e., a function such that, for any $B \in \mathcal{B}(\mathbf{R})$, $\rho^{-1}(B) \in \mathcal{B}$. Here, $\mathcal{B}(\mathbf{R})$ is a particular set of subsets of $\mathbf{R}$, more precisely the sigma field of

Borel sets of $\mathbf{R}$. The *distribution* of Y is the probability measure $\mathbf{P}_Y$ defined by

$$\mathbf{P}_Y(B) = \mathbf{P}[Y^{-1}(B)] \tag{3.1.2}$$

for $B \in \mathcal{B}(\mathbf{R})$. For abbreviation, we write

$$\mathcal{L}(Y) = \mathbf{P}_Y \tag{3.1.3}$$

If the range space of the random variable Y is discrete, then p_Y defined by

$$p_Y(y) = \mathbf{P}_Y(\{y\}), \quad \sum_y p_Y(y) = 1 \tag{3.1.4}$$

is the *probability mass function* or *pmf* associated with the distribution $\mathbf{P}_Y$. The following relation holds :

$$\mathbf{P}_Y(B) = \sum_{x \in B} p_Y(x), \quad B \in \mathcal{B}(\text{range space}) \tag{3.1.5}$$

When the range space of the random variable Y is continuous, Y is said to have a *probability density function* or *pdf* f_Y with respect to a probability measure μ, when the distribution $\mathbf{P}_Y$ can be written as

$$\mathbf{P}_Y(B) = \int_{x \in B} f_Y(x)\, d\mu(x), \quad B \in \mathcal{B}(\text{range space}) \tag{3.1.6}$$

Note that the density f_Y satisfies

$$\int_{\mathbf{R}} f_Y(x)\, d\mu(x) = 1 \tag{3.1.7}$$

and is only defined up to an almost sure equality with respect to μ. In the case of a density with respect to Lebesgue measure on the real line, we have

$$\mathbf{P}_Y(B) = \int_{x \in B} f_Y(x)\, dx, \quad B \in \mathcal{B}(\text{range space}) \tag{3.1.8}$$

In other words, we have the analog of the previous formula for the pmf. From now on, we consider only this case of Lebesgue measure. If a real random variable Y has a density f_Y with respect to Lebesgue measure, then a statistic $S(Y)$ has a density that is given by the transformation lemma, as we show later.

The *cumulative distribution function* or *cdf*, is defined by

$$F_Y(y) = \mathbf{P}_Y((-\infty, y]) \tag{3.1.9}$$

If the distribution $\mathbf{P}_Y$ has a probability density function f_Y, we have

$$F_Y(y) = \int_{-\infty}^{y} f_Y(x)\, dx \tag{3.1.10}$$

In the present chapter, we use the symbol Y for a scalar or vector random variable, and the symbol y for the argument of the pdf and cdf. We do *not* distinguish between the random variable Y and its actual value when it is observed. In subsequent chapters, we distinguish between a vector random variable Y and a scalar random variable y, while keeping the notation y for the argument of the pdf and cdf. This should not introduce any confusion.

The space $(\mathbf{R}, \mathcal{B}(\mathbf{R}), \mathbf{P}_Y)$ is also a probability space, and is often the useful canonical description of the random variable. In what follows, the subscript Y is often omitted. Furthermore, most of the time we deal with *parametric distributions* and densities, and we use the notation $\mathbf{P}_\theta$ for distributions and p_θ or f_θ for densities. We now introduce an important concept that is used throughout the book.

Definition 3.1.1 (Likelihood function). *The* likelihood function *of one* observation Y *is equal to the probability density* $p_\theta(Y)$ *of the underlying random variable. It should be clear that the likelihood function is in fact a function of the parameter* θ.

The relevance of this concept for change detection lies in the fact that this function selects the most likely values of the parameter θ and can be used for deciding between several possible values for hypotheses testing, as we explain in the next chapter.

The following parametric family of distributions plays an important part in mathematical statistics, as we show in the three next sections.

Example 3.1.1 (Exponential family). *The Koopman-Darmois* exponential family *of distributions plays an important role in mathematical statistics. These densities have the following form :*

$$f_\theta(y) = h(y)e^{c(\theta)T(y)-d(\theta)} \tag{3.1.11}$$

where all the functions on the right side are finite and measurable. Such an expression is not affected by one-to-one transformations of the variable or the parameter. If c *and* T *are monotonic,* $\tilde{\theta} = c(\theta)$ *and* $\tilde{y} = T(y)$ *lead to a density* $f_{\tilde{\theta}}(\tilde{y})$ *of the above form, with* c *replaced by the identity function [Cox and Hinkley, 1986]. Such a parameter is called a* natural parameter.

The *expectation* of a continuous random variable Y which has a pdf f_Y is defined by

$$\mathbf{E}(Y) = \int_{\mathbf{R}} y\, f_Y(y)\, dy \tag{3.1.12}$$

when this integral exists. The expectation of a discrete random variable with pmf p_Y is defined in a similar manner, with the integral replaced by a sum. The moments of order $k > 1$ are defined in a similar way using the successive powers of the random variable. The *variance* of a continuous random variable Y is the second-order moment of the centered variable $Y - \mathbf{E}(Y)$:

$$\mathrm{var}(Y) = \int_{\mathbf{R}} [y - \mathbf{E}(Y)]^2\, f_Y(y)\, dy \tag{3.1.13}$$

Example 3.1.2 (Gaussian distribution). *The density and cumulative distribution function of the scalar* Gaussian distribution $\mathcal{N}(0,1)$ *with mean* 0 *and variance* 1 *are denoted by* φ *and* ϕ*, respectively :*

$$\begin{aligned} \varphi(y) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \\ \phi(y) &= \int_{-\infty}^{y} \varphi(x)\, dx \end{aligned} \tag{3.1.14}$$

The density of the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ *with mean* μ *and variance* σ^2 *is* $\frac{1}{\sigma}\varphi(\frac{y-\mu}{\sigma})$.

The *characteristic function* of a random variable is

$$\Phi_Y(t) = \mathbf{E}(e^{itY}) \tag{3.1.15}$$

for real t. When the distribution of Y has a density f, it becomes

$$\Phi_Y(t) = \int_{-\infty}^{\infty} e^{ity} f(y)\, dy \tag{3.1.16}$$

Under some regularity conditions, the moments of the random variable can be obtained by derivating this characteristic function :

$$\mathbf{E}(Y^k) = \frac{1}{i^k}\Phi_Y^{(k)}(0) \tag{3.1.17}$$

More generally, the *moment generating function* (mgf) of a random variable Y with density f is defined as the two-sided *Laplace transform* of the pdf [Cox and Miller, 1965, Feller, 1966, Ghosh, 1970] :

$$\psi_Y(\varsigma) = \mathbf{E}(e^{-\varsigma Y}) = \int_{-\infty}^{\infty} e^{-\varsigma y} f(y)\, dy \tag{3.1.18}$$

for any complex number ς. When ς is purely imaginary, $\varsigma = it$, the mgf reduces to the characteristic function $\Phi_Y(-t)$. The Laplace transform is also characteristic of the distribution and, for some processes, can be easily computed.

The following distributions play an important role in hypotheses testing, as we show in chapter 4.

Example 3.1.3 (Gamma distributions). *For $a > 0$ and $b > 0$, the law $\gamma(a, b)$ has the following density :*

$$\frac{1}{\Gamma(a)} b^a e^{-by} y^{a-1} \mathbf{1}_{\{y>0\}} \tag{3.1.19}$$

where

$$\Gamma(a) = \int_0^{\infty} x^{a-1} e^{-x}\, dx \tag{3.1.20}$$

For nonzero integer values,

$$\Gamma(n) = (n-1)! \tag{3.1.21}$$

The Laplace transform of this distribution is

$$\left(\frac{b}{\varsigma + b}\right)^a \tag{3.1.22}$$

for $\varsigma > -b$. Its mean is $\frac{a}{b}$ and its variance is $\frac{a}{b^2}$.

For $a > 0$, $b > 0$ and $c \geq 0$, we define the law $\gamma(a, c, b)$ by

$$\gamma(a, c, b) = \sum_{i=0}^{\infty} \frac{e^{-c} c^i}{i!} \gamma(a + i, b) \tag{3.1.23}$$

Its Laplace transform is

$$\left(\frac{b}{\varsigma + b}\right)^a e^{-\frac{c\varsigma}{\varsigma + b}} \tag{3.1.24}$$

for $\varsigma > 0$. The mean and variance of this distribution are $\frac{a+c}{b}$ and $\frac{a+2c}{b^2}$, respectively.

One of the most important uses of the gamma distributions concerns the so-called χ^2 distributions, which are a central issue in statistical hypotheses testing, as we show in section 4.2.

Example 3.1.4 (χ^2 distributions). *If Y is distributed as $\mathcal{N}(0, \sigma^2)$, the Laplace transform of the distribution of Y^2 is $\frac{1}{\sqrt{2\varsigma\sigma^2+1}}$ and Y^2 is distributed according to the law $\gamma(\frac{1}{2}, \frac{1}{2\sigma^2})$.*

If $Y_1, \ldots, Y_n$ are independent and distributed as $\mathcal{N}(0, 1)$, then the law of $\xi = Y_1^2 + \ldots + Y_n^2$ is denoted by $\chi^2(n)$ and called χ^2 with n degrees of freedom. *Its Laplace transform is*

$$\left(\frac{1}{2\varsigma + 1}\right)^{\frac{n}{2}} \tag{3.1.25}$$

It is a law $\gamma(\frac{n}{2}, \frac{1}{2})$ *and thus it has mean* n *and variance* $2n$.

When the Y_i *have mean* μ_i *and common variance* 1*, their sum of squares* ξ *is said to have a law* $\chi'^2(n, \lambda)$*, with mean*

$$m = n + \lambda \tag{3.1.26}$$

where $\lambda = \sum_{i=1}^{n} \mu_i^2$ *is called* noncentrality parameter*, and with variance*

$$\sigma^2 = 2n + 4\lambda \tag{3.1.27}$$

The density of this distribution is $\gamma(\frac{n}{2}, \frac{\lambda}{2}, \frac{1}{2})$*. Its Laplace transform is*

$$\psi_\xi(\varsigma) = \left(\frac{1}{2\varsigma + 1}\right)^{\frac{n}{2}} e^{-\frac{\lambda\varsigma}{2\varsigma+1}} \tag{3.1.28}$$

The following alternative and useful expressions of the densities of a central and a noncentral χ^2 *distributions are used in chapters 4 and 7. The density of the law* $\gamma(\frac{n}{2}, \frac{1}{2})$ *can be written as [Ghosh, 1970]*

$$p_0(y) = \frac{y^{\frac{n}{2}-1} e^{-\frac{y}{2}}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \tag{3.1.29}$$

Similarly the density of the law $\gamma(\frac{n}{2}, \frac{\lambda}{2}, \frac{1}{2})$ *can be written as [Ghosh, 1970]*

$$p_\lambda(y) = p_0(y)\ e^{-\frac{\lambda}{2}}\ \ G\left(\frac{n}{2}, \frac{\lambda y}{4}\right) \tag{3.1.30}$$

where G *is the hypergeometric function :*

$$G(a, y) = \sum_{i=0}^{\infty} \frac{\Gamma(a) y^i}{\Gamma(a + i) i!} \tag{3.1.31}$$

The following convergence result is useful in chapter 7. When the noncentrality parameter λ *goes to infinity, the distribution* $\chi'^2(n, \lambda)$ *is asymptotically equivalent to the Gaussian distribution* $\mathcal{N}(m, \sigma^2)$*, where* m *and* σ *are defined in (3.1.26)-(3.1.27). The proof of this fact is based upon the Laplace transform of the distribution of* $\xi' = \frac{\xi - m}{\sigma}$*, which is* $e^{-\frac{m}{\sigma}} \psi_\xi(\frac{\zeta}{\sigma})$ *and which converges to* $e^{-\frac{\zeta^2}{2}}$ *when* $\lambda \to \infty$.

Finally, we define the following extremum.

Definition 3.1.2 (Essential supremum). *Let* $(Y_i)_{i \in I}$ *be a family of real-valued random variables on* $(\Omega, \mathcal{B}, \mathbf{P})$*, bounded by another variable. We say that* Y *is an* essential supremum *for* $(Y_i)_{i \in I}$*, and we write* $Y = \operatorname{ess\,sup}_I Y_i$*, if*

$$(\forall i \in I)\ \ Y_i \leq Z\ \ \mathbf{P}\text{-almost surely} \ \ \Leftrightarrow \ \ Y \leq Z\ \ \mathbf{P}\text{-almost surely} \tag{3.1.32}$$

If I *is countable, then* $\operatorname{ess\,sup}_I Y_i = \sup_I Y_i$.

3.1.1.2 Vector Random Variables

A *random vector* Y of dimension r is a finite collection of r random variables $(Y_1, ..., Y_r)$. The cdf and pdf are defined as before :

$$\mathbf{P}_Y(B) = \int_B dF_Y(x), \quad B \in \mathcal{B}(\mathbf{R}^r) \tag{3.1.33}$$

and are referred to by the term *joint distribution*. The *marginal distribution* of each random variable is defined by

$$\mathbf{P}_{Y_i}(B) = \mathbf{P}[Y_i^{-1}(B)], \quad B \in \mathcal{B}(\mathbf{R}) \tag{3.1.34}$$

When probability densities exist, they are connected through relations of the following type :

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\, dy_2 \tag{3.1.35}$$

Example 3.1.5 (Gaussian vector distributions). *A random vector Y of dimension r is said to be Gaussian if any linear combination of its components is Gaussian. The Laplace transform of a Gaussian vector distribution with mean μ and covariance matrix Σ is*

$$e^{-\varsigma^T \mu + \frac{1}{2}\varsigma^T \Sigma \varsigma} \tag{3.1.36}$$

When Σ is positive definite, the Gaussian distribution has the following probability density :

$$\varphi_{\mu,\Sigma}(y) = \frac{1}{\sqrt{(2\pi)^r (\det \Sigma)}} e^{-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)} \tag{3.1.37}$$

and the log-likelihood function of an observation Y_k is

$$-\frac{r}{2}\ln(2\pi) - \frac{1}{2}\ln(\det \Sigma) - \frac{1}{2}(Y_k - \mu)^T \Sigma^{-1}(Y_k - \mu) \tag{3.1.38}$$

In this case, we use the following notation for the cumulative distribution function :

$$\Phi_{\mu,\Sigma}(B) = \int_B \varphi_{\mu,\Sigma}(y)\, dy, \quad B \in \mathcal{B}(\mathbf{R}^r) \tag{3.1.39}$$

When Σ is degenerated with rank $\tilde{r} < r$, let D be the diagonal matrix filled with the $\tilde{r}$ nonzero eigenvalues of Σ, and A be the matrix of size $r \times \tilde{r}$ filled with the corresponding eigenvectors. We have

$$\begin{aligned} \Sigma &= ADA^T \\ Y &= \mu + AD^{\frac{1}{2}}X \end{aligned} \tag{3.1.40}$$

where X is a normalized Gaussian random vector of size $\tilde{r}$. In this case, the log-likelihood function of an observation Y_k is

$$-\frac{r}{2}\ln(2\pi) - \frac{1}{2}\ln(\det D) - \frac{1}{2}(Y_k - \mu)^T AD^{-1}A^T(Y_k - \mu) \tag{3.1.41}$$

The marginal *distribution of a Gaussian distribution is also Gaussian. More precisely, if Y, μ and Σ are partitioned as*

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{3.1.42}$$

then Y_1 is a Gaussian vector with mean μ_1 and covariance Σ_{11}. If Y is an r-dimensional vector such that $\mathcal{L}(Y) = \mathcal{N}(\mu, \Sigma)$, then $(Y - \mu)^T \Sigma^{-1}(Y - \mu)$ is distributed as a $\chi^2(r)$.

Definition 3.1.3 (Independent variables). *Random variables $Y_1, \ldots, Y_n$ are said to be* independent *if their joint distribution is the product of the n marginal distributions, that is, if*

$$\mathbf{P}(Y_1 \in B_1, \ldots, Y_n \in B_n) = \prod_{i=1}^{n} \mathbf{P}(Y_i \in B_i), \quad B_i \in \mathcal{B}(\mathbf{R}) \tag{3.1.43}$$

If the n marginal distributions are all identical, the variables are said to be independent identically distributed, *abbreviated as* i.i.d.

We often make use of the following lemma.

Lemma 3.1.1 (Transformation). *Let Y be a vector* random variable *taking its values in an open set A of $\mathbf{R}^r$, and with density f_Y with respect to Lebesgue measure. Let ρ be a diffeomorphism from A into an open set B of $\mathbf{R}^r$, and note J_ρ its Jacobian matrix. Then the densities of Y and $\rho(Y)$ are related through*

$$f_Y = (\det J_\rho) f_{\rho(Y)} \circ \rho \tag{3.1.44}$$

where $\circ$ denotes the composition of functions. If Y is a random process, this lemma is valid for the joint distribution of a sample of size N.

A typical example of the use of this lemma concerns the computation of the probability density of a statistic $S(Y)$ as a function of the density of Y. Another particularly useful application of this transformation lemma is shown next when computing the likelihood function in terms of the innovations.

3.1.1.3 Random Processes

A *random process* is an indexed family $(Y_t)_{t\in I}$ of random variables, which may be discrete time if I is a set of integers, or continuous time if I is a set of real numbers. This book is mainly devoted to discrete time random processes or time series or signals. But continuous time random processes are often useful for deriving approximations for performance evaluation of the change detection algorithms.

Let $(Y_i)_{i\geq 1}$ be an r-dimensional random process with mean value μ_i. The covariance R_{ij} between the variables Y_i and Y_j is

$$R_{ij} = \mathbf{E}\left[(Y_i - \mu_i)(Y_j - \mu_j)^T\right] \tag{3.1.45}$$

The *covariance function* of the process $(Y_i)_i$ is R_{ij} considered as a function of the time instant i and lag $i - j$.

Example 3.1.6 (Gaussian process). *Assume that $(Y_i)_{1\leq i\leq n}$ is a random process made of n r-dimensional Gaussian vectors with laws $\mathcal{N}(\mu_i, \Sigma_i)$. Then the joint distribution of the Y_i's is $\mathcal{N}(\mu, \Sigma)$, where*

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \tag{3.1.46}$$

and

$$\Sigma = \begin{pmatrix} \Sigma_1 & \ldots & R_{1j} & \ldots & R_{1n} \\ R_{21} & \Sigma_2 & \ldots & \ldots & R_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ R_{n1} & \ldots & R_{nj} & \ldots & \Sigma_n \end{pmatrix} \tag{3.1.47}$$

When the random variables Y_i are independent, this reduces to

$$\Sigma = \operatorname{diag}(\Sigma_i) \tag{3.1.48}$$

We now investigate a very important particular class of random processes.

Definition 3.1.4 (Stationary process). *A discrete time random process $(Y_n)_n$ is said to be* stationary *if, for every k and N, the distribution of $Y_{k+1}, Y_{k+2}, \ldots, Y_{k+N}$ is the same as the distribution of $Y_1, Y_2, \ldots, Y_N$. If $(Y_n)_n$ is Gaussian, the time invariance of the mean and of the covariance function is a sufficient condition for stationarity.*

The *covariance* function of a stationary random process Y with mean μ is denoted by

$$R_l^Y = \mathbf{E}\left[(Y_k - \mu)(Y_{k-l} - \mu)^T\right] \tag{3.1.49}$$

and depends only upon the time lag.

The *power spectrum* or *spectral density* of the process Y is the Fourier transform of the covariance sequence :

$$\Phi_Y(\omega) = \sum_{l=-\infty}^{l=+\infty} R_l^Y e^{-il\omega} \tag{3.1.50}$$

where ω is in $(-\pi, \pi)$. The subscript Y is omitted when no confusion is possible. The power spectrum of a state-space model representation and the power spectrum of an ARMA process are given in section 3.2, in formulas (3.2.4) and (3.2.36), respectively.

Definition 3.1.5 (Exit time). *Let $(Y_t)_t$ be any random process, and assume $Y_0 = y$. Let $T_{-a,b}$ be the first time at which the process reaches $-a$ or b, where $-a < y < b$. The instant $T_{-a,b}$ is the first* exit time *from the interval $(-a, b)$. Similarly, we note T_b the first time at which the process reaches b.*

3.1.2 Conditional Probability and Expectation

We now define the conditional probability and conditional expectation which are of interest for nonindependent random variables or events, and which are necessary for the definition of transition probabilities of Markov chains, and for the formal definition of the delay for detection. This concept is also useful in the framework of filtering and estimation, as we show in section 3.2.

3.1.2.1 Conditional Probability

Let $(\Omega, \mathcal{B}, \mathbf{P})$ be a probability space and A be an event such that $\mathbf{P}(A) > 0$. Then the *conditional probability* of any event B given the event A is defined by

$$\mathbf{P}(B|A) = \frac{\mathbf{P}(B \cap A)}{\mathbf{P}(A)} \tag{3.1.51}$$

This extends easily to conditioning by random variables taking only countably many values. In that case, the *conditional cdf* of a random variable X given another random variable Y is defined by

$$F_{X|Y}(x|y) = \frac{\mathbf{P}(X \leq x, Y = y)}{\mathbf{P}(Y = y)} \tag{3.1.52}$$

when $\mathbf{P}(Y = y) > 0$.

For conditioning by random variables with noncountably many values, conditional distributions are naturally defined as *random variables* [Breiman, 1968], and thus up to a set of $\mathbf{P}$-probability zero. These random variables are often called *determinations of the conditional distribution*.

When Y and X are jointly distributed continuous random variables, the conditional distribution function is

$$F_{X|Y}(x|y) = \frac{\int_{-\infty}^{x} f_{Y,X}(y,z)\,dz}{f_Y(y)} \tag{3.1.53}$$

The following property is known under the name of *law of total probability* :

$$F_X(x) = \int_{-\infty}^{+\infty} F_{X|Y}(x|y) f_Y(y)\,dy \tag{3.1.54}$$

Definition 3.1.6 (Conditional density). *When the two random variables are jointly continuously distributed, the* conditional density function *is defined by*

$$f_{X|Y}(x|y) = \frac{f_{Y,X}(y,x)}{f_Y(y)} \tag{3.1.55}$$

As we said before for ordinary distributions and densities, for simplicity we often omit the subscripts Y, X and $X|Y$ when no confusion is possible.

The main use we make of this formula in this book is in giving the joint probability density function of a sequence of identically distributed observations $(Y_k)_{1\leq k\leq n}$:

$$f(y_1,\ldots,y_n) = f(y_1) \prod_{k=2}^{n} f(y_k|\mathcal{Y}_1^{k-1}) \tag{3.1.56}$$

In the case of an *independent* sequence (i.i.d.), this reduces to

$$f(y_1,\ldots,y_n) = \prod_{k=1}^{n} f(y_k) \tag{3.1.57}$$

When dealing with parametric densities $f_\theta(y)$ for the random variable Y, the Bayes approach consists of considering the parameter θ as being a value of a random variable Θ with density $f_\Theta(\theta)$, often called the *prior* distribution. Let us write the parametric density of the observations as $f_{Y|\Theta}(y|\theta)$. For inferring or testing about the value of Θ realized in the available observation, it is of interest to consider the conditional density of Θ given $Y = y$, called the *posterior* distribution. We make use of the following well-known result [De Groot, 1970, Cox and Hinkley, 1986] :

Theorem 3.1.1 (Bayes rule). *The posterior distribution is given by*

$$f_{\Theta|Y}(\theta|y) = \frac{f_{Y|\Theta}(y|\theta) f_\Theta(\theta)}{\int_\Omega f_{Y|\Theta}(y|\theta') f_\Theta(\theta') d\theta'} \tag{3.1.58}$$

3.1.2.2 Conditional Expectation

The conditional expectation can be defined with respect to the conditional probability in the same way that the ordinary expectation is defined with respect to the probability. More precisely, let X be a random variable with finite expectation, and Y be a random variable or vector.

Definition 3.1.7 (Conditional expectation). *The* conditional expectation *of X given Y*

$$\mathbf{E}(X|y) = \mathbf{E}(X|Y = y) \tag{3.1.59}$$

is defined by

$$\mathbf{E}(X|y) = \int x f_{X|Y}(x|y)\, dx \tag{3.1.60}$$

if X is continuous, and by

$$\mathbf{E}(X|y) = \sum_x x f_{X|Y}(x|y) \tag{3.1.61}$$

if X is discrete.

If we note $g(y) = \mathbf{E}(X|Y = y)$, then $\mathbf{E}(X|Y)$ can be defined as $g(Y)$, and thus conditional expectations are random variables [Breiman, 1968].

As a particular case, we can define the conditional expectation of a random variable Y with respect to an event of the sigma algebra $\mathcal{B}$ defined by Y in the following manner :

$$\mathbf{E}(Y|a < Y \leq b) = \frac{\int_a^b y f_Y(y)\, dy}{\int_a^b f_Y(y)\, dy} \tag{3.1.62}$$

The main useful properties of conditional expectations are

$$\mathbf{E}(X) = \mathbf{E}\left[\mathbf{E}(X|Y)\right] \tag{3.1.63}$$

and

$$\mathbf{E}\left[h_1(Y)h_2(X,Y)\right] = \mathbf{E}\left[h_1(Y)\mathbf{E}(h_2(X,Y)|Y)\right] \tag{3.1.64}$$

for any bounded functions h_1 and h_2.

The conditional expectation is necessary, for example, for defining the delay for detection of change detection algorithms, as explained in section 4.4 of the next chapter, and for the derivation of the Kalman filter as recalled in section 3.2. The property given in the following example is useful for this latter issue.

Example 3.1.7 (Gaussian distributions - contd.). *The conditional distribution of a Gaussian vector is also Gaussian. Keeping the notation of example 3.1.5, the conditional distribution of Y_1 given $Y_2 = y_2$ is Gaussian with mean $\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2)$.*

We now use the conditional expectation for introducing two key concepts, namely the innovation and the residual, and for computing the likelihood function.

Definition 3.1.8 (Innovation and residual in a random process). *Let $(Y_k)_{k\geq 1}$ be a random process with distribution $\mathbf{P}_\theta$. The* innovation *ε_k is*

$$\varepsilon_k = Y_k - \mathbf{E}_\theta(Y_k|\mathcal{Y}_1^{k-1}) \tag{3.1.65}$$

where $\mathcal{Y}_1^{k-1}$ is the vector made of the past observations. Let us note

$$\hat{Y}_k = \mathbf{E}_\theta(Y_k|\mathcal{Y}_1^{k-1}) \tag{3.1.66}$$

Because of (3.1.64), for all $i > 0$, ε_k is uncorrelated with Y_{k-i}, in the sense that

$$\mathbf{E}_\theta(\varepsilon_k Y_{k-i}) = 0 \tag{3.1.67}$$

Consequently, we also have

$$\mathbf{E}_\theta(\varepsilon_k \hat{Y}_k) = 0 \tag{3.1.68}$$

Furthermore, for the same reason, the innovation sequence is an uncorrelated *sequence, namely,*

$$\mathbf{E}_\theta(\varepsilon_k \varepsilon_{k-i}) = 0 \tag{3.1.69}$$

If Y_k is Gaussian, these orthogonality properties are equivalent to independence.

For $\tilde{\theta}$ possibly different from θ, we define the residual e_k *by*

$$e_k = Y_k - \mathbf{E}_{\tilde{\theta}}(Y_k | \mathcal{Y}_1^{k-1}) \tag{3.1.70}$$

For $\tilde{\theta} = \theta$, the residual is nothing but the innovation.

The innovations play a key role when computing the likelihood function, as we show now.

Example 3.1.8 (Likelihood of a Gaussian process). *Let us discuss the likelihood of a Gaussian random process. Because of the definition of the conditional density, the log-likelihood is given by*

$$l_N = \ln p_Y(Y_1, \ldots, Y_N) = \sum_{k=1}^{N} \ln p_Y(Y_k | Y_{k-1}, \ldots, Y_1) \tag{3.1.71}$$

The transformation lemma (3.1.44) implies that

$$l_N = \sum_{k=1}^{N} \ln p_\varepsilon(\varepsilon_k | Y_{k-1}, \ldots, Y_1) \tag{3.1.72}$$

This comes from the result of example 3.1.7 concerning the conditional distribution of a Gaussian process : In this case, the Jacobian matrix corresponding to the transformation from the observation Y_k to the innovation ε_k has its determinant obviously equal to 1. The independence properties of the innovation sequence in the Gaussian case results in

$$l_N = \sum_{k=1}^{N} \ln p_\varepsilon(\varepsilon_k) \tag{3.1.73}$$

In the particular case of a process modeled by a regression model, namely,

$$Y_k = HX_k + V_k \tag{3.1.74}$$

where X is an unknown vector of size n and V a white noise sequence of size $r > n$, the likelihood is as in (3.1.73) where ε_k is replaced by the residual e_k of the least-squares estimation of X.

3.1.2.3 Markov Chains

Definition 3.1.9 (Markov process). *A* Markov process *is a random process such that the following property holds: Given the present value, the future values of the process are independent of the past values. More precisely [Feller, 1966], for all finite collection $Y_{t_1}, \ldots, Y_{t_l}$ and for all $k \leq l$, the following property holds:*

$$\mathbf{P}(Y_{t_k} \in B | Y_{t_1}, \ldots, Y_{t_{k-1}}) = \mathbf{P}(Y_{t_k} \in B | Y_{t_{k-1}}) \tag{3.1.75}$$

The structure of a Markov process basically depends upon the *transition probability function* :

$$P_{s,t}(B|y) = \mathbf{P}(Y_t \in B | Y_s = y) \tag{3.1.76}$$

for $s < t$.

A *Markov chain* is a discrete time Markov process with discrete range space - or state space. The state space is usually labeled with the aid of the set of natural integers. The transition probability function of a stationary Markov chain is defined by the probabilities $p(i|j)$ of transition from state j to state i. A Markov chain is often used in Bayesian approaches to model the transition behavior of a dynamic process between the normal functioning mode and different abnormal ones. An elementary example is considered in section 2.3.

A *Markov process of order* p is such that

$$\mathbf{P}(Y_{t_k} \in B|Y_{t_{k-1}}, \ldots, Y_{t_1}) = \mathbf{P}(Y_{t_k} \in B|Y_{t_{k-1}}, \ldots, Y_{t_{k-p}}) \tag{3.1.77}$$

Of course, a Markov process of order 1 is what we called a Markov process before. An AR(p) process is a Markov process of order p. The corresponding state introduced in subsection 3.2.4, made of the past p observations, is a Markov process (of order 1). Similarly, the state underlying an ARMA(p, q) process - which has dimension $\tilde{p} = \max\{p, q + 1\}$ - is a Markov process (of order 1), but the ARMA process itself is not a Markov process of any finite order; it can be viewed only as a Markov process with infinite order. A more detailed investigation of this question can be found in [Benveniste *et al*., 1990].

3.1.3 Martingales and Stopping Times

In this subsection, we define martingales and give an important stopping formula for them, which is useful for deriving properties of exit times for Brownian motion and diffusion processes, and consequently for evaluating the performances of some change detection algorithms, as we explain in chapter 5. Let us first consider *discrete time processes*.

Definition 3.1.10 (Martingale). *Let* $(S_n)_{n\geq 1}$ *and* $(Y_n)_{n\geq 1}$ *be two stochastic processes.* $(S_n)_n$ *is said to be a* martingale *with respect to* $(Y_n)_n$ *if* $\mathbf{E}|S_n| < \infty$ *and*

$$\mathbf{E}(S_{n+1}|Y_1, \ldots, Y_n) = S_n \tag{3.1.78}$$

for $n \geq 1$. *If* $S_n = Y_n$, *then* $(Y_n)_n$ *is simply said to be a martingale*.

By the law of total probability and by induction, we have

$$\mathbf{E}(S_k) = \mathbf{E}(S_1) \tag{3.1.79}$$

and

$$\mathbf{E}(S_{n+k}|Y_1, \ldots, Y_n) = S_n \tag{3.1.80}$$

for all $k \geq 1$.

The following concept is also useful for the investigation of the properties of the change detection algorithms.

Definition 3.1.11 (Submartingale). *Let* $(S_n)_{n\geq 1}$ *and* $(Y_n)_{n\geq 1}$ *be two stochastic processes.* $(S_n)_n$ *is said to be a* submartingale *with respect to* $(Y_n)_n$ *if* $\mathbf{E}|S_n| < \infty$ *and*

$$\mathbf{E}(S_{n+1}|Y_1, \ldots, Y_n) \geq S_n \tag{3.1.81}$$

for $n \geq 1$.

In the case of *continuous time random processes* $(S_t)_t$ and $(Y_t)_t$, a martingale $(S_t)_t$ can be defined in a similar way in terms of dependency with respect to the trajectory of the random process $(Y_t)_t$, or more precisely to the sigma algebra generated by $(Y_\tau)_{\tau \leq t}$. The condition (3.1.78) then becomes

$$\mathbf{E}[S_t | (Y_\tau)_{\tau < t}] = S_t \tag{3.1.82}$$

Examples of martingales that are of interest within the framework of this book are the following.

Example 3.1.9 (Cumulative sum and likelihood ratio). *The cumulative sum*

$$S_n = \sum_{k=1}^{n} Y_k \tag{3.1.83}$$

of independent and zero mean random variables $(Y_k)_{k \geq 1}$ is a martingale with respect to $(Y_n)_n$. Note that here we use the abbreviation S_n for the quantity S_1^n which is introduced in chapter 2. If these random variables are identically distributed with probability density f_0, then, for any density f, the likelihood ratio

$$\Lambda_n = \frac{\prod_{k=1}^{n} f(Y_k)}{\prod_{k=1}^{n} f_0(Y_k)} = \prod_{k=1}^{n} \frac{f(Y_k)}{f_0(Y_k)} \tag{3.1.84}$$

is a martingale with respect to $(Y_n)_n$. We will see later that the normalized Brownian motion $(Y_t)_{t \geq 0}$ is a martingale, and that several functions of Y_t, which are martingales with respect to $(Y_t)_t$, are useful for the investigation of exit times.

We now define stopping times and give a useful stopping formula for martingales. Again we begin with *discrete time* random processes.

Definition 3.1.12 (Stopping time). *A random variable T is called a* stopping time *- or Markov time - with respect to a process $(Y_n)_{n \geq 1}$ if T takes only integer values and if, for every $n \geq 1$, the event $\{T = n\}$ is determined by $(Y_1, \ldots, Y_n)$. In other words,*

$$\mathbf{1}_{\{T=n\}} = \mathbf{1}_{\{T=n\}}(Y_1, \ldots, Y_n) \tag{3.1.85}$$

For example, for any fixed k, the kth time at which the process $(Y_t)_t$ visits a set A is a stopping time. Note that the *last* visit time is not a stopping time.

By the law of total probability, it can be shown that, if $(Y_n)_n$ is a martingale and T is a stopping time, then for all $n \geq k$,

$$\mathbf{E}(Y_n \mathbf{1}_{\{T=k\}}) = \mathbf{E}(Y_k \mathbf{1}_{\{T=k\}}) \tag{3.1.86}$$

Theorem 3.1.2 (Optional stopping theorem). *Let $(Y_n)_{n \geq 1}$ be a martingale and T a stopping time. If*

$$\begin{aligned} \mathbf{P}(T < \infty) &= 1 \\ \mathbf{E}(|Y_T|) &< \infty \\ \lim_{n \to \infty} \mathbf{E}(Y_n \mathbf{1}_{\{T > n\}}) &= 0 \end{aligned}$$

then

$$\mathbf{E}(Y_T) = \mathbf{E}(Y_1) \tag{3.1.87}$$

A stopping time with respect to a *continuous time* random process can be defined in a similar way : A positive random variable T is a stopping time if, for every positive t, the event $\{T \leq t\}$ is in the sigma algebra generated by $(Y_\tau)_{\tau \leq t}$ [Breiman, 1968]. The optional stopping theorem can then be stated in a similar way and can be used for deriving Wald's fundamental identity and the Laplace transform of the distribution of some stopping times and maximum values related to processes such as Brownian motion and diffusions, as we show in the next subsection.

3.1.4 Some Results for Brownian Motion and Diffusion Processes

We now consider *continuous time* stochastic processes, for which we use the notation $(Y_t)_{t\geq 0}$. We concentrate on Brownian motion and diffusion processes, for which we give some properties of exit times which will be used for getting approximations of performances of several change detection algorithms.

3.1.4.1 Brownian Motion with Boundaries

Brownian motion can be seen as the limit of cumulative sums of independent, identically distributed random variables. A precise statement of this limiting property in terms of a representation and a convergence theorem for cumulative sums can be found in [Breiman, 1968, Billingsley, 1968]. Because of this, Brownian motion is of key interest in this book because of the central role played by sequential hypotheses testing and CUSUM algorithm for change detection. Moreover, Brownian motion can be seen as the limit of some martingales [P.Hall and Heyde, 1980]. This fact is of crucial interest for deriving asymptotic expansions in sequential analysis and analyzing cumulative sum with a random number of increments.

Definition 3.1.13 (Brownian motion). Brownian motion *is a continuous time process* $(Y_t)_{t\geq 0}$ *starting from* 0*, with Gaussian stationary independent increments :*

$$\mathbf{P}(Y_t - Y_s \leq y) = \frac{1}{\sigma\sqrt{2\pi(t-s)}} \int_{-\infty}^{y} e^{-\frac{(u-\mu(t-s))^2}{2\sigma^2(t-s)}}\, du \tag{3.1.88}$$

for $t > s$*. If* $\mu \neq 0$*,* $(Y_t)_t$ *is said to be Brownian motion with drift* μ*. If* $\mu = 0$ *and* $\sigma^2 = 1$*,* $(Y_t)_t$ *is called* normalized *Brownian motion*.

Let $p_t^0(dy|y)$ be its transition probability. Note that the property of being a normalized Brownian motion is invariant with respect to the following transformations [Breiman, 1968] : symmetry, change in time origin, time inversion, scale change, and time reversal.

For any $t_n > t_{n-1} > t_0 \geq 0$, the random variables $Y_{t_n}, \ldots, Y_{t_0}$ have a joint normal distribution with mean $\mathbf{E}(Y_{t_n}) = \mu t_n$ and covariances

$$\Sigma(t_j, t_k) = \mathbf{E}(Y_{t_j} - \mu t_j)(Y_{t_k} - \mu t_k) = \sigma^2 \min(t_j, t_k) \tag{3.1.89}$$

If $(Y_t)_t$ is a normalized Brownian motion, let us define

$$U_t = Y_t^2 - t \tag{3.1.90}$$

$$V_t = e^{\lambda Y_t - \frac{1}{2}\lambda^2 t} \tag{3.1.91}$$

where λ is any real constant. Then the three processes $(Y_t)_t$, $(U_t)_t$, and $(V_t)_t$ are also martingales with respect to $(Y_t)_t$.

Now we investigate the probability of reaching a boundary and the mean time before reaching this boundary. We use the notation for exit time that we introduced in subsection 3.1.1. We first consider two parallel horizontal boundaries, and then other boundaries, straight or not. We distinguish between absorbing and reflecting boundaries, which can be intuitively defined as follows. An *absorbing* boundary is a boundary that the process does not leave after having reached it. A *reflecting* boundary is a boundary that the process can leave after having reached it, but only on the same side (namely without crossing) and possibly after a finite sojourn in this state.

Horizontal absorbing boundaries Consider first the case of a *normalized* Brownian motion ($\mu = 0$) starting from zero. Let $-a < 0 < b$. Then, using the fact that $(Y_t)_t$ and $(Y_t^2 - t)_t$ are martingales and the optional stopping theorem, it is easy to show [Breiman, 1968] that

$$\begin{aligned} \mathbf{P}(Y_{T_{-a,b}} = -a | Y_0 = 0) &= \frac{b}{b+a} \\ \mathbf{P}(Y_{T_{-a,b}} = b | Y_0 = 0) &= \frac{a}{b+a} \\ \mathbf{E}(T_{-a,b} | Y_0 = 0) &= ab \end{aligned} \tag{3.1.92}$$

If $Y_0 = y \neq 0$, using the invariance with respect to a change in the time origin, we find

$$\begin{aligned} \mathbf{P}(Y_{T_{-a,b}} = -a | Y_0 = y) &= \frac{b-y}{b+a} \\ \mathbf{P}(Y_{T_{-a,b}} = b | Y_0 = y) &= \frac{y+a}{b+a} \\ \mathbf{E}(T_{-a,b} | Y_0 = y) &= (y+a)(b-y) \end{aligned} \tag{3.1.93}$$

Coming back to the case $Y_0 = 0$ and using the above mentioned transformation invariance properties of Brownian motion and their effect on the Laplace transform of the distribution of exit times, it is possible to show [Breiman, 1968] that the stopping time $T_{-a,b}$ satisfies

$$\mathbf{E}\left(e^{-\varsigma T_{-a,b}} \mathbf{1}_{\{Y_{T_{-a,b}} = b\}}\right) = \frac{\sinh(\sqrt{2\varsigma}a)}{\sinh\sqrt{2\varsigma}(b+a)} \tag{3.1.94}$$

Similarly,

$$\mathbf{E}\left(e^{-\varsigma T_{-a,b}} \mathbf{1}_{\{Y_{T_{-a,b}} = -a\}}\right) = \frac{\sinh(\sqrt{2\varsigma}b)}{\sinh\sqrt{2\varsigma}(b+a)} \tag{3.1.95}$$

The sum of these two last quantities is nothing but the Laplace transform $\mathbf{E}(e^{-\varsigma T_{-a,b}})$. This transform contains complete information concerning the probability distribution of the stopping time $T_{-a,b}$, and thus it can be used for computing the performances of CUSUM-type algorithms [Basseville, 1978, Basseville, 1981] in terms of expectations and variances of this stopping time.

For a *general* Brownian motion starting from $Y_0 = y \neq 0$, namely when $\mu \neq 0$ and $\sigma^2 \neq 1$, the probability that the process reaches the level $b > y$ before hitting $-a < y$ is [Karlin and Taylor, 1975]

$$\mathbf{P}(Y_{T_{-a,b}} = b | Y_0 = y) = \frac{e^{-2\gamma y} - e^{2\gamma a}}{e^{-2\gamma b} - e^{2\gamma a}} \tag{3.1.96}$$

where $\gamma = \frac{\mu}{\sigma^2}$. Using the optional stopping theorem for the martingale $(Y_t - \mu t - Y_0)_t$, we get the conditional expectation of $T_{-a,b}$:

$$\mathbf{E}(T_{-a,b} | Y_0 = y) = \frac{1}{\mu}\left(b\,\frac{e^{-2\gamma y} - e^{2\gamma a}}{e^{-2\gamma b} - e^{2\gamma a}} - a\,\frac{e^{-2\gamma b} - e^{-2\gamma y}}{e^{-2\gamma b} - e^{2\gamma a}} - y\right) \tag{3.1.97}$$

When $y = 0$, this reduces to

$$\mathbf{P}(Y_{T_{-a,b}} = b | Y_0 = 0) = \frac{1 - e^{2\gamma a}}{e^{-2\gamma b} - e^{2\gamma a}} \tag{3.1.98}$$

which in turn results in (3.1.92) in the limit case $\mu = 0$. When $y = 0$, the expectation of $T_{-a,b}$ reduces to

$$\mathbf{E}(T_{-a,b} | Y_0 = 0) = \frac{1}{\mu}\left(b\,\frac{1 - e^{2\gamma a}}{e^{-2\gamma b} - e^{2\gamma a}} - a\,\frac{e^{-2\gamma b} - 1}{e^{-2\gamma b} - e^{2\gamma a}}\right) \tag{3.1.99}$$

Furthermore, $T_{-a,b}$ satisfies [Breiman, 1968, Karlin and Taylor, 1975, Taylor, 1975]

$$\mathbf{E}\left(e^{-\varsigma T_{-a,b}}\,\mathbf{1}_{\{Y_{T_{-a,b}}=b\}}\right) = e^{\gamma b}\frac{\sinh(\delta a)}{\sinh\delta(b+a)} \tag{3.1.100}$$

where $\delta = \sqrt{\gamma^2 + \frac{2\varsigma}{\sigma^2}}$.

One absorbing and one reflecting boundary Consider now the process $g_t = M_t - Y_t$, where M_t is the maximum $\max_{s\leq t} Y_s$ of the Brownian motion $(Y_t)_t$. For investigating the properties of some *alarm times* (or exit times or boundary crossing), such as the CUSUM stopping time,

$$t_a = \inf\{t : g_t \geq h\} \tag{3.1.101}$$

some martingale properties are useful. More precisely, exponential types of martingales are often used for evaluating the performances of sequential detection algorithms through Laplace transforms of the distribution of stopping times, as we show now.

If $(Y_t)_{t\geq 0}$ is a Brownian motion with drift μ and variance σ^2, let

$$V_t = e^{\varsigma Y_t - (\varsigma\mu + \frac{1}{2}\varsigma^2\sigma^2)t} \tag{3.1.102}$$

where ς is any real number. Then $(V_t)_t$ is a martingale with respect to $(Y_t)_t$. Another martingale of interest with respect to this point of view is $(W_t)_t$ where [Kennedy, 1976]

$$W_t = \left[\zeta\cosh(\zeta g_t) - (\varsigma + \frac{\mu}{\sigma^2})\sinh(\zeta g_t)\right] e^{\varsigma M_t + \frac{\mu}{\sigma^2}g_t - \frac{\sigma^2}{2}t(\zeta^2 - \frac{\mu^2}{\sigma^4})} \tag{3.1.103}$$

and where ς, ζ are any real numbers. The application of the optional stopping theorem to these two martingales, using the stopping time of the CUSUM algorithm given before, provides us with the Laplace transforms of the distributions of the pairs (t_a, Y_{t_a}) and (t_a, M_{t_a}) respectively, which are a way of characterizing the properties of the change detection algorithm. For example, the optional stopping theorem applied to the second martingale results in

$$\left[\zeta\cosh(\zeta h) - (\varsigma + \frac{\mu}{\sigma^2})\sinh(\zeta h)\right]\mathbf{E}\left(e^{\varsigma M_{t_a} + \frac{\mu}{\sigma^2}h - \frac{\sigma^2}{2}t_a(\zeta^2 - \frac{\mu^2}{\sigma^4})}\right) = \zeta \tag{3.1.104}$$

From this we deduce in particular

$$\mathbf{E}(t_a) = \frac{1}{\mu}\left(\frac{e^{-\frac{2\mu h}{\sigma^2}} - 1}{2\frac{\mu}{\sigma^2}} + h\right) \tag{3.1.105}$$

This result is used in [R.Johnson and Bagshaw, 1974, Bagshaw and R.Johnson, 1975a, Reynolds, 1975].

Other boundaries Consider first a linear boundary. Assume that $(Y_t)_{t\geq t_1}$ is a normalized Brownian motion that starts from zero at time t_1, and let $a \geq 0$ and $b \geq 0$ be two positive numbers. The probability that a sample path Y_t crosses the line $h(t) = a + b(t - t_1)$ for some $t \in (t_1, t_2)$ is as follows. Define the crossing time :

$$T_h = \inf\{t : Y_t \geq h(t)\} \tag{3.1.106}$$

The above-mentioned probability is [Durbin, 1971, Karlin and Taylor, 1975]

$$\mathbf{P}(t_1 < T_h < t_2) = (1 + e^{-2ab}) - \left\{\phi\left[\frac{a + b(t_2 - t_1)}{\sqrt{t_2 - t_1}}\right] + e^{-2ab}\phi\left[\frac{a - b(t_2 - t_1)}{\sqrt{t_2 - t_1}}\right]\right\} \tag{3.1.107}$$

where ϕ is the cumulative distribution function of the normalized Gaussian distribution defined in (3.1.14). When t_2 goes to infinity, the probability that a sample path Y_t crosses the line $y = a + b(t - t_1)$ for some $t \geq t_1$ is [Breiman, 1968]

$$\mathbf{P}(T_h < \infty) = e^{-2ab} \tag{3.1.108}$$

which is known as the Doob formula.

Moreover, for $b \leq 0$, the crossing time T_h has the following distribution function [Lorden, 1973] :

$$\mathbf{P}(t_1 < T_h < t) = \int_{t_1}^{t} \frac{a}{[2\pi(\tau - t_1)^3]^{\frac{1}{2}}} \exp\left\{-\frac{[a + b(\tau - t_1)]^2}{2(\tau - t_1)}\right\} d\tau \tag{3.1.109}$$

Now, when the boundary $h(t)$ is no longer a straight line, but a concave and piecewise continuously differentiable function on $(0, \infty)$, then the distribution of the stopping time is [Lorden, 1973]

$$\mathbf{P}(t_1 < T_h < t) = \int_{t_1}^{t} \frac{h(\tau) - \tau h'(\tau)}{(2\pi\tau^3)^{\frac{1}{2}}} e^{-\frac{h^2(\tau)}{2\tau}} d\tau \tag{3.1.110}$$

for $0 \leq t_1 < t \leq \infty$.

3.1.4.2 Diffusion Processes with Boundaries

A diffusion process is a Markov process for which all sample paths $(Y_t)_{t \geq 0}$ are continuous functions. In some sense, a diffusion process has drift and variance parameters at time t, which do depend upon the value Y_t. Brownian motion is a special type of diffusion process, with constant parameters. An Ornstein-Uhlenbeck process is a Brownian motion submitted to an elastic force [Feller, 1966]. The investigation of first crossing probabilities and expectations can be done with the aid of the Ito formula for stochastic differential equations. For example, it is possible to compute $\mathbf{P}(Y_{T_{ab}} = a | Y_0 = y)$ and $\mathbf{E}(e^{\varsigma T_{ab}} | Y_{T_{ab}} = a, Y_0 = y)$ (see [Lehoczky, 1977, Basseville, 1978]).

Note that there is no martingale formula for deriving the Laplace transform of the distribution of T_{ab} and the maximum of the process [Basseville, 1978], in opposition to the case of Brownian motion [Kennedy, 1976]. The main case in chapter 5 where a diffusion process is necessary and Brownian motion not sufficient concerns the investigation of the properties of the geometric moving average algorithm, as we discuss in section 5.1.

3.2 Some Results from System Theory

In this section, we report several results about estimation and identification in state-space models that are useful for the investigation of two different approaches for the design of change detection algorithms : the *statistical approach* based on the Kalman filter and likelihood techniques, and the *geometrical approach* based upon various observers and analytical redundancy relationships. We also recall the equivalence between state-space and ARMA models, which is used in sections 7.4 and 9.3. In many aspects of the following four subsections, we follow [Goodwin and Sin, 1984].

3.2.1 State-Space Models

A linear time invariant (LTI) stochastic state-space model in discrete time has the following form :

$$\left\{ \begin{array}{rcl} X_{k+1} & = & FX_k + GU_k + W_k \\ Y_k & = & HX_k + JU_k + V_k \end{array} \right. \tag{3.2.1}$$

where

X, U, Y are the state, input, and observation vectors, having dimensions n, m, r, respectively,

$(W_k)_k$ and $(V_k)_k$ are two independent Gaussian white noise sequences, having covariance matrices Q and R, respectively,

the initial state X_0 has mean x_0 and covariance P_0, and is independent of the two noise sequences $(W_k)_k$ and $(V_k)_k$,

F is the state transition matrix, H the observation matrix, and G and J the control matrices.

Such a model is a Markov model, namely the pair (X_{k+1}, Y_k) is a Markov process in the sense defined in section 3.1.

Using the forward shift operator z, namely $zY_k = Y_{k+1}$, (3.2.1) can be rewritten as

$$Y_k = \left[H(zI_n - F)^{-1}G + J\right] U_k + H(zI_n - F)^{-1}W_k + V_k \tag{3.2.2}$$

Thus, let

$$\begin{aligned} \mathcal{T}(z) &= \begin{bmatrix} \mathcal{T}_U(z) & \mathcal{T}_W(z) & \mathcal{T}_V(z) \end{bmatrix} \\ &= \begin{bmatrix} H(zI_n - F)^{-1}G + J & H(zI_n - F)^{-1} & I_r \end{bmatrix} \end{aligned} \tag{3.2.3}$$

be the transfer function of this system [Ljung, 1987]. Note that we consider here two different types of transfer functions : the input–output transfer function $\mathcal{T}_U(z)$, and the noise–output transfer functions $\mathcal{T}_W(z)$ and $\mathcal{T}_V(z)$. The input–output transfer function $\mathcal{T}_U(z)$ is said to be *proper* when the degree of the numerator is less than or equal to the degree of the denominator. It is said to be *stable* when all its poles are inside the unit circle. The power spectrum of the output observations Y is

$$\Phi_Y(\omega) = \mathcal{T}_U(e^{i\omega})\ \Phi_U(\omega)\ \mathcal{T}_U^T(e^{-i\omega}) + \mathcal{T}_W(e^{i\omega})\ Q\ \mathcal{T}_W^T(e^{-i\omega}) + R \tag{3.2.4}$$

where $\Phi_U(\omega)$ is the power spectrum of the possibly deterministic input U.

We recall now some basic deterministic notions that are useful in the subsequent chapters, especially chapter 7. In the deterministic case, the noises W_k and V_k in (3.2.1) do not exist.

Definition 3.2.1 *A state x of this system is said to be* controllable *if there exists a time instant n and an input sequence $(U_k)_{0 \le k \le n-1}$ that drives the system from $X_0 = x$ to $X_n = 0$. The system is* completely controllable *if every state is controllable.*

A state x is said to be unobservable *if for any $n > 0$ and $U_k = 0,\ 0 \le k \le n$, the initial state $X_0 = x$ results in a zero output $Y_k = 0,\ 0 \le k \le n$. The system is said to be* completely observable *if no state (except 0) is unobservable.*

The system is said to be stabilizable *if all uncontrollable modes have corresponding eigenvalues strictly inside the unit circle.*

The system is said to be detectable *if all unobservable modes have corresponding eigenvalues strictly inside the unit circle.*

We now define two important matrices.

Definition 3.2.2 (Controllability and observability matrices). *The* controllability matrix *of the system (3.2.1) is*

$$\mathcal{C}_n(F, G) = \begin{pmatrix} G & FG & F^2G & \dots & F^{n-1}G \end{pmatrix} \tag{3.2.5}$$

The observability matrix *of the system (3.2.1) is*

$$\mathcal{O}_n(H,F) = \begin{pmatrix} H \\ HF \\ HF^2 \\ \vdots \\ HF^{n-1} \end{pmatrix} \tag{3.2.6}$$

When no confusion is possible, we simply write $\mathcal{C}_n$ and $\mathcal{O}_n$. The observability index *is the rank of the observability matrix.*

The following results do hold.

Theorem 3.2.1 (Complete controllability and observability). *The system (3.2.1) of order n is*

- *completely controllable if rank $\mathcal{C}_n = n$; the condition is necessary if F is nonsingular;*
- *completely observable if and only if rank $\mathcal{O}_n = n$; this condition is equivalent to the following : the spectral density $\mathcal{T}_W(z)\mathcal{T}_W^T(\frac{1}{z})$ is invertible [Kailath, 1980, Caines, 1988].*

Considering the ranks of the controllability and observability matrices, it is possible to find linear transformations of the state-space that result in completely observable subsystems corresponding to part of the transformed state-space model. For completely controllable systems, it is possible to define the *controllability form* (the *controller form*) which has block upper triangular state transition matrix F with 1 on the first lower diagonal and nonzero coefficients on the last column (first row) of each block. Similarly, for completely observable systems, it is possible to define the *observability form* (the *observer form*) which has block upper triangular state transition matrix F with 1 on the first upper diagonal and nonzero coefficients on the last raw (first column) of each block. The interested reader is referred to [Kailath, 1980, Goodwin and Sin, 1984] for further details.

Finally, let us introduce two definitions of stability which will be useful later.

Definition 3.2.3 (Stability). *A state x_e is said to be an* equilibrium state *of the deterministic part of the system (3.2.1) if $x_e = Fx_e$ (for $U_k = 0$).*

An equilibrium state x_e is said to be stable *if for any k_0 and $\epsilon > 0$, there exists $\delta(\epsilon, k_0)$ such that $\|X_{k_0} - x_e\| < \delta \Rightarrow \|X_k - x_e\| < \epsilon \ \forall k \geq k_0$.*

An equilibrium state x_e is said to be asymptotically stable *if it is stable and if for any k_0 there exists $\delta(k_0)$ such that $\|X_{k_0} - x_e\| < \delta \Rightarrow \lim_{k\to\infty} \|X_k - x_e\| = 0$. The linearity of the system implies that if one equilibrium state is (asymptotically) stable, then all other equilibrium states are (asymptotically) stable, and thus these definitions of stability are also definitions of stability for the system.*

An important result [Aström and Wittenmark, 1984] is the following.

Theorem 3.2.2 (Stability). *A discrete-time, linear, time-invariant system is asymptotically stable if and only if all eigenvalues of F are strictly inside the unit circle, or equivalently if the poles of the transfer function are also inside the unit circle. This is the link between stability and stabilizability defined before.*

Other possible definitions of stability are investigated in [Goodwin and Sin, 1984, Aström and Wittenmark, 1984].

3.2.2 Observers

Observers are dynamic systems that are aimed at the reconstruction of the state X of a state-space model (3.2.1) and that work on the basis of the measured inputs U and outputs Y.

3.2.2.1 Direct Estimation

Consider first a direct computation [Aström and Wittenmark, 1984]. A repeated use of the deterministic part of the equations (3.2.1) leads to the following :

$$\begin{pmatrix} Y_{k-n+1} \\ Y_{k-n+2} \\ \vdots \\ Y_k \end{pmatrix} = \mathcal{O}_n(H,F)X_{k-n+1} + \mathcal{J}_n(G,J) \begin{pmatrix} U_{k-n+1} \\ U_{k-n+2} \\ \vdots \\ U_k \end{pmatrix} \tag{3.2.7}$$

where $\mathcal{O}_n(H,F)$ is as defined in (3.2.6) and

$$\mathcal{J}_n(G,J) = \begin{pmatrix} J & \dots & \dots & \dots & \dots & \dots \\ HG & J & \dots & 0 & \dots & \dots \\ HFG & HG & J & \dots & \dots & \dots \\ HF^2G & HFG & HG & J & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ HF^{n-2}G & \dots & \dots & HFG & HG & J \end{pmatrix} \tag{3.2.8}$$

is the block Toeplitz matrix associated with the impulse response of (3.2.1). This can be rewritten as

$$\mathcal{Y}_{k-n+1}^k = \mathcal{O}_n(H,F)X_{k-n+1} + \mathcal{J}_n(G,J)\,\mathcal{U}_{k-n+1}^k \tag{3.2.9}$$

where

$$\mathcal{Y}_{k-n+1}^k = \begin{pmatrix} Y_{k-n+1} \\ Y_{k-n+2} \\ \vdots \\ Y_k \end{pmatrix} \text{ and } \mathcal{U}_{k-n+1}^k = \begin{pmatrix} U_{k-n+1} \\ U_{k-n+2} \\ \vdots \\ U_k \end{pmatrix} \tag{3.2.10}$$

If the system (3.2.1) is observable, we can compute

$$\tilde{\mathcal{O}}_n(H,F) = \left[\mathcal{O}_n^T(H,F)\mathcal{O}_n(H,F)\right]^{-1}\ \mathcal{O}_n^T(H,F) \tag{3.2.11}$$

Then X_k can be estimated by

$$X_{k-n+1} = \tilde{\mathcal{O}}_n(H,F)\mathcal{Y}_{k-n+1}^k - \tilde{\mathcal{O}}_n(H,F)\mathcal{J}_n(G,J)\,\mathcal{U}_{k-n+1}^k \tag{3.2.12}$$

and (3.2.1) can be used again, leading to

$$\begin{aligned} X_k &= F^{n-1}\tilde{\mathcal{O}}_n^T(H,F)\mathcal{Y}_{k-n+1}^k \\ &\quad + \left[\begin{pmatrix} F^{n-2}G & \dots & FG & G & 0 \end{pmatrix} - F^{n-1}\tilde{\mathcal{O}}_n(H,F)\mathcal{J}_n(G,J)\right]\,\mathcal{U}_{k-n+1}^k \end{aligned} \tag{3.2.13}$$

This observer is sometimes called a *deadbeat observer* for a completely observable system. The advantage is that the state variable X can be computed in at most n steps. The drawback is that it may be sensitive to disturbances.

3.2.2.2 Luenberger Observers

Alternative solutions for state estimation do make use of the state transition equation in (3.2.1) and the difference between the measured and estimated outputs Y, in the following manner.

Definition 3.2.4 *An* observer *is a dynamic system of the form*

$$\hat{X}_{k+1|k} = F\hat{X}_{k|k-1} + GU_k + K\left(Y_k - H\hat{X}_{k|k-1} - JU_k\right) \tag{3.2.14}$$

where $\hat{X}_{k|k-1}$ is an estimate, or prediction, of X_k given the measurements $\mathcal{Y}_1^{k-1}$ available up to time $k-1$. The state estimation error $\tilde{X}_{k+1|k} = X_{k+1} - \hat{X}_{k+1|k}$ is given by

$$\tilde{X}_{k+1|k} = (F - KH)\tilde{X}_{k|k-1} \tag{3.2.15}$$

If K is chosen such that the system (3.2.15) is asymptotically stable, then the reconstruction error converges to zero, even if the system (3.2.1) is not stable.

It is possible to avoid the delay in (3.2.14) by using the following observer :

$$\hat{X}_{k|k} = F\hat{X}_{k-1|k-1} + GU_{k-1} + K\left[Y_k - H\left(F\hat{X}_{k-1|k-1} + GU_{k-1}\right) - JU_k\right] \tag{3.2.16}$$

The estimation error is then

$$\tilde{X}_{k|k} = (I_n - KH)F\tilde{X}_{k-1|k-1} \tag{3.2.17}$$

Because the pair (F, HF) is observable if the pair (F, H) is detectable, it results that $F - KHF$ can be given arbitrary eigenvalues by selecting K. If these eigenvalues are chosen inside the unit circle, a zero estimation error can be achieved. Furthermore,

$$Y_k - H\hat{X}_{k|k} - JU_k = (I_r - HK)HF\tilde{X}_{k-1|k-1} \tag{3.2.18}$$

If rank$(H) = r$, then K may be chosen such that $HK = I_r$, so that the outputs can be estimated without error. Thus, it is possible to eliminate r equations from (3.2.16). The resulting reduced order observer is called a *Luenberger observer*. The choice of the matrix K can be shown to be dual to a problem of feedback pole placement, and thus this choice is simple if the system is in observable form [Aström and Wittenmark, 1984].

One key issue concerning the *residuals* (3.2.18) is that, in case of noise disturbances W_k and V_k, they *are not an independent sequence*, except if the observer is in fact the Kalman filter - viewed as a full order state observer.

We see in chapter 7 that observers are a useful way to obtain analytical redundancy relationships, which are one of the geometrical approaches for change detection.

3.2.3 Kalman Filter

The Kalman filter provides one particular estimate of the state X_k of the system (3.2.1). This filter gives the *minimum variance estimate* of the state, namely the *conditional mean* of X_k given the past observations $Y_{k-1}, Y_{k-2}, \ldots$, which we denote also by $\hat{X}_{k|k-1}$ and call *one-step ahead prediction*. When the Gaussian assumption concerning the noises is removed, the Kalman filter gives the linear minimum variance estimate of the state, namely the smallest unconditional error covariance among all linear estimates, but, in general, this estimate is not the conditional mean [Goodwin and Sin, 1984].

3.2.3.1 Kalman Filter, Innovation Model, Whitening Filter

The one-step ahead prediction, the estimated state, and the innovation ε_k are given by

$$\left\{ \begin{array}{rcrcl} \hat{X}_{k+1|k} & = & F\hat{X}_{k|k} & + & GU_k \\ \hat{X}_{k|k} & = & \hat{X}_{k|k-1} & + & K_k\varepsilon_k \\ \varepsilon_k & = & Y_k & - & H\hat{X}_{k|k-1} - JU_k \end{array} \right. \tag{3.2.19}$$

where K_k is the Kalman gain. In other words, the Kalman filter is of the form (3.2.14), where the gain K is chosen as FK_k. The Kalman gain K_k, the state estimation error covariance $P_{k|k-1}$, and the covariance of the innovation Σ_k are given by

$$\begin{array}{rcl} K_k & = & P_{k|k-1}H^T\Sigma_k^{-1} \\ P_{k+1|k} & = & FP_{k|k}F^T + Q \\ P_{k|k} & = & (I_n - K_kH)P_{k|k-1} \\ \Sigma_k & = & HP_{k|k-1}H^T + R \end{array} \tag{3.2.20}$$

These formulas can be proven by induction, using the joint conditional distribution of $(X_{k+1}^T, Y_k^T)^T$, given $\mathcal{Y}_1^{k-1}$ and the result about the transformation of Gaussian random variables given in example 3.1.7. According to definition 3.1.8, the innovation is orthogonal to all past observations, that is,

$$\mathbf{E}\left(\varepsilon_k|\mathcal{Y}_1^{k-1}\right) = 0 \tag{3.2.21}$$

The Kalman filter equations (3.2.19) can be rewritten as

$$\left\{ \begin{array}{rcrcrcl} \hat{X}_{k+1|k} & = & F\hat{X}_{k|k-1} & + & GU_k & + & FK_k\varepsilon_k \\ Y_k & = & H\hat{X}_{k|k-1} & + & JU_k & + & \varepsilon_k \end{array} \right. \tag{3.2.22}$$

which is called the *innovation model*. This in turn can be rewritten as the following *whitening filter* :

$$\left\{ \begin{array}{rcrcrcl} \hat{X}_{k+1|k} & = & F(I_n - K_kH)\hat{X}_{k|k-1} & + & GU_k & + & FK_kY_k \\ \varepsilon_k & = & -H\hat{X}_{k|k-1} & - & JU_k & + & Y_k \end{array} \right. \tag{3.2.23}$$

The design of the Kalman gain K is known to be related to the problem of spectral factorization of the power spectrum of the observations Y [B.Anderson and Moore, 1979], as in

$$\begin{array}{r} \left[I_r + H(zI_n - F)^{-1}FK\right]\left(R + HPH^T\right)\left[I_r + K^TF^T(z^{-1}I_n - F^T)^{-1}H^T\right] \\ = R + H(zI_n - F)^{-1}Q(z^{-1}I_n - F^T)^{-1}H^T \end{array} \tag{3.2.24}$$

3.2.3.2 Stability of the Kalman Filter

The asymptotic time invariance and stability of the Kalman filter are important for applications, of course, and also are used in the computation of the likelihood function of observations modeled by state-space models (3.2.1). We thus briefly recall a useful stability result [Goodwin and Sin, 1984].

Because of (3.2.23) and (3.2.20), the Kalman filter can be summarized as

$$\begin{array}{rcl} \hat{X}_{k+1|k} & = & \bar{F}_k\hat{X}_{k|k-1} + GU_k + FK_kY_k \\ \bar{F}_k & = & F(I_n - K_kH) \\ K_k & = & P_{k|k-1}H^T(HP_{k|k-1}H^T + R)^{-1} \\ P_{k+1|k} & = & FP_{k|k-1}F^T - FP_{k|k-1}H^T(HP_{k|k-1}H^T + R)^{-1}HP_{k|k-1}F^T + Q \\ P_{1|0} & = & P_0 \end{array} \tag{3.2.25}$$

Assuming that the error covariance $P_{k|k-1}$ converges to a steady-state value P (conditions are given later), it results from (3.2.25) that P is solution of the so-called *algebraic Riccati equation* :

$$P - FPF^T + FPH^T(HPH^T + R)^{-1}HPF^T - Q = 0 \tag{3.2.26}$$

We assume here that $HPH^T + R$ is invertible; sufficient conditions for this are R nonsingular, or P nonsingular and H full rank. In this case, the filter state transition matrix $\bar{F}_k$ and the Kalman gain K_k also converge to steady-state values $\bar{F}$ and K, given by

$$\begin{aligned} \bar{F} &= F(I_n - KH) \\ K &= PH^T(HPH^T + R)^{-1} \end{aligned} \tag{3.2.27}$$

Note that the following identity holds [B.Anderson and Moore, 1979] :

$$\left[I_r + H(zI_n - F)^{-1}FK\right]^{-1} = I_r - H(zI_n - \bar{F})^{-1}FK \tag{3.2.28}$$

which is useful when discussing detectability conditions in chapter 7.

A real symmetric positive semidefinite solution of the algebraic Riccati equation (3.2.26) is said to be a *strong solution* if the corresponding filter state transition matrix $\bar{F}$ has all its eigenvalues inside or on the unit circle. Now let P_s be the (unique) strong solution of (3.2.26), and let K_s and $\bar{F}_s$ be the corresponding steady-state filter gain and state transition matrix given by (3.2.27). The following result holds :

Theorem 3.2.3 (Stability). *Provided that* (H, F) *is observable and* $(P_0 - P_s) > 0$ *or* $P_0 = P_s$*, then*

$$\begin{aligned} \lim_{k\to\infty} P_{k|k-1} &= P_s \\ \lim_{k\to\infty} K_k &= K_s \\ \lim_{k\to\infty} \bar{F}_k &= \bar{F}_s \end{aligned} \tag{3.2.29}$$

Note that if $Q = DD^T$ and if (F, D) has uncontrollable modes on the unit circle, then $\bar{F}_s$ will have the same roots on the unit circle, but the convergence and stability of the Kalman filter are still ensured in this case.

3.2.3.3 Likelihood Function

Let us compute the log-likelihood function of a sequence of Gaussian observations $Y_1, \ldots, Y_N$ modeled by the state-space model (3.2.1). Since we discuss in example 3.1.8, the likelihood of the observations is the likelihood of the innovations. since the innovation sequence here is a Gaussian process, then from (3.1.37) we have

$$-2l_N = \sum_{k=1}^{N} \left[\ln\left(\det \Sigma_k\right) + \varepsilon_k \Sigma_k^{-1} \varepsilon_k^T\right] \tag{3.2.30}$$

which asymptotically for N large becomes

$$-2l_N = N \ln \det(HPH^T + R) + \text{trace}\left[(HPH^T + R)^{-1} \sum_{k=1}^{N} \varepsilon_k \varepsilon_k^T\right] \tag{3.2.31}$$

where P is the steady-state solution of (3.2.26).

3.2.4 Connections Between ARMA and State-Space Models

In this subsection, we investigate the equivalence between state-space models and ARMA representations. This equivalence is of interest from both identification and detection/diagnosis points of view, as we show in chapter 9. Actually, even if the final desired diagnosis is in terms of spectral properties, it can be useful to design and use detection/diagnosis parametric algorithms working in the time domain.

We call the *autoregressive moving average model with auxiliary input (ARMAX)* an input-output model of the form

$$A(z)Y_k = C(z)U_k + B(z)\varepsilon_k \tag{3.2.32}$$

where A, B, C are polynomial matrices in the backward shift operator z^{-1} :

$$\begin{aligned} A(z) &= A_0 - \sum_{i=1}^{p} A_i z^{-i} \\ B(z) &= \sum_{j=0}^{q} B_j z^{-j} \\ C(z) &= \sum_{\ell=1}^{l} C_\ell z^{-\ell} \end{aligned} \tag{3.2.33}$$

such that A has a nonsingular constant term A_0, and where $(\varepsilon_k)_k$ is a white noise sequence with covariance matrix R.

The *transfer function* representation of an ARMAX model is given by [Ljung, 1987]

$$Y_k = \mathcal{T}_U(z)U_k + \mathcal{T}_\varepsilon(z)\varepsilon_k \tag{3.2.34}$$

where

$$\begin{aligned} \mathcal{T}_U(z) &= A^{-1}(z)C(z) \\ \mathcal{T}_\varepsilon(z) &= A^{-1}(z)B(z) \end{aligned} \tag{3.2.35}$$

Because of (3.2.34), the *power spectrum* corresponding to an ARMAX model is defined in the same way as in (3.2.4) for a state-space model from the transfer function representation (3.2.3) :

$$\Phi_Y(\omega) = \mathcal{T}_U(e^{i\omega})\ \Phi_U(\omega)\ \mathcal{T}_U^T(e^{-i\omega}) + \mathcal{T}_\varepsilon(e^{i\omega})\ R\ \mathcal{T}_\varepsilon^T(e^{-i\omega}) \tag{3.2.36}$$

with $\mathcal{T}_U$ and $\mathcal{T}_\varepsilon$ defined in (3.2.35).

3.2.4.1 From State-Space Models to ARMA Models

Following [Goodwin and Sin, 1984], we show that a state-space model written in the innovation form (3.2.22) can be compacted into an ARMAX model. Considering first the single-input single-output case, and assuming that the state-space model (3.2.1) is in observer form, the innovations model (3.2.22) can be rewritten as

$$\left\{ \begin{array}{rcl} \hat{X}_{k+1|k} &=& \begin{pmatrix} a_1 & 1 & & \\ \vdots & \ddots & \ddots & \\ \vdots & & \ddots & 1 \\ a_n & & & 0 \end{pmatrix} \hat{X}_{k|k-1} + \begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix} U_k + \begin{pmatrix} \kappa_1(k) \\ \vdots \\ \kappa_n(k) \end{pmatrix} \varepsilon_k \\ Y_k &=& \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix} \hat{X}_{k|k-1} + JU_k + \varepsilon_k \end{array} \right. \tag{3.2.37}$$

Using successive substitutions, this can be rewritten as a time-varying ARMAX model :

$$A(z)Y_k = C(z)U_k + B(k,z)\varepsilon_k \tag{3.2.38}$$

where

$$\begin{array}{rcl} A(z) & = & 1 - a_1 z^{-1} - \cdots - a_n z^{-n} \\ C(z) & = & g_1 z^{-1} + \cdots + g_n z^{-n} + JA(z^{-1}) \\ B(k,z) & = & 1 + \left[\kappa_1(k-1) - a_1\right] z^{-1} + \cdots + \left[\kappa_n(k-n) - a_n\right] z^{-n} \end{array} \tag{3.2.39}$$

B is time-varying because the Kalman filter gain K_k is time-varying. But, under the conditions of the stability theorem, K and B are asymptotically constant. Thus,

$$A(z)Y_k \quad = \quad C(z)U_k + B(z)\varepsilon_k \tag{3.2.40}$$

$B(z)$ is the denominator polynomial matrix in the model giving ε_k in terms of Y_k. As is obvious from (3.2.23), this whitening filter has a state transition matrix $F(I_n - KH)$, which, from the stability theorem again, generally has eigenvalues inside or on the unit circle.

In the multiple-input multiple-output case, following [Akaike, 1974], we start from the innovations model :

$$\left\{ \begin{array}{rcl} X_{k+1} & = & FX_k + K\varepsilon_k \\ Y_k & = & HX_k \end{array} \right. \tag{3.2.41}$$

where the state X is of dimension n again. The ARMA representation of the output Y can be derived in the following way. Let

$$\det\left(\lambda I_n - F\right) = 1 - \sum_{l=1}^{n} a_l \lambda^{n-l} \tag{3.2.42}$$

be the characteristic polynomial of F. By the Cayley-Hamilton theorem, we have

$$F^n - \sum_{l=1}^{n} a_l F^{n-l} = 0 \tag{3.2.43}$$

On the other hand, from (3.2.41) we deduce

$$X_{k+l} = F^l X_k + F^{l-1} K\varepsilon_k + \cdots + K\varepsilon_{k+l-1} \tag{3.2.44}$$

namely,

$$X_{k+l} = F^l X_k + \sum_{i=0}^{l-1} F^{l-1-i} K\varepsilon_{k+i} \tag{3.2.45}$$

Therefore, Y_k has the following ARMA representation :

$$Y_{k+n} - \sum_{l=1}^{n} a_l Y_{k+n-l} = \sum_{l=0}^{n} B_l \varepsilon_{k+n-l} \tag{3.2.46}$$

where $B_0 = 0$ and

$$B_l = H\left(F^{l-1} - a_1 F^{l-2} - \cdots - a_l I_n\right) K \tag{3.2.47}$$

for $1 \leq l \leq n$. In (3.2.46), the autoregressive coefficients are scalars or equivalently diagonal matrices.

Note that if the innovations model (3.2.41) is replaced by

$$\left\{ \begin{array}{rcl} X_{k+1} & = & FX_k + K\varepsilon_k \\ Y_k & = & HX_k + \varepsilon_k \end{array} \right. \tag{3.2.48}$$

which is closer to (3.2.22), the corresponding ARMA model is given by (3.2.46), where $B_0 = a_0 I_r$ and $B_l = H\left(F^{l-1} - a_1 F^{l-2} - \cdots - a_l I_n\right) K - a_l I_r$ for $1 \leq l \leq n$.

If the state-space model contains inputs as in (3.2.1), then the corresponding ARMA model is in fact an ARMAX model, as shown before in the scalar case. This is a standard issue in deterministic systems [Goodwin and Sin, 1984].

Thus, any state-space model can be represented with the aid of an ARMA model.

3.2.4.2 From ARMA Models to State-Space Models

We now show the converse statement, namely that there exists a state-space representation of any ARMA model, which ends the proof of the equivalence between both representations. Here we follow [Akaike, 1974] again. Consider the ARMA model,

$$Y_k - \sum_{i=1}^{p} A_i Y_{k-i} = \sum_{j=0}^{q} B_j \varepsilon_{k-j} \tag{3.2.49}$$

where Y and ε are of dimension r, $B_0 = I_r$, and $(\varepsilon_k)_k$ is a white noise sequence with covariance matrix R. Assume that the two characteristic equations

$$\begin{aligned} \det\left(\lambda^p I_r - \sum_{i=1}^{p} \lambda^{p-i} A_i\right) &= 0 \\ \det\left(\sum_{j=0}^{q} \lambda^{q-j} B_j\right) &= 0 \end{aligned} \tag{3.2.50}$$

have zeroes outside the unit circle, in other words that the process Y is stable. This ensures that Y has the following Wold decomposition :

$$Y_k = \sum_{l=0}^{\infty} D_l \varepsilon_{k-l} \tag{3.2.51}$$

with $D_0 = I_r$, and ε_k is the innovation of Y_k at time k, namely $\varepsilon_k = Y_k - \hat{Y}_{k|k-1}$, where $\hat{Y}_{k|k-1}$ is the one-step ahead predictor of Y_k at time $k-1$. Considering projections on the past observations $\mathcal{Y}_1^k$, we have, with obvious notation, the following relationship

$$\hat{Y}_{k+l|k} - \sum_{i=1}^{p} A_i \hat{Y}_{k+l-i|k} = \sum_{j=0}^{q} B_j \varepsilon_{k+l-j|k} \tag{3.2.52}$$

where $\hat{Y}_{k+l|k} = Y_{k+l}$ for $l = 0, -1, \ldots$ and $\varepsilon_{k+l|k} = 0$ for $l = 0, 1, \ldots$. For $l \geq q+1$ the right side of (3.2.52) is zero. Thus, for any i, $\hat{Y}_{k+i|k}$ is a linear combination of $\hat{Y}_{k|k}, \hat{Y}_{k+1|k}, \ldots, \hat{Y}_{k+s-1|k}$ where $s = \max(p, q+1)$. For example,

$$\hat{Y}_{k+s|k} = \sum_{l=1}^{s} A_l \hat{Y}_{k+s-l|k} \tag{3.2.53}$$

where, from (3.2.49), $A_i = 0$ for $p+1 \leq i \leq s$. From (3.2.51), we get

$$\hat{Y}_{k+i+1|k+1} = \hat{Y}_{k+i+1|k} + D_i \varepsilon_{k+1} \tag{3.2.54}$$

From (3.2.53) and (3.2.54), we deduce that the vector

$$X_k = \begin{pmatrix} \hat{Y}_{k|k} \\ \hat{Y}_{k+1|k} \\ \vdots \\ \hat{Y}_{k+s-1|k} \end{pmatrix} \tag{3.2.55}$$

provides us with the state-space model

$$\left\{ \begin{array}{rcl} X_{k+1} & = & FX_k + K\varepsilon_k \\ Y_k & = & HX_k \end{array} \right. \tag{3.2.56}$$

where

$$F = \begin{pmatrix} 0 & I_r & 0 & \dots & 0 \\ 0 & 0 & I_r & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & I_r \\ A_s & A_{s-1} & A_{s-2} & \dots & A_1 \end{pmatrix} \tag{3.2.57}$$

$$K = \begin{pmatrix} D_0 \\ D_1 \\ \vdots \\ D_{s-2} \\ D_{s-1} \end{pmatrix} \tag{3.2.58}$$

$$H = \begin{pmatrix} I_r & 0 & 0 & \dots & 0 \end{pmatrix} \tag{3.2.59}$$

Note that the D_l in (3.2.51) and (3.2.58) are the impulse response matrices of the system (3.2.49) and can be obtained from the ARMA coefficients A_i and B_j in the following manner. For any matrix M, let $M(\tilde{\iota})$ be the ιth column of M. Then, for $l \geq 0$, $D_l(\tilde{\iota})$ obeys the following relation :

$$D_l(\tilde{\iota}) - \sum_{i=1}^{p} A_i D_{l-i}(\tilde{\iota}) = \sum_{j=0}^{q} B_j \Delta_{l-j}(\tilde{\iota}) \tag{3.2.60}$$

where $D_l(\tilde{\iota}) = 0$ for $l < 0$ and $\Delta_l = \mathbf{1}_{\{l=0\}} I_r$.

This completes the derivation of a state-space model (3.2.56) from an ARMA model (3.2.49). More thorough investigations of this question can be found in [Caines, 1988].

3.3 Notes and References

In this section, we give basic textbooks and important papers concerning the topics investigated in this chapter.

Section 3.1

The main textbooks in probability that we think are useful to the reader are [Loeve, 1964, Cox and Miller, 1965, Feller, 1966, Breiman, 1968, Karlin and Taylor, 1975, Shiryaev, 1984, Gray and Davisson, 1986]. More advanced books are [Billingsley, 1968, P.Hall and Heyde, 1980]. For the problem of boundary crossing, we refer to the books [Leadbetter *et al.*, 1983, Siegmund, 1985b] and to the papers [Robbins and Siegmund, 1970, Durbin, 1971, Blake and Lindsey, 1973, Lerche, 1980, Durbin, 1985].

Section 3.2

The textbooks related to the topics described in this subsection on system theory background are [Box and Jenkins, 1970, B.Anderson and Moore, 1979, Goodwin and Sin, 1984, Aström and Wittenmark, 1984, Ljung, 1987, Söderström and Stoïca, 1989]. More advanced books are [Kailath, 1980, Caines, 1988, Hannan and Deistler, 1988].

4

Statistical Background and Criteria

In this chapter, we provide the reader with the statistical backgrounds necessary for the design and analysis of change detection algorithms. Section 4.1 is devoted to results concerning statistical inference. We first introduce two basic definitions of information and their connection with sufficiency and efficiency, and report some results about estimation. Then in sections 4.2 and 4.3, we discuss in more detail the key issues of on-line change detection, namely hypotheses testing and sequential analysis. We also include another important tool for designing detection algorithms : the expansion of the likelihood ratio leading to the so-called asymptotic local approach.

In section 4.4, we give a formal definition of the criteria for the design and performance evaluation of change detection algorithms. (These criteria were informally introduced in section 1.4). We follow here the three formal problem statements given in subsection 1.1.2. Finally, we give further notes and bibliographical references on all these topics in section 4.5.

4.1 Statistical Inference and Information

This section, together with the two following sections, introduces the key elements of mathematical statistics that will be used throughout the book for the design and performance evaluation of change detection algorithms. The present section is devoted to statistical inference. We introduce the key concepts of sufficiency, efficiency, and information.

4.1.1 Sufficient Statistics

Let Y be a random variable with distribution $\mathbf{P}_\theta$ belonging to the parametric family $\mathcal{P} = \{\mathbf{P}_\theta\}$, where $\theta \in \mathbf{R}^\ell$. In this book, we are mainly interested in distributions for which a probability density function (pdf) f_θ exists. Assume that a sample of observations $\mathcal{Y}_1^N = (Y_1^T, \ldots, Y_N^T)^T$ is available. This sample will be the only source of information for all subsequent inferences about $\mathbf{P}_\theta$. One of the most important concepts in mathematical statistics is the concept of *sufficient statistics*, which was introduced in [Fisher, 1925]. Let $S = S(\mathcal{Y}_1^N)$ be a scalar or vector measurable function of $\mathcal{Y}_1^N$ and let us consider the distribution of $\mathcal{Y}_1^N$ conditioned by S, which we note $\mathbf{P}_\theta(\mathcal{Y}_1^N \in B|S)$, where $B \in \mathcal{B}^N$.

Definition 4.1.1 (Sufficient statistic). *We say that S is a* sufficient statistic *for the parametric family $\mathcal{P} = \{\mathbf{P}_\theta\}$ (or a sufficient statistic for the parameter θ characterizing the family $\mathcal{P}$) , if there exists a determination of the conditional distribution $\mathbf{P}_\theta(\mathcal{Y}_1^N \in B|S)$ that is independent of θ.*

In other words, the information about θ contained in the sample $\mathcal{Y}_1^N$ is concentrated in the statistic S, hence the name *sufficient*. For this reason, if the sufficient statistic S is available, it is not necessary to know the

whole sample $\mathcal{Y}_1^N$ to make inference about θ. The existence of a sufficient statistic can be investigated with the aid of the following criterion [Lehmann, 1986] :

Theorem 4.1.1 (Neyman-Fisher factorization). *A sufficient statistic S for the parametric family $\mathcal{P} = \{\mathbf{P}_\theta \ : \theta \in \Theta\}$ exists if and only if the pdf can be factorized as*

$$f_\theta(y) = \psi[S(y), \theta] \; h(y) \tag{4.1.1}$$

where ψ and h are nonnegative functions depending only upon their arguments, ψ is measurable in S, and h is measurable in y.

We now investigate four examples that will be useful in the other chapters. The first is a central topic in this and the two subsequent sections.

Example 4.1.1 (Likelihood ratio). *Let us consider two fixed values θ_0 and θ_1, and let $\Theta = \{\theta_0\} \cup \{\theta_1\}$. The likelihood ratio*

$$\Lambda(y) = \frac{f_{\theta_1}(y)}{f_{\theta_0}(y)} \tag{4.1.2}$$

is a sufficient statistic, as is obvious from

$$\begin{aligned} f_\theta(y) &= \psi[S(y), \theta] \; h(y) && (4.1.3)\\ \psi[S(y), \theta] &= \begin{cases} \Lambda(y) & \textit{when} \quad \theta = \theta_1 \\ 1 & \textit{when} \quad \theta = \theta_0 \end{cases} && (4.1.4)\\ h(y) &= f_{\theta_0}(y) && (4.1.5) \end{aligned}$$

and the above-mentioned factorization theorem. This basic fact is of key importance throughout the book for the design of change detection algorithms.

Example 4.1.2 (Exponential family - contd.). *In the case of a Koopman-Darmois exponential family (3.1.11), the statistic $T(y)$ is a sufficient statistic.*

We now apply this general situation to two particular cases.

Example 4.1.3 (Mean of a Gaussian sequence). *Let $\mathcal{L}(y) = \mathcal{N}(\theta, \sigma^2)$ and assume that $\mathcal{Y}_1^N$ is an independent sample of size N, and that σ^2 is known. A direct computation shows that the pdf can be written as*

$$\begin{aligned} f_\theta(\mathcal{Y}_1^N) &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \; e^{-\frac{1}{2\sigma^2}(y_i-\theta)^2} \\ &= (2\pi\sigma^2)^{-\frac{N}{2}} \; e^{\frac{1}{2\sigma^2}(2S\theta - N\theta^2)} \; e^{-\frac{1}{2\sigma^2}\sum_{i=1}^N y_i^2} \end{aligned} \tag{4.1.6}$$

where

$$S(y) = \sum_{i=1}^N y_i = N\bar{y} \tag{4.1.7}$$

and $\bar{y} = \frac{1}{N}\sum_{i=1}^N y_i$. In this example, the above-mentioned factorization is obtained with

$$\begin{aligned} \psi(S, \theta) &= e^{\frac{1}{2\sigma^2}(2S\theta - N\theta^2)} && (4.1.8)\\ h(y) &= (2\pi\sigma^2)^{-\frac{N}{2}} \; e^{-\frac{1}{2\sigma^2}\sum_{i=1}^N y_i^2} \end{aligned}$$

and S given before is therefore the sufficient statistic for the parameter θ. Writing the Neyman-Fisher factorization of the density and using the result of the previous example obviously gives the same result, as we emphasize in the next subsection.

Example 4.1.4 (Scalar AR model). *We consider here a scalar Gaussian AR process*

$$y_k = \sum_{i=1}^{p} a_i y_{k-i} + \varepsilon_k \tag{4.1.9}$$

where $(\varepsilon_k)_k$ is an independent Gaussian sequence with zero mean and variance σ^2. Now, following [Anderson, 1971], we recall that an AR model belongs to an exponential family, with respect to a particular parameterization, and thus we exhibit a sufficient statistic. Note that an ARMA model is not member of any exponential family. This is related to the fact that an ARMA process is not *a Markov process, and moreover this is obvious from the proof we derive now for the AR case. In what follows, we will see another sufficient statistic for AR models coming from the efficient score.*

Assuming the stability and thus the stationarity of the AR model, when N goes to infinity, the joint likelihood function of $\mathcal{Y}_1^N$ can be approximated by an exponential family (3.1.11) :

$$f_\theta(\mathcal{Y}_1^N) \approx e^{-\frac{1}{2}\sum_{i=0}^{p} \theta_i T_i(\mathcal{Y}_1^N) - d(\theta)} \tag{4.1.10}$$

with the natural parameter $\theta^T = (\ \theta_0 \quad \dots \quad \theta_p\ )$ defined by

$$\begin{aligned}
\theta_0 &= +\tfrac{1}{\sigma^2}(1 + a_1^2 + \cdots + a_p^2) \\
\theta_1 &= +\tfrac{2}{\sigma^2}(-a_1 + a_1 a_2 + \cdots + a_{p-1} a_p) \\
&\vdots \qquad \vdots \\
\theta_{p-1} &= \tfrac{2}{\sigma^2}(-a_{p-1} + a_1 a_p) \\
\theta_p &= -\tfrac{2}{\sigma^2} a_p
\end{aligned} \tag{4.1.11}$$

and with different possible choices of the quadratic forms :

$$T_i(\mathcal{Y}_1^N) = (\mathcal{Y}_1^N)^T E_i\ \mathcal{Y}_1^N \tag{4.1.12}$$

where the E_i are symmetric matrices and $E_0 = I$, and with

$$d(\theta) = \frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln\det\left(\sum_{i=0}^{p} \theta_i E_i\right) \tag{4.1.13}$$

The sufficient statistic T is thus the vector of T_i. The different quadratic forms $T_i(\mathcal{Y}_1^N)$ are all approximately equal to $\sum_{k=i+1}^{N} y_k y_{k-i}$. One such choice of weighting matrix in these quadratic forms is given by

$$E_i = \frac{1}{2}(C^i + C^{-i}) \tag{4.1.14}$$

where C is the circulant matrix :

$$C = \begin{pmatrix}
0 & 1 & 0 & 0 & \dots & 0 & 0 \\
0 & 0 & 1 & 0 & \dots & 0 & 0 \\
0 & 0 & 0 & 1 & \dots & 0 & 0 \\
0 & 0 & 0 & 0 & \dots & 0 & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & 0 & 0 & \dots & 0 & 1 \\
1 & 0 & 0 & 0 & \dots & 0 & 0
\end{pmatrix} \tag{4.1.15}$$

Other choices of E_i are given in [Anderson, 1971].

4.1.2 Information

In this subsection, we follow [Borovkov, 1984, Cox and Hinkley, 1986, Blahut, 1987]. We introduce several definitions of information, and investigate the relationships between information, exponential family, sufficiency, and efficiency, which are central issues in statistical inference and hypotheses testing. The three additional main reasons for this discussion here are :

- the role played by information concepts in asymptotic expansions of the likelihood ratio which we use for designing change detection algorithms;
- the important role played by information concepts in the theoretical results concerning the performances of the algorithms;
- the use of information concepts which we make in the subsequent chapters when we discuss detectability issues for different types of changes.

4.1.2.1 Scalar Parameter

We use the following notation. The log-likelihood function is

$$l_\theta(y) = \ln f_\theta(y) \tag{4.1.16}$$

and the log-likelihood ratio is

$$s(y) = \ln \frac{f_{\theta_1}(y)}{f_{\theta_0}(y)} = l_{\theta_1}(y) - l_{\theta_0}(y) \tag{4.1.17}$$

For small values of $\theta_1 - \theta_0$, we will give simple approximations to the first and second moments of s after the definition of the Fisher information.

Definition 4.1.2 (Efficient score). *When θ is a scalar parameter, we define the* efficient score *for the* random variable Y *as the quantity*

$$z = \frac{\partial l_\theta(y)}{\partial \theta} \tag{4.1.18}$$

Similarly, the efficient score for a sample of size N of a random process $(Y_n)_n$ *is denoted and defined by*

$$\mathcal{Z}_N = \frac{\partial l_\theta(\mathcal{Y}_1^N)}{\partial \theta} \tag{4.1.19}$$

If we note

$$z_i = \frac{\partial l_\theta(y_i|\mathcal{Y}_1^{i-1})}{\partial \theta} \tag{4.1.20}$$

we get

$$\mathcal{Z}_N = \sum_{i=1}^{N} z_i \tag{4.1.21}$$

This concept was introduced in [Fisher, 1925].

When the dependence on θ is of interest, we very often use the following notation

$$s^* = \ln \frac{f_\theta(y)}{f_{\theta^*}(y)} \tag{4.1.22}$$

$$z^* = \left.\frac{\partial l_\theta(y)}{\partial \theta}\right|_{\theta=\theta^*} \tag{4.1.23}$$

Now it is obvious that the efficient score is zero mean :

$$\mathbf{E}_{\theta^*}(z^*) = 0 \tag{4.1.24}$$

A simple approximation to $\mathbf{E}_\theta(z^*)$ for small values of $\theta - \theta^*$ will be given after the definition of the Fisher information. Note that in the particular case of the mean in a Gaussian random variable $\mathcal{L}(y) = \mathcal{N}(\mu, \sigma^2)$, the parameter of interest is $\theta = \mu$ and thus the efficient score is nothing but

$$z = \frac{y - \mu}{\sigma^2} \tag{4.1.25}$$

In this subsection, we investigate several issues connected to the *information* in the case of a scalar parameter. We extend these definitions and results to the vector parameter case in the next subsection.

Information We first recall the definition of the entropy and then consider two definitions of the information.

Definition 4.1.3 (Shannon entropy). *The* Shannon entropy *of the probability distribution of a* random variable *is defined as*

$$\mathbf{N}(p_\theta) = \mathbf{N}(\theta) = -\mathbf{E}_\theta[\ln p_\theta(Y)] \tag{4.1.26}$$

Similarly, the entropy contained in a sample of size N of a random process *is*

$$\mathbf{N}_N(\theta) = -\frac{1}{N} \int p_\theta(\mathcal{Y}_1^N) \ln p_\theta(\mathcal{Y}_1^N) d\mathcal{Y}_1^N \tag{4.1.27}$$

Definition 4.1.4 (Fisher information). *The* Fisher information *about θ contained in the* random variable Y *is*

$$\begin{aligned} \mathbf{I}(\theta) &= \mathbf{E}_\theta \left[\frac{\partial l_\theta(Y)}{\partial \theta} \right]^2 > 0 && (4.1.28) \\ &= \mathbf{E}_\theta \left[-\frac{\partial^2 l_\theta(Y)}{\partial \theta^2} \right] && (4.1.29) \\ &= \operatorname{var} \left[\frac{\partial l_\theta(Y)}{\partial \theta} \right] = \operatorname{var}(z) && (4.1.30) \end{aligned}$$

Note that

$$\mathbf{E}_\theta \left[-\frac{\partial^2 l_\theta(y)}{\partial \theta^2} \right] = \mathbf{E}_\theta \left[\frac{1}{f_\theta(y)} \frac{\partial^2 f_\theta(y)}{\partial \theta^2} \right] \tag{4.1.31}$$

and thus, from (4.1.29), the Fisher information is the expectation of the inverse of the curvature radius of the likelihood function.

Similarly, the Fisher information about the parameter θ contained in a sample of size N of a random process $(Y_n)_n$ *is*

$$\mathbf{I}_N(\theta) = \frac{1}{N} \operatorname{var} \left[\frac{\partial l_\theta(\mathcal{Y}_1^N)}{\partial \theta} \right] = \frac{1}{N} \operatorname{var} (\mathcal{Z}_N) \tag{4.1.32}$$

and in this case the Fisher information is defined to be

$$\mathbf{I}(\theta) = \lim_{N \to \infty} \mathbf{I}_N(\theta) \tag{4.1.33}$$

Note that the existence of this limit, which is always true, is a nontrivial fact [Pinsker, 1964]. Another possible definition of $\mathbf{I}_N(\theta)$ could have been $\operatorname{var}(\mathcal{Z}_N)$ *as in (4.1.30). We prefer to choose the* mean *information (4.1.32).*

In several chapters of this book, we make use of the following properties of the log-likelihood ratio s and efficient score z.

Lemma 4.1.1 (Approximations for the log-likelihood ratio). *Let f_θ be any probability density satisfying some regularity conditions. Note that f_θ does not need to belong to an exponential family of distributions, and especially does not need to be Gaussian. For small values of $(\theta_1 - \theta_0)$, we have*

$$\begin{aligned}
\mathbf{E}_{\theta_0}(s) &\approx -\frac{1}{2}\,\mathbf{I}(\theta_0)\,(\theta_1-\theta_0)^2 && (4.1.34)\\
\mathbf{E}_{\theta_1}(s) &\approx \frac{1}{2}\,\mathbf{I}(\theta_1)\,(\theta_1-\theta_0)^2 && (4.1.35)\\
&\approx \frac{1}{2}\,\mathbf{I}(\theta_0)\,(\theta_1-\theta_0)^2 && (4.1.36)\\
&\approx -\mathbf{E}_{\theta_0}(s)\\
\mathbf{E}_{\theta_0}(s^2) &\approx \mathbf{I}(\theta_0)\,(\theta_1-\theta_0)^2 && (4.1.37)\\
&\approx \mathbf{E}_{\theta_1}(s^2)
\end{aligned}$$

The proof of this lemma relies upon the following second-order Taylor expansion of l_θ :

$$s = l_{\theta_1} - l_{\theta_0} \approx (\theta_1-\theta_0)\;\left.\frac{\partial l_\theta}{\partial\theta}\right|_{\theta=\theta_0} + \frac{1}{2}(\theta_1-\theta_0)^2\;\left.\frac{\partial^2 l_\theta}{\partial\theta^2}\right|_{\theta=\theta_0} \tag{4.1.38}$$

Taking the expectation $\mathbf{E}_{\theta_0}$ of both sides of (4.1.38) leads to (4.1.34) [Borovkov, 1984], because

$$\mathbf{E}_{\theta_0}\left(\left.\frac{\partial l_\theta}{\partial\theta}\right|_{\theta=\theta_0}\right) = 0 \tag{4.1.39}$$

The approximation (4.1.35) is deduced by symmetry, and (4.1.36) comes from the approximation $\mathbf{I}(\theta_1) \approx \mathbf{I}(\theta_0)$. Moreover, raising (4.1.38) to the power 2 and keeping only second-order terms results in (4.1.37).

Lemma 4.1.2 (Approximation for the efficient score). *Let f_θ be as before. For small values of $(\theta - \theta^*)$, we have*

$$\mathbf{E}_\theta(z^*) \approx \mathbf{I}(\theta^*)\,(\theta-\theta^*) \tag{4.1.40}$$

The proof of this lemma relies upon the first term of the Taylor expansion (4.1.38), which we rewrite as

$$s^* = l_\theta - l_{\theta^*} \approx (\theta-\theta^*)\,z^* \tag{4.1.41}$$

Using (4.1.36) results in (4.1.40).

Definition 4.1.5 (Kullback information). *The* Kullback-Leibler information *between two probability densities f_{θ_0} and f_{θ_1} of a* random variable Y *is defined by*

$$\begin{aligned}
\mathbf{K}(\theta_0,\theta_1) &= \int \ln\frac{f_{\theta_0}(y)}{f_{\theta_1}(y)} f_{\theta_0}(y)\,dy\\
&= \mathbf{E}_{\theta_0}[-s(Y)]\\
&\geq 0
\end{aligned} \tag{4.1.42}$$

The Kullback information is zero only when the two densities are equal.

Similarly, in the case of a random process*, we define the Kullback information contained in a sample of size N by*

$$\mathbf{K}_N(\theta_0, \theta_1) = \frac{1}{N} \int \ln \frac{f_{\theta_0}(\mathcal{Y}_1^N)}{f_{\theta_1}(\mathcal{Y}_1^N)} f_{\theta_0}(\mathcal{Y}_1^N) d\mathcal{Y}_1^N \tag{4.1.43}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \int \ln \frac{f_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})}{f_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})} f_{\theta_0}(\mathcal{Y}_1^N) d\mathcal{Y}_1^N \tag{4.1.44}$$

and in this case the Kullback information is defined to be

$$\mathbf{K}(\theta_0, \theta_1) = \lim_{N \to \infty} \mathbf{K}_N(\theta_0, \theta_1) \tag{4.1.45}$$

Note again that the existence of this limit, which is always true, is a nontrivial fact [Pinsker, 1964].

The following asymptotic approximation is of interest. Let

$$s_i = \ln \frac{f_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{f_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})} \tag{4.1.46}$$

When $N \to \infty$, by the law of large numbers under $\mathbf{P}_{\theta_0}$ we have

$$\frac{1}{N} \sum_{i=1}^{N} \int \ln \frac{f_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})}{f_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})} f_{\theta_0}(\mathcal{Y}_1^N) d\mathcal{Y}_1^N \approx \frac{1}{N} \sum_{i=1}^{N} s_i \tag{4.1.47}$$

Therefore, the Kullback information (4.1.43) can be approximated by

$$\mathbf{K}_N(\theta_0, \theta_1) \approx \frac{1}{N} \sum_{i=1}^{N} s_i \tag{4.1.48}$$

This second information is not a distance, basically because it is not symmetric. A symmetrized version

$$\mathbf{J}(\theta_0, \theta_1) = \mathbf{K}(\theta_0, \theta_1) + \mathbf{K}(\theta_1, \theta_0) \tag{4.1.49}$$

is called the *Kullback divergence* and will be used in several places in this book, for example, for measuring a magnitude of change in chapters 7 and 8. Kullback information and divergence will be used for defining the detectability of a given change in section 6.3. It is also of interest that the maximum likelihood estimate $\hat{\theta}$ of θ minimizes the Kullback information $K(\theta, \hat{\theta})$ [Kullback, 1959].

The Fisher and Kullback information do have strong connections in several particular cases of interest in this book, as we explain in sections 4.2, 7.2, and 8.2. One basic general connection is the following.

Lemma 4.1.3 (Approximation of the Kullback information). *From the approximation (4.1.34) and the definition (4.1.42) we find that for small values of* $(\theta_1 - \theta_0)$

$$\mathbf{K}(\theta_0, \theta_1) \approx \frac{1}{2}(\theta_1 - \theta_0)^2 \, \mathbf{I}(\theta_0) \tag{4.1.50}$$

Note again that this approximation is fairly general, and does not require that the distribution belong to an exponential family.

Example 4.1.5 (Exponential family - contd.). *In the case of a Koopman-Darmois exponential family of distributions,*

$$f_\theta(y) = h(y)e^{c(\theta)T(y)-d(\theta)} \tag{4.1.51}$$

which we introduced in section 3.1, the efficient score and Fisher and Kullback information are as follows. The efficient score is

$$z = \dot{c}(\theta)T(y) - \dot{d}(\theta) \tag{4.1.52}$$

and from this we deduce that

$$\mathbf{E}_\theta[T(y)] = \frac{\dot{d}(\theta)}{\dot{c}(\theta)} \tag{4.1.53}$$

because z is zero mean. Moreover, the Fisher information is

$$\mathbf{I}(\theta) = \dot{c}(\theta)\left[\frac{\partial}{\partial\theta}\frac{\dot{d}(\theta)}{\dot{c}(\theta)}\right] \tag{4.1.54}$$

On the other hand, because of (4.1.17), (4.1.42), and (4.1.51), the Kullback information is given by

$$\begin{aligned}\mathbf{K}(\theta_0,\theta_1) &= [c(\theta_1)-c(\theta_0)]\ \ \mathbf{E}_{\theta_0}[T(y)] - [d(\theta_1)-d(\theta_0)] \\ &= \dot{d}(\theta_0)\left[\frac{d(\theta_1)-d(\theta_0)}{\dot{d}(\theta_0)} - \frac{c(\theta_1)-c(\theta_0)}{\dot{c}(\theta_0)}\right]\end{aligned} \tag{4.1.55}$$

In the case of a natural parameter, $c(\theta) = \theta$ and thus

$$\begin{aligned} z &= T(y) - \dot{d}(\theta) \\ \mathbf{I}(\theta) &= \ddot{d}(\theta) && \textit{because of (4.1.28)} \\ &= \mathrm{var}[T(y)] && \textit{because of (4.1.30)} \\ \mathbf{K}(\theta_0,\theta_1) &= d(\theta_1) - d(\theta_0) - (\theta_1-\theta_0)\dot{d}(\theta_0) \end{aligned} \tag{4.1.56}$$

Information and sufficiency Considering a conditional distribution, we said before that if S is a sufficient statistic, then the information about the parameter θ contained in the observation Y is concentrated in S. This statement can be reinforced in a more formal way using the two definitions of information, as we show next.

Let S be a statistic with density g_θ induced by f_θ; we will define *Fisher information contained in S* by the following quantity :

$$\mathbf{I}^S(\theta) = \mathbf{E}_\theta\left[\frac{\partial}{\partial\theta}\ln g_\theta(S)\right]^2 \tag{4.1.57}$$

The following inequality holds :

$$\mathbf{I}^S(\theta) \leq \mathbf{I}(\theta) \tag{4.1.58}$$

Remembering the Neyman-Fisher factorization formula as a necessary and sufficient condition of existence of a sufficient statistic S, and noting that

$$g_\theta[S(y)] = \psi[S(y),\theta] \tag{4.1.59}$$

is then the density of this sufficient statistic, it is possible to show that the previous inequality is an *equality* if and only if S is a sufficient statistic [Borovkov, 1984]. In other words, a sufficient statistic keeps the whole Fisher information.

It is also possible to show that a sufficient statistic keeps the whole Kullback information [Blahut, 1987], namely that

$$\mathbf{K}(g_{\theta_0}, g_{\theta_1}) = \mathbf{K}(f_{\theta_0}, f_{\theta_1}) \tag{4.1.60}$$

Information and efficiency Let $\hat{\theta}$ be an estimate of θ having a bias $b(\theta)$. Then, under some regularity assumptions, the precision of this estimate is bounded from below, according to the following inequality, which is known as the *Cramer-Rao inequality* :

$$\mathrm{var}(\hat{\theta}) \geq \frac{[1+\dot{b}(\theta)]^2}{\mathbf{I}(\theta)} \tag{4.1.61}$$

In particular, when the estimate is unbiased, we have

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{\mathbf{I}(\theta)} \tag{4.1.62}$$

which has to be compared with (4.1.56). When the equality is attained in (4.1.61), the estimate $\hat{\theta}$ is said to be *efficient*.

We now give three useful examples. The first two belong to the case of exponential family. The last, concerned with χ^2 distributions, does not belong to that case; it is useful for the χ^2-CUSUM algorithms presented in chapter 2.

Example 4.1.6 (Mean of a Gaussian sequence - contd.). *Consider the Gaussian distribution $\mathcal{N}(\theta, \sigma^2)$, where σ^2 is assumed to be known. The corresponding exponential family (4.1.51) with natural parameter θ is then given by*

$$\begin{aligned} d(\theta) &= \frac{\theta^2}{2\sigma^2} \\ T(y) &= \frac{y}{\sigma^2} \\ h(y) &= \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{y^2}{2\sigma^2}} \end{aligned} \tag{4.1.63}$$

The Fisher information is given by the second derivative of d

$$\mathbf{I}(\theta) = \frac{1}{\sigma^2} \tag{4.1.64}$$

and does not depend upon the unknown parameter θ, but is only inversely proportional to the level of noise, and the Kullback information, computed with the aid of (4.1.56), is

$$\mathbf{K}(\theta_0, \theta_1) = \frac{(\theta_1 - \theta_0)^2}{2\sigma^2} \tag{4.1.65}$$

The Kullback divergence in this case is nothing but the *signal-to-noise ratio*, usually measured as $10\ \ln \frac{(\theta_1-\theta_0)^2}{\sigma^2}$ on the decibel scale. As already mentioned, the Kullback divergence is used as a measure of the magnitude of the change throughout the book.

Example 4.1.7 (Variance of a Gaussian sequence). *The exponential family corresponding to the natural parameter $\theta = \frac{1}{\sigma^2}$ of the law $\mathcal{N}(\mu, \sigma^2)$ is specified by*

$$\begin{aligned} d(\theta) &= \frac{\mu^2}{2}\,\theta - \frac{1}{2}\,\ln\theta \\ T(y) &= -y^2 + 2\mu y \\ h(y) &= \frac{1}{\sqrt{2\pi}} \end{aligned} \tag{4.1.66}$$

The second derivative of d gives the Fisher information :

$$\mathbf{I}(\theta) = \frac{1}{2\theta^2} = \frac{\sigma^4}{2} \tag{4.1.67}$$

and the Kullback information is again given by (4.1.56), which results in

$$\mathbf{K}(\theta_0, \theta_1) = \frac{1}{2}\left(\ln\frac{\theta_0}{\theta_1} + \frac{\theta_1 - \theta_0}{\theta_0}\right) = \frac{1}{2}\left[\ln\frac{\sigma_1^2}{\sigma_0^2} + \left(\frac{\sigma_0^2}{\sigma_1^2} - 1\right)\right] \tag{4.1.68}$$

Let us now discuss the two types of information for a distribution that does not belong to an exponential family.

Example 4.1.8 (χ^2 distribution - contd.). *Using the definition, the Fisher information about the parameter θ in a gamma distribution $\gamma(\theta, b)$ defined in (3.1.19) can be shown to be*

$$\mathbf{I}(\theta) = \frac{\ddot{\Gamma}(\theta)\Gamma(\theta) - \dot{\Gamma}(\theta)}{\Gamma^2(\theta)} \tag{4.1.69}$$

The Kullback information between a $\chi^2(n)$ and a $\chi'^2(n, \lambda)$ distribution is discussed in section 7.3.

4.1.2.2 Vector Case

The previous results do extend to the case of a vector observation of dimension r in a trivial manner, and to the case of a vector parameter θ of dimension ℓ. In the latter case, the efficient score is defined as

$$Z = \frac{\partial l_\theta(y)}{\partial \theta} \tag{4.1.70}$$

and the Fisher information is defined as an $\ell \times \ell$ matrix with elements

$$\mathbf{I}_{ij}(\theta) = \int_{-\infty}^{+\infty} \left[\frac{\partial}{\partial \theta_i} \ln f_\theta(y)\right] \left[\frac{\partial}{\partial \theta_j} \ln f_\theta(y)\right] f_\theta(y) dy \tag{4.1.71}$$

which is an obvious extension of (4.1.28). If the observation Y is a vector, the elements of the Fisher information matrix are expressed as multiple integrals.

Similarly, the efficient score for a sample of size N of a *random process* $(Y_n)_n$ is denoted and defined by

$$\mathcal{Z}_N = \frac{\partial l_\theta(\mathcal{Y}_1^N)}{\partial \theta} \tag{4.1.72}$$

If we note

$$Z_i = \frac{\partial l_\theta(y_i | \mathcal{Y}_1^{i-1})}{\partial \theta} \tag{4.1.73}$$

we get

$$\mathcal{Z}_N = \sum_{i=1}^{N} Z_i \tag{4.1.74}$$

The efficient score can also be viewed as an ℓ-dimensional vector :

$$\mathcal{Z}_N = \begin{pmatrix} \mathcal{Z}_1 \\ \vdots \\ \mathcal{Z}_\ell \end{pmatrix} \tag{4.1.75}$$

The Fisher information matrix is then an $\ell \times \ell$ matrix with elements

$$\mathbf{I}_{ij}(\theta) = \frac{1}{N}\mathbf{E}_\theta(\mathcal{Z}_i \mathcal{Z}_j^T) \tag{4.1.76}$$

Lemma 4.1.4 (Approximations). *Under some regularity conditions on f_θ and for small values of $\|\theta_1-\theta_0\|$, we have the following approximations :*

$$\mathbf{E}_{\theta_1}(s) \approx \frac{1}{2}(\theta_1-\theta_0)^T\mathbf{I}(\theta_0)(\theta_1-\theta_0) \quad (4.1.77)$$

$$\mathbf{E}_{\theta_0}(s^2) \approx (\theta_1-\theta_0)^T\mathbf{I}(\theta_0)(\theta_1-\theta_0) \quad (4.1.78)$$

for the log-likelihood ratio, and

$$\mathbf{E}_\theta(\mathcal{Z}^*) \approx \mathbf{I}(\theta^*)(\theta-\theta^*) \quad (4.1.79)$$

for the efficient score. Furthermore, the Kullback information (4.1.42) can be approximated [Blahut, 1987] as

$$\mathbf{K}(\theta_0,\theta_1) \approx \frac{1}{2}(\theta_0-\theta_1)^T\ \mathbf{I}(\theta_0)\ (\theta_0-\theta_1) \quad (4.1.80)$$

which is the extension of (4.1.50).

Exponential families are defined as in (4.1.51), where c is a *row vector* and T is a *column vector* and their product is thus understood as a scalar product. To decrease the ambiguity of this representation, the components of c are chosen to be linearly independent [Borovkov, 1984]. The density of a sufficient statistic S is $g_\theta[S(y)] = \psi[S(y),\theta]$ as before. The only general formula for the Fisher information matrix in this exponential case is the definition (4.1.71) given before; the reason is that, unlike in the scalar case, there is no analytic formula for $\mathbf{E}_\theta[T(y)]$. The general formula for the Kullback information is (4.1.55) as in the scalar case, with the same meaning of the product cT as in the vector counterpart of (4.1.51).

Information and sufficiency The following property is also of interest. When the Fisher information matrix is block-diagonal, we can easily deduce sufficient statistic for *subsets* of parameters. Partitioning θ and z as

$$\theta = \begin{pmatrix} \theta_a \\ \theta_b \end{pmatrix}\ ,\quad z = \begin{pmatrix} z_a \\ z_b \end{pmatrix} \quad (4.1.81)$$

and using the definition $\mathbf{I}(\theta) = \mathbf{E}_\theta(zz^T)$, we get

$$\begin{pmatrix} \mathbf{I}(\theta_a) & 0 \\ 0 & \mathbf{I}(\theta_b) \end{pmatrix} = \begin{pmatrix} \mathbf{E}_\theta(z_a z_a^T) & 0 \\ 0 & \mathbf{E}_\theta(z_b z_b^T) \end{pmatrix} \quad (4.1.82)$$

which shows that z_a (respectively z_b) is a sufficient statistic for θ_a (respectively θ_b). In the AR case, we use this result to show that the innovation is a sufficient statistic only for changes in the standard deviation of the input excitation.

Information and efficiency The Cramer-Rao bound (4.1.61) for an estimate $\hat{\theta}$ with bias $b(\theta)$ is now characterized by the following inequality between matrices :

$$\mathrm{var}(\hat{\theta}) \geq \left[I_\ell + \dot{b}(\theta)\right]^T\ \mathbf{I}^{-1}(\theta)\ \left[I_\ell + \dot{b}(\theta)\right] \quad (4.1.83)$$

We now give three examples of computation of both types of information in the case of a vector parameter. The first two belong to exponential families of distribution, but the third does not.

Example 4.1.9 (Mean and variance of a scalar Gaussian sequence). *We consider the parameterization $\theta = (\ \theta_1\quad \theta_2\ )$ with $\theta_1 = \mu$ and $\theta_2 = \sigma^2$ in the exponential family*

$$f_\theta(y) = h(y)\ e^{\sum_{i=1}^2 c_i(\theta)T_i(y) - d(\theta)} \quad (4.1.84)$$

Noting that in this case

$$\begin{aligned} d(\theta) &= \frac{1}{2}\ln\theta_2 + \frac{\theta_1^2}{2\theta_2} \\ \sum_{i=1}^{2} c_i(\theta)T_i(y) &= -\frac{y^2}{2\theta_2} + y\frac{\theta_1}{\theta_2} \end{aligned} \tag{4.1.85}$$

and using the definition (4.1.71), it is easy to show that the Fisher information matrix is given by

$$\mathbf{I}(\mu,\sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \tag{4.1.86}$$

and thus is diagonal and independent of the mean value. The Kullback information is given by (4.1.55)

$$\mathbf{K}(\theta_0,\theta_1) = \frac{(\mu_1-\mu_0)^2}{2\sigma_1^2} + \frac{1}{2}\left(\ln\frac{\sigma_1^2}{\sigma_0^2} + \frac{\sigma_0^2}{\sigma_1^2} - 1\right) \tag{4.1.87}$$

Note that the second term on the right of this formula is strictly positive as long as $\sigma_1 \neq \sigma_0$. This means that the Kullback information in the case of joint changes in the mean and *in the variance of a Gaussian variable is greater than the Kullback information corresponding to a change in the single mean. As we show in the next chapter, for a fixed time between false alarms, the delay for detection is inversely proportional to the Kullback information, and thus it is easier to detect a change in the mean arising together with a change in the variance than to detect a change in the single mean.*

Example 4.1.10 (Mean of a vector Gaussian sequence). *In this case, Y is of dimension r and law $\mathcal{L}(Y) = \mathcal{N}(\mu,\Sigma)$. As in example 3.1.5, we distinguish two cases according to the rank of Σ. When Σ is positive definite, this law has a density (3.1.37), the natural parameter is $\theta = \mu$, and $d(\theta) = \frac{1}{2}\theta^T\Sigma^{-1}\theta$. Thus, the Fisher information matrix is then simply*

$$\mathbf{I}(\theta) = \Sigma^{-1} \tag{4.1.88}$$

The Kullback information is

$$\mathbf{K}(\theta_0,\theta_1) = \frac{1}{2}(\theta_0-\theta_1)^T\Sigma^{-1}(\theta_0-\theta_1) \tag{4.1.89}$$

and the Kullback divergence is

$$\mathbf{J}(\theta_0,\theta_1) = (\theta_0-\theta_1)^T\Sigma^{-1}(\theta_0-\theta_1) \tag{4.1.90}$$

When Σ is degenerated with rank $\tilde{r} < r$, it results from (3.1.41) that the Kullback divergence is given by

$$\mathbf{J}(\theta_0,\theta_1) = (\theta_0-\theta_1)^T AD^{-1}A^T(\theta_0-\theta_1) \tag{4.1.91}$$

where D is the diagonal matrix of the nonzero eigenvalues of Σ, and the columns of A are the corresponding eigenvectors.

We now discuss the case of scalar AR and ARMA models. For computing the Fisher information matrices with respect to the AR and MA parameters, we use the definition in terms of the efficient score, and not the general result which is available for exponential families. The reason is that ARMA processes do not belong to this family, and the same is true of AR models for the particular parameterization that we consider here (see the third example of subsection 4.1.1).

Example 4.1.11 (Fisher information in scalar AR and ARMA processes). *Consider here the Gaussian stable ARMA model*

$$y_k = \sum_{i=1}^{p} a_i y_{k-i} + \sum_{j=1}^{q} b_j v_{k-j} + v_k \tag{4.1.92}$$

where $(v_k)_k$ is a Gaussian white noise sequence with variance σ^2. Let

$$\theta^T = \begin{pmatrix} a_1 & \dots & a_p & b_1 & \dots & b_q & \sigma \end{pmatrix} \tag{4.1.93}$$

be the vector of parameters of interest, and let ε_k be the innovation, which can be computed recursively with the aid of

$$\varepsilon_k = y_k - \sum_{i=1}^{p} a_i y_{k-i} - \sum_{j=1}^{q} b_j \varepsilon_{k-j} \tag{4.1.94}$$

The conditional probability distribution of the observation y_k is given by

$$p_\theta(y_k|\mathcal{Y}_1^{k-1}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(A^T \check{\mathcal{Y}}_{k-p}^{k} - B^T \check{\mathcal{E}}_{k-q}^{k-1})^2} \tag{4.1.95}$$

$$= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\varepsilon_k^2} \tag{4.1.96}$$

where we use the following notation :

$$A^T = \begin{pmatrix} 1 & -a_1 & \dots & -a_p \end{pmatrix} \tag{4.1.97}$$

$$B^T = \begin{pmatrix} b_1 & \dots & b_q \end{pmatrix} \tag{4.1.98}$$

for the sets of AR and MA parameters, and

$$(\check{\mathcal{Y}}_{k-p}^{k})^T = \begin{pmatrix} y_k & y_{k-1} & \dots & y_{k-p} \end{pmatrix} \tag{4.1.99}$$

$$(\check{\mathcal{E}}_{k-q}^{k-1})^T = \begin{pmatrix} \varepsilon_{k-1} & \varepsilon_{k-2} & \dots & \varepsilon_{k-q} \end{pmatrix} \tag{4.1.100}$$

for the sets of past observations and innovations ordered backward.

Let us compute the Fisher information matrix (4.1.71) about θ contained in a large sample of observations of size N, with the aid of the efficient score defined before. First, we investigate the case of an AR model. *We recall that for a stable AR model, the influence, on the likelihood function and thus on the efficient score, of the initial values $y_0, \dots, y_{1-p}$ of the observations is negligible. Using the definition, it is easy to show that the efficient score is*

$$\mathcal{Z}_N = \begin{pmatrix} \frac{1}{\sigma^2} \sum_{i=1}^{N} \check{\mathcal{Y}}_{i-p}^{i-1} \varepsilon_i \\ \\ \frac{1}{\sigma} \sum_{i=1}^{N} \left(\frac{\varepsilon_i^2}{\sigma^2} - 1 \right) \end{pmatrix} \tag{4.1.101}$$

The Fisher information matrix is, by definition, the covariance matrix of the efficient score. Because the innovation is independent from the past observations, the first consequence of (4.1.101) is that the Fisher information matrix of an AR model with respect to the p AR parameters on one hand and the standard deviation σ of the innovation on the other hand is block diagonal. One straightforward consequence of this fact is that the innovation of an AR process is not a sufficient statistic for detecting changes in the AR parameters. Furthermore, using (4.1.33), straightforward computations result in

$$\mathbf{I}(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} \mathbf{E}_\theta(\check{\mathcal{Y}}_{k-p}^{k-1} (\check{\mathcal{Y}}_{k-p}^{k-1})^T) & 0 \\ \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}$$

$$= \frac{1}{\sigma^2} \begin{pmatrix} \mathbf{T}_p & 0 \\ 0 & 2 \end{pmatrix} \tag{4.1.102}$$

where the $p \times p$ matrix $\mathbf{T}_p$ defined by

$$\mathbf{T}_p = \begin{pmatrix} R_0 & R_1 & \dots & R_{p-1} \\ R_1 & R_0 & \dots & R_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ R_{p-1} & R_{p-2} & \dots & R_0 \end{pmatrix} \tag{4.1.103}$$

is nothing but the Toeplitz matrix *associated with the AR process*.

Now, let us consider the case of an ARMA model. *Here the derivation of the efficient score is less simple, because of the derivatives of the innovation, which are defined by*

$$\begin{aligned} \alpha_{k-i} &= -\frac{\partial \varepsilon_k}{\partial a_i} \\ \beta_{k-j} &= -\frac{\partial \varepsilon_k}{\partial b_j} \end{aligned} \tag{4.1.104}$$

and are the outputs of the same AR model :

$$\begin{aligned} \alpha_k &= -\sum_{j=1}^{q} b_j \alpha_{k-j} + y_k \\ \beta_k &= -\sum_{j=1}^{q} b_j \beta_{k-j} + \varepsilon_k \end{aligned} \tag{4.1.105}$$

Then similar computations starting from the previous conditional distribution result in

$$\mathcal{Z}_N = \begin{pmatrix} \frac{1}{\sigma^2} \sum_{i=1}^{N} \check{\mathcal{A}}_{i-p}^{i-1} \varepsilon_i \\ \\ \frac{1}{\sigma^2} \sum_{i=1}^{N} \check{\mathcal{B}}_{i-q}^{i-1} \varepsilon_i \\ \\ \frac{1}{\sigma} \sum_{i=1}^{N} \left(\frac{\varepsilon_i^2}{\sigma^2} - 1 \right) \end{pmatrix} \tag{4.1.106}$$

and

$$\mathbf{I}(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} \mathbf{E}_\theta [\check{\mathcal{A}}_{k-p}^{k-1} (\check{\mathcal{A}}_{k-p}^{k-1})^T] & \frac{1}{\sigma^2} \mathbf{E}_\theta [\check{\mathcal{A}}_{k-p}^{k-1} (\check{\mathcal{B}}_{k-q}^{k-1})^T] & 0 \\ \\ \frac{1}{\sigma^2} \mathbf{E}_\theta [\check{\mathcal{B}}_{k-q}^{k-1} (\check{\mathcal{A}}_{k-p}^{k-1})^T] & \frac{1}{\sigma^2} \mathbf{E}_\theta [\check{\mathcal{B}}_{k-q}^{k-1} (\check{\mathcal{B}}_{k-q}^{k-1})^T] & 0 \\ \\ 0 & 0 & \frac{2}{\sigma^2} \end{pmatrix} \tag{4.1.107}$$

where $\check{\mathcal{A}}_{k-p}^{k-1}$ and $\check{\mathcal{B}}_{k-q}^{k-1}$ are the sets of α and β ordered backward. Because of (4.1.105), $\check{\mathcal{A}}_{k-p}^{k-1}$ and $\check{\mathcal{B}}_{k-q}^{k-1}$ are not independent. Therefore, it results from formula (4.1.107) that the Fisher information matrix is not block diagonal with respect to AR coefficients on one hand and MA coefficients on the other hand. *This property means that, in the log-likelihood function of an ARMA process, there is a basic coupling between the AR and MA parts. This fact is used in chapters 8 and 9.*

Furthermore, a closed-form expression of the Fisher information matrix with respect to the parameterization of an ARMA process in terms of the magnitudes and angles of the poles and the zeroes *is proposed in [Bruzzone and Kaveh, 1984]. From this expression, it can be deduced that the Fisher information matrix of a scalar AR process tends to a diagonal matrix when the poles go to the unit circle. This means that the log-likelihood function ensures an approximate decoupling of weakly damped poles. This fact can be used for solving the isolation or diagnosis problem, once the change has been detected.*

We now give two general results concerning the Kullback information between two vector Gaussian processes, and several formulas for computing the Kullback information between two AR models. These results will be useful for discussing the detectability issue in chapters 8 and 9.

Example 4.1.12 (Kullback information for vector Gaussian processes). *Let $(Y_k)_k$ be a zero mean r-dimensional Gaussian process having two possible probability density functions f_0, f_1 and corresponding power spectra $\Phi_0(\omega), \Phi_1(\omega)$. Then the Kullback information is given by*

$$\mathbf{K}(f_0, f_1) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \operatorname{tr} \Phi_1^{-1}(\omega)\Phi_0(\omega) - \operatorname{tr} I_r - \ln \det[\Phi_1^{-1}(\omega)\Phi_0(\omega)] \right\} d\omega \tag{4.1.108}$$

[Kazakos and Papantoni-Kazakos, 1980] and the Kullback divergence is thus

$$\mathbf{J}(f_0, f_1) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \operatorname{tr} \left[\Phi_1^{-1}(\omega)\Phi_0(\omega) + \Phi_0^{-1}(\omega)\Phi_1(\omega) - 2I_r \right] d\omega \tag{4.1.109}$$

Another useful expression of the Kullback information and divergence contained in a sample of size N of a Gaussian process is the following [Pinsker, 1964, Basseville, 1989]. Let $\mathbf{T}_N(\Phi)$ be the $(N+1) \times (N+1)$ block-Toeplitz matrix filled with the covariance matrices associated with the power spectrum Φ through inverse Fourier transform. The (i,j)th block of this matrix is the covariance matrix of order $i-j$. Then the Kullback information is

$$2\, \mathbf{K}_N(f_0, f_1) = \operatorname{tr} \left[\mathbf{T}_N^{-1}(\Phi_1)\mathbf{T}_N(\Phi_0) \right] + \ln \det \left[\mathbf{T}_N^{-1}(\Phi_1)\mathbf{T}_N(\Phi_0) \right] - N \tag{4.1.110}$$

and the Kullback divergence is

$$2\, \mathbf{J}_N(f_0, f_1) = \operatorname{tr} \left[\mathbf{T}_N^{-1}(\Phi_1)\mathbf{T}_N(\Phi_0) + \mathbf{T}_N^{-1}(\Phi_0)\mathbf{T}_N(\Phi_1) \right] - 2N \tag{4.1.111}$$

We now consider the particular case of AR processes, which is discussed in chapters 8 and 9.

Example 4.1.13 (Kullback information for AR processes). *When the process is an AR process,*

$$Y_k = \sum_{i=1}^{p} A_i Y_{k-i} + V_k \tag{4.1.112}$$

where $(V_k)_k$ is a Gaussian white noise sequence with covariance matrix R, then from (3.2.36) the formula (4.1.109) for Kullback divergence becomes

$$\begin{aligned} 2\, \mathbf{J}(f_0, f_1) \;=\; & \frac{1}{2\pi} \int_{-\pi}^{\pi} \operatorname{tr} \Big[A_1^T(e^{-i\omega}) R_1^{-1} A_1(e^{i\omega}) A_0^{-1}(e^{i\omega}) R_0 A_0^{-T}(e^{-i\omega}) \\ & + A_0^T(e^{-i\omega}) R_0^{-1} A_0(e^{i\omega}) A_1^{-1}(e^{i\omega}) R_1 A_1^{-T}(e^{-i\omega}) - 2I_r \Big] d\omega \end{aligned} \tag{4.1.113}$$

where we note

$$A(z) = I_r - \sum_{i=1}^{p} A_i z^{-i} \tag{4.1.114}$$

and where $A_l(z)$ and R_l correspond to the model with index l ($l = 0, 1$). We also use the following notation :

$$A^T = \begin{pmatrix} I_r & -A_1 & \dots & -A_p \end{pmatrix} \tag{4.1.115}$$

In the case of a scalar *AR process with input variance $R = \sigma^2$, the following expression [Pinsker, 1964, Basseville, 1989] is also useful :*

$$2\, \mathbf{K}(f_0, f_1) = \frac{1}{\sigma_1^2} (A^1)^T \mathbf{T}_p(\Phi_0) A^1 - \ln \frac{\sigma_0^2}{\sigma_1^2} - 1 \tag{4.1.116}$$

where A^1 is the block-row vector of parameters defined in (4.1.115) and corresponding to the model with index 1.

4.2 Hypotheses Testing

In this section, we begin the systematic investigation of the main ideas of hypotheses testing, which will be continued in the next section. These two sections are of key importance for the subsequent chapters, because the hypotheses testing theory is the *main* background of change detection. We begin by investigating the case of *fixed* sample size, or equivalently off-line detection algorithms. Then in the next section, we investigate sequential analysis, which is related to random sample size and on-line detection algorithms, as will be discussed in detail later. Here we follow [Lehmann, 1986, Borovkov, 1984].

4.2.1 Notation and Main Criteria

Let us introduce the main definitions and criteria of the hypotheses testing framework.

Definition 4.2.1 (Simple hypothesis). *We call* simple hypothesis $\mathbf{H}$ *any assumption concerning the distribution* $\mathbf{P}$ *that can be reduced to a single value in the space of probability distributions.*

Assume that we have M distributions $\mathbf{P}_0, \ldots, \mathbf{P}_{M-1}$, and let $\mathcal{Y}_1^N$ be an N-size sample generated by one of these distributions. The problem of hypotheses testing is to decide which distribution is the true one. The parametric version of this testing problem is the following. Let $\mathbf{P}_\theta \in \mathcal{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta}$ and consider the simple hypotheses $\mathbf{H}_j : \mathcal{L}(\mathcal{Y}) = \mathbf{P}_{\theta_j} (j = 0, \ldots, M-1)$, where $\theta_0, \ldots, \theta_{M-1}$ are fixed points in the parameter space. In the subsequent chapters, we shall use mainly the parametric case.

Definition 4.2.2 *We call a* statistical test *for testing between hypotheses* $\mathbf{H}_0, \ldots, \mathbf{H}_{M-1}$ *any measurable mapping* $g : \Omega^N \to \{\mathbf{H}_0, \ldots, \mathbf{H}_{M-1}\}$.

In other words, $g(\mathcal{Y}_1^N)$ is a random variable that takes its values in the set of hypotheses. If $g(\mathcal{Y}_1^N) = \mathbf{H}_k$, then hypothesis $\mathbf{H}_k$ is accepted. In the parametric case, we simply say that $\theta = \theta_k$. We also call the function $g(\mathcal{Y}_1^N)$ a *decision function*. Giving the decision function g is equivalent to giving a partition of Ω^N into M nonintersecting Borel sets $\Omega_0, \ldots, \Omega_{M-1}$, inside which exactly one of the hypotheses is accepted. When $M = 2$, the set $\Omega_1 \subset \Omega^N$ is said to be the *critical region* of the test g.

The quality of a statistical test is usually defined with the aid of a set of error probabilities as follows :

$$\alpha_i = \mathbf{P}(\mathcal{Y}_1^N \notin \Omega_i | \mathbf{H}_i) = \mathbf{P}[g(\mathcal{Y}_1^N) \neq \mathbf{H}_i | \mathbf{H}_i] \tag{4.2.1}$$

where α_i is the probability of rejecting hypothesis $\mathbf{H}_i$ when it is true. Obviously, all α_i should be small. But, because the sample is of finite length, all α_i are strictly positive for nondeterministic decisions. The question then arises of how to compare two different statistical tests. Let us consider three well-known approaches for this purpose.

1. **Most powerful approach :** Let us define a class of tests with $M-1$ given errors :

$$K_{\alpha_0,\ldots,\alpha_{M-2}} = \{g : \alpha_j(g) = \alpha_j; j = 0, \ldots, M-2\} \tag{4.2.2}$$

Definition 4.2.3 (Most powerful test). *We say that the test* $g^* \in K_{\alpha_0,\ldots,\alpha_{M-2}}$ *is the* most powerful *(MP) in this class if, for all* $g \in K_{\alpha_0,\ldots,\alpha_{M-2}}$*, the following inequality holds for the* M*-th error :*

$$\alpha_{M-1}(g^*) \leq \alpha_{M-1}(g) \tag{4.2.3}$$

2. **Bayesian approach :** Assume that hypotheses $\mathbf{H}_i$ $(i = 0, \ldots, M-1)$ have known *a priori* probabilities q_i such that $\sum_{i=0}^{M-1} q_i = 1$. For a given statistical test g, we then define the *weighted error probability* $\bar{\alpha}(g)$ by

$$\bar{\alpha}(g) = \sum_{j=0}^{M-1} \mathbf{P}(\mathbf{H}_j)\mathbf{P}[g(\mathcal{Y}_1^N) \neq \mathbf{H}_j | \mathbf{H}_j] \tag{4.2.4}$$

$$= \sum_{j=0}^{M-1} q_j \alpha_j \tag{4.2.5}$$

Definition 4.2.4 (Bayes test). *The test $\bar{g}$ is said to be a* Bayes test *if it minimizes the error probability $\bar{\alpha}(g)$ for given a priori probabilities $q_j = \mathbf{P}(\mathbf{H}_j)$ $(j = 0, \ldots, M-1)$.*

3. **Minmax approach :** Let us define the *maximum error probability* of a test g as follows :

$$\alpha(g) = \max_{j=0,\ldots,M-1} \alpha_j(g) \tag{4.2.6}$$

Definition 4.2.5 (Minimax test). *We say that the test $\tilde{g}$ is* minimax *if the following condition holds :*

$$\alpha(\tilde{g}) = \min_g \alpha(g) \tag{4.2.7}$$

Minimax and Bayes tests have strong connections. Sometimes it is possible to find M *a priori* probabilities q_j, which maximize the weighted error probability of all the Bayesian tests. Such a set of *a priori* probabilities is called a least favorable distribution. Then the Bayes test that corresponds to this least-favorable distribution is the minimax test. Similarly, MP and Bayes tests also have strong connections. For an appropriate choice of the *a priori* probabilities q_j, a Bayes test $\bar{g}$ is a MP test in the class K. See [Lehmann, 1986, Borovkov, 1984] for further details.

4.2.2 Testing Between Two Simple Hypotheses

Testing between two simple hypotheses $\mathbf{H}_0$ and $\mathbf{H}_1$ is an important special case of the problem of testing between M simple hypotheses. In this case, the error probability of type I $\alpha_0(g)$ is called the *size* of the test. The value $1 - \alpha_0(g)$ is called the *level* of the test g. The value $\beta(g) = 1 - \alpha_1(g)$ is called the *power* of the test. Let us define the *critical function* $g(\mathcal{Y}_1^N)$, for which we use the same notation as for the statistical test, because this function completely characterizes the test g. The test assigns a real number $g(\mathcal{Y}_1^N)$ such that $0 \leq g(\mathcal{Y}_1^N) \leq 1$, to the conditional probability $\mathbf{P}_\theta$(test $g(\mathcal{Y}_1^N)$ accepts $\mathbf{H}_1 | \mathcal{Y}_1^N$) for each point $\mathcal{Y}_1^N \in \Omega^N$. This function defines the probability of acceptance of hypothesis $\mathbf{H}_1$. We assume that $g(\mathcal{Y}_1^N)$ is the indicator function of the critical region Ω_1 and has only two values, 0 and 1. The size and the power of the test g can be computed as follows :

$$\alpha_0(g) = \mathbf{E}_0[g(\mathcal{Y}_1^N)] \tag{4.2.8}$$

$$\beta(g) = 1 - \alpha_1(g)$$

$$= \mathbf{E}_1[g(\mathcal{Y}_1^N)] \tag{4.2.9}$$

where $\mathbf{E}_0$ and $\mathbf{E}_1$ are the expectations under hypotheses $\mathbf{H}_0$ and $\mathbf{H}_1$, respectively.

We now give the fundamental result known as the Neyman-Pearson lemma.

Theorem 4.2.1 (Neyman-Pearson). *The three following statements hold.*

- *Let $\mathbf{P}_0$ and $\mathbf{P}_1$ be two probability distributions with densities p_0 and p_1 with respect to some probability measure μ (for example, $\mathbf{P}_0 + \mathbf{P}_1$). For testing between $\mathbf{H}_0$ and $\mathbf{H}_1$, there exists a test $g(\mathcal{Y}_1^N)$ and a constant λ_α such that*

$$\mathbf{E}_0[g(\mathcal{Y}_1^N)] = \alpha_0 \tag{4.2.10}$$

$$g(\mathcal{Y}_1^N) = \begin{cases} 1, & \textit{when} \quad \frac{p_1(\mathcal{Y}_1^N)}{p_0(\mathcal{Y}_1^N)} \geq \lambda_\alpha \\ 0, & \textit{when} \quad \frac{p_1(\mathcal{Y}_1^N)}{p_0(\mathcal{Y}_1^N)} < \lambda_\alpha \end{cases} \tag{4.2.11}$$

- *If a test g satisfies the relations (4.2.10) and (4.2.11) for some constant λ_α, then this test is the MP test with level $1 - \alpha_0$; thus, this gives a sufficient condition for the existence of a MP test.*
- *If a test g is a MP test with level $1 - \alpha_0$, then for some constant λ_α, it satisfies the relation (4.2.11) almost surely with respect to μ. This test also satisfies (4.2.10), except if another test exists with size lower than α_0 and power 1. This gives a necessary condition for MP test.*

It follows from this theorem that an optimal MP test is necessarily based upon the *likelihood ratio* (LR) :

$$\Lambda(\mathcal{Y}_1^N) = \frac{p_1(\mathcal{Y}_1^N)}{p_0(\mathcal{Y}_1^N)} \tag{4.2.12}$$

But the LR test is also optimal with respect to the two other criteria mentioned above, namely Bayes and minimax [Borovkov, 1984, Lehmann, 1986].

Example 4.2.1 (Mean in a Gaussian sequence - contd.). *Consider again the example of testing the mean value θ in an independent Gaussian sequence $\mathcal{Y}_1^N$ with variance σ^2. The two hypotheses are $\mathbf{H}_i : \theta = \theta_i$, $(i = 0, 1)$. From the Neyman-Pearson lemma and the independence property (3.1.57) the optimal test with level $1 - \alpha$ can be written as*

$$\prod_{i=1}^N \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} = \prod_{i=1}^N \frac{\varphi\left(\frac{y_i - \theta_1}{\sigma}\right)}{\varphi\left(\frac{y_i - \theta_0}{\sigma}\right)} \underset{\mathbf{H}_0}{\overset{\mathbf{H}_1}{\gtrless}} \lambda_\alpha \tag{4.2.13}$$

where φ is the Gaussian density, or

$$\frac{\theta_1 - \theta_0}{\sigma^2} \left(\sum_{i=1}^N y_i - N \, \frac{\theta_1 + \theta_0}{2} \right) \underset{\mathbf{H}_0}{\overset{\mathbf{H}_1}{\gtrless}} \ln \lambda_\alpha \tag{4.2.14}$$

Note here that the critical function of the optimal test is based upon the sufficient statistic $S_N(y) = \sum_{i=1}^N y_i$, which is the log-likelihood ratio. This illustrates the fact that the likelihood ratio is a sufficient statistic.

We now generalize the Neyman-Pearson result to the case of an exponential family, as in (4.1.51).

Example 4.2.2 (Exponential family - contd.). *Testing the parameter θ of an exponential family (4.1.51) can be achieved in an optimal manner through the use of the log-likelihood ratio of the observations $\mathcal{Y}_1^N$ and thus the sufficient statistic $S_N(y) = \sum_{i=1}^N T(y_i)$.*

Example 4.2.3 (Testing between two χ^2 distributions). *Using the two expressions given in example 3.1.4 for the density of a χ^2 distribution, it results from Neyman-Pearson lemma that testing against zero the*

noncentrality parameter λ of a χ^2 distribution with n degrees of freedom can be optimally achieved through the use of the likelihood ratio, which is the sufficient statistic

$$\begin{aligned}\Lambda(y) &= \frac{p_\lambda(y)}{p_0(y)} \\ &= e^{-\frac{\lambda}{2}} \sum_{i=0}^{\infty} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n}{2}+i)i!} \left(\frac{\lambda y}{4}\right)^i \\ &= e^{-\frac{\lambda}{2}} \left[1 + \sum_{i=1}^{\infty} \frac{1}{\frac{n}{2}(\frac{n}{2}+1)\ldots(\frac{n}{2}+i-1)i!} \left(\frac{\lambda y}{4}\right)^i\right] \end{aligned} \tag{4.2.15}$$

It is important for the subsequent chapters to introduce asymptotic points of view for the investigation of the properties of the tests.

4.2.3 Asymptotic Points of View

The investigation of the properties of a test g consists of computing the value λ_α and the probabilities $\alpha_0(g)$ and $\beta(g)$. It should be clear that these computations require the knowledge of the cumulative distribution function of the likelihood ratio Λ. However, the distribution of the likelihood ratio is non-Gaussian in general and, moreover, it is difficult to compute; thus, the tuning of the threshold and the computation of the error probabilities are tricky. It is therefore of interest to discuss some approximate solutions to this problem for large samples of i.i.d. random variables. We have to compute the two following probabilities :

$$\begin{aligned}\alpha_0(g_\lambda) &= \mathbf{P}_0\left[\sum_{i=1}^{N} \ln \frac{p_1(y_i)}{p_0(y_i)} \geq \ln \lambda\right] \\ \alpha_1(g_\lambda) &= \mathbf{P}_1\left[\sum_{i=1}^{N} \ln \frac{p_1(y_i)}{p_0(y_i)} < \ln \lambda\right] \end{aligned} \tag{4.2.16}$$

Let $N \to +\infty$. The asymptotic point of view for getting approximations for these probabilities consists of replacing the test g by a sequence of tests for each N. There are two possible asymptotic approaches, which we describe now [Borovkov, 1984]. The first is called the *large deviation approach* and assumes that the distributions $\mathbf{P}_0$ and $\mathbf{P}_1$ are fixed. In other words, the distance between them does not depend upon the sample size N. The second approach is called the *local (hypotheses) approach* and assumes that the distance between p_0 and p_1 depends on N in such a way that the two hypotheses get closer to each other when N grows to infinity, which we formalize as

$$\mathbf{K}(p_0, p_1) + \mathbf{K}(p_1, p_0) = \mathbf{J}(p_0, p_1) \to 0 \text{ when } N \to \infty \tag{4.2.17}$$

We concentrate on this second approach because we use it extensively throughout the book. Let us first begin with an informal presentation of this local approach in the *scalar* parametric case. Assume that $\theta_1 = \theta_0 + \nu(N)$ with $\nu(N) \to 0$ when $N \to \infty$. As in subsection 4.1.2, let us consider the Taylor expansion of the logarithm of the likelihood ratio with respect to the small variable ν :

$$\sum_{i=1}^{N} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \approx \nu(N) \sum_{i=1}^{N} \frac{\partial \ln p_{\theta_0}(y_i)}{\partial \theta_0} \tag{4.2.18}$$

In this formula, we recover the efficient score for θ, which we introduced in subsection 4.1.2 :

$$z_i = \frac{\partial \ln p_{\theta_0}(y_i)}{\partial \theta_0} \tag{4.2.19}$$

Let $\mathcal{Z}_N = \sum_{i=1}^N z_i$. As shown in subsection 4.1.2,

$$\begin{aligned} \mathbf{E}_{\theta_0}(\mathcal{Z}_N) &= 0 \\ \mathbf{E}_{\theta_1}(\mathcal{Z}_N) &\approx \nu(N)\ N\ \mathbf{I}(\theta_1) \approx \nu(N)\ N\ \mathbf{I}(\theta_0) \\ \mathrm{var}_{\theta_0}(\mathcal{Z}_N) &= N\ \mathbf{I}(\theta_0) \\ \mathrm{var}_{\theta_1}(\mathcal{Z}_N) &\approx N\ \mathbf{I}(\theta_1) \approx N\ \mathbf{I}(\theta_0) \end{aligned} \tag{4.2.20}$$

where $\mathbf{I}(\theta) = \mathbf{E}_\theta(z_i^2)$ is the Fisher information. Now let us discuss the relevant speed of convergence of $\nu(N)$ for which this asymptotic framework remains meaningful. It is obvious that the two local hypotheses $\mathbf{H}_0$ and $\mathbf{H}_1$ are separable when the order of magnitude of the quantity

$$\mathbf{E}_{\theta_1}(\mathcal{Z}_N) - \mathbf{E}_{\theta_0}(\mathcal{Z}_N) \approx \nu(N)\ N\ \mathbf{I}(\theta_1) \tag{4.2.21}$$

is greater than or equal to $\sqrt{\mathrm{var}_{\theta_i}(\mathcal{Z}_N)} \approx \sqrt{N\ \mathbf{I}(\theta_0)}$ for $i = 1, 2$. In other words, the following condition must hold : $\nu(N) = \frac{\nu}{\sqrt{N}}$. This concludes our informal introduction to local hypotheses testing.

A more formal mathematical derivation of this approach is based upon the asymptotic expansion of the likelihood ratio (see subsection 4.2.9 for the vector parameter case), which holds under some regularity assumptions :

$$\begin{aligned} \sum_{i=1}^N \ln \frac{p_{\theta+\frac{\nu}{\sqrt{N}}}(y_i)}{p_\theta(y_i)} &= \frac{\nu}{\sqrt{N}} \sum_{i=1}^N \frac{\partial \ln p_\theta(y_i)}{\partial \theta} - \frac{1}{2}\nu^2\left(\mathbf{I}(\theta) + \epsilon_N\right) \\ &= \frac{\nu}{\sqrt{N}}\mathcal{Z}_N - \frac{1}{2}\nu^2\left(\mathbf{I}(\theta) + \epsilon_N\right) \end{aligned} \tag{4.2.22}$$

where $\epsilon_N \to 0$ almost surely. It follows from the asymptotic local theory [Roussas, 1972] that

$$\mathcal{L}\left(\frac{1}{\sqrt{N}}\mathcal{Z}_N\right) \rightsquigarrow \begin{cases} \mathcal{N}(0, \mathbf{I}(\theta)) & \text{when} \quad \mathcal{L}(Y) = \mathbf{P}_\theta \\ \mathcal{N}(\nu\ \mathbf{I}(\theta), \mathbf{I}(\theta)) & \text{when} \quad \mathcal{L}(Y) = \mathbf{P}_{\theta+\frac{\nu}{\sqrt{N}}} \end{cases} \tag{4.2.23}$$

where $\rightsquigarrow$ corresponds to the weak convergence. In other words, the distribution of the random variable $\frac{1}{\sqrt{N}}\mathcal{Z}_N$ weakly converges to the normal one when $N \to \infty$.

This normal approximation together with the expansion (4.2.22) leads to the following approximations for $\alpha_0(g_\lambda)$ and $\alpha_1(g_\lambda)$ defined in (4.2.16), which are based on

$$\begin{aligned} \alpha_0(g_\lambda) &\approx 1 - \phi\left[\frac{\nu^2\mathbf{I}(\theta_0) + 2\ln\lambda}{2|\nu|\sqrt{\mathbf{I}(\theta_0)}}\right] \\ \alpha_1(g_\lambda) &\approx \phi\left[\frac{-\nu^2\mathbf{I}(\theta_0) + 2\ln\lambda}{2|\nu|\sqrt{\mathbf{I}(\theta_0)}}\right] \end{aligned} \tag{4.2.24}$$

where $\phi(x)$ is the Gaussian cdf defined in (3.1.14).

Definition 4.2.6 (Asymptotic equivalence of tests). *Two tests $g_0(y)$ and $g_1(y)$ are said to be* asymptotically equivalent *when*

$$\limsup_{N\to\infty} |\alpha_j(g_0) - \alpha_j(g_1)| = 0 \quad (j = 0, 1) \tag{4.2.25}$$

A test that is asymptotically equivalent to another test that is MP is called an asymptotically MP test or AMP.

It results from the local approach and the approximation (4.2.18) that the efficient score is a sufficient statistic for testing between $\mathbf{H}_0$ and $\mathbf{H}_1$. Therefore, using the Gaussian approximation (4.2.23), we deduce that the test defined by

$$\text{sign}(\nu)\ \frac{1}{\sqrt{N}}\ \mathcal{Z}_N \underset{\mathbf{H}_0}{\overset{\mathbf{H}_1}{\gtrless}} \text{sign}(\nu)\ \frac{\nu^2\ \mathbf{I}(\theta) + 2\ln\lambda}{2|\nu|} \tag{4.2.26}$$

is asymptotically equivalent to the MP test $g(\mathcal{Y}_1^N)$ in (4.2.11) and thus is an AMP test.

The two asymptotic approaches, namely the large deviation and the asymptotic local approaches, are compared for testing the parameter of an exponential distribution in [Borovkov, 1984]. The observations are supposed to be distributed according to the following pdf $p_\theta(x) = \theta e^{-\theta x}, x \geq 0$, and the problem is to test between hypotheses $\mathbf{H}_0 = \{\theta_0 = 1\}$ and $\mathbf{H}_1 = \{\theta_1 = 1 - \nu\}$, where $\nu = 0.5, 0.2, 0.1$. Using fixed size $\alpha_0(g_\lambda) = 0.023$ and for sample sizes $N = 30, 100, 300, 1000$, it was shown that the best approximation of $\alpha_1(g_\lambda)$ is given by the large deviations approach when $(\theta_0 - \theta_1)\sqrt{N} > 3$, and by the local approach when $(\theta_0 - \theta_1)\sqrt{N} < 3$.

4.2.4 Composite Hypotheses Testing Problems

Definition 4.2.7 (Composite hypothesis). *Any nonsimple hypothesis is called a* composite hypothesis.

Let us define a composite hypotheses testing problem in the vector parametric case in the following manner :

$$\mathbf{H}_i = \{\mathcal{L}(\mathcal{Y}_1^N) = \mathbf{P}_\theta; \theta \in \Theta_i\},\ \ \Theta_i \subset \Theta,\ \ i = 0, 1\ \ \Theta_0 \cap \Theta_1 = \emptyset \tag{4.2.27}$$

The quality of a composite hypotheses test can be defined by generalization of the criteria used for the simple hypotheses case. The *size* $\alpha_0(g)$ of a test is defined by

$$\alpha_0(g) = \sup_{\theta \in \Theta_0} \mathbf{E}_\theta[g(\mathcal{Y}_1^N)] \tag{4.2.28}$$

The *level* of a test is $1 - \alpha_0(g)$ as before. The *power* of a test $g(\mathcal{Y}_1^N)$ is now a function of θ and is defined by

$$\beta_g(\theta) = \mathbf{E}_\theta[g(\mathcal{Y}_1^N)]\ ,\ \theta \in \Theta_1 \tag{4.2.29}$$

This function is often called the *power function* of the test.

Definition 4.2.8 (UMP test). *A test $g^*(\mathcal{Y}_1^N)$ is said to be* uniformly most powerful (UMP) *in the class of tests K_ϵ with fixed size $\epsilon = \alpha_0(g)$ if, for all other tests $g \in K_\epsilon$, the following relationship holds :*

$$\forall \theta \in \Theta_1,\ \ \beta_{g^*}(\theta) \geq \beta_g(\theta) \tag{4.2.30}$$

This definition is illustrated in figure 4.1. In this figure, the power function $\beta_{g^*}(\theta)$ of the UMP test corresponds to the solid line, and the dotted line corresponds to the power function of any other test in the class K_ϵ.

Definition 4.2.9 (Monotone LR). *A parametric family of densities $\mathcal{P}$ with* scalar *parameter θ is said to have a* monotone likelihood ratio *(LR) if there exists a function $T(\mathcal{Y}_1^N)$ such that, for all θ and θ_0 where $\theta > \theta_0$, the ratio*

$$\Lambda(\mathcal{Y}_1^N) = \frac{p_\theta(\mathcal{Y}_1^N)}{p_{\theta_0}(\mathcal{Y}_1^N)} = \Lambda[T(\mathcal{Y}_1^N)] \tag{4.2.31}$$

is a nondecreasing (respectively nonincreasing) function of $T(\mathcal{Y}_1^N)$.

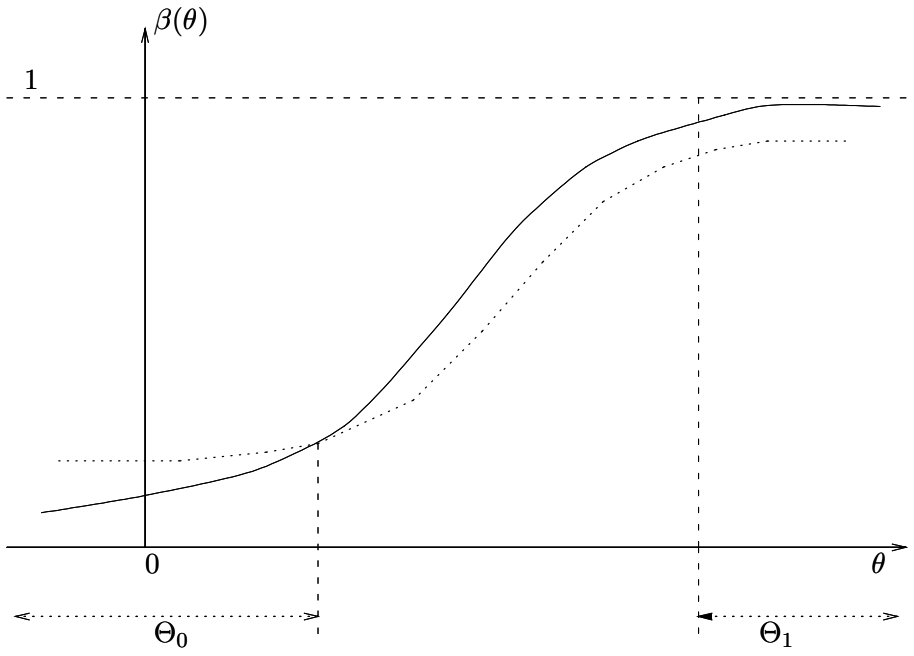

Figure 4.1 The power function of a UMP test (solid line) and of another test (dotted line).

Note that $T(\mathcal{Y}_1^N)$ is a sufficient statistic.

Theorem 4.2.2 *[Lehmann, 1986] Assume that the random variable Y has a density p_θ where θ is scalar and the family $(p_\theta)_\theta$ has a monotone likelihood ratio. The following results hold :*

- *For testing between hypothesis $\mathbf{H}_0 = \{\theta \leq \theta_0\}$ and the* one-sided *alternative hypothesis $\mathbf{H}_1 = \{\theta > \theta_0\}$, there exists a UMP test in the class of tests K_ϵ defined by*

$$g^*(\mathcal{Y}_1^N) = \begin{cases} 1 & \textit{when} \quad T(\mathcal{Y}_1^N) > \lambda \\ p & \textit{when} \quad T(\mathcal{Y}_1^N) = \lambda \\ 0 & \textit{when} \quad T(\mathcal{Y}_1^N) < \lambda \end{cases} \tag{4.2.32}$$

  where the constants p and λ are such that

$$\mathbf{E}_{\theta_0}[g^*(\mathcal{Y}_1^N)] = \mathbf{P}_{\theta_0}[T(\mathcal{Y}_1^N) > \lambda] + p\, \mathbf{P}_{\theta_0}[T(\mathcal{Y}_1^N) = \lambda] = \epsilon \tag{4.2.33}$$

- *The power function $\beta_{g^*}(\theta)$ is a strictly increasing function for all points θ for which $\beta_{g^*}(\theta) < 1$.*
- *For all $\tilde{\theta}$, the test $g^*(\mathcal{Y}_1^N)$ is UMP in the class $K_{\beta_{g^*}(\tilde{\theta})}$ for testing between $\mathbf{H}_0 = \{\theta \leq \tilde{\theta}\}$ and $\mathbf{H}_1 = \{\theta > \tilde{\theta}\}$.*
- *For any $\theta < \theta_0$, the test $g^*(\mathcal{Y}_1^N)$ minimizes the power function $\beta_g(\theta) = \mathbf{E}_\theta[g(\mathcal{Y}_1^N)]$ in the class K_ϵ.*

Let us discuss an important consequence of theorem 4.2.2. Assume that the family $\mathcal{P}$ with monotone LR is a single-parameter exponential family with pdf :

$$p_\theta(y) = h(y)e^{c(\theta)T(y)-d(\theta)} \tag{4.2.34}$$

In the case of an i.i.d. sample $\mathcal{Y}_1^N$, the LR is

$$\Lambda(\mathcal{Y}_1^N) = \frac{p_\theta(\mathcal{Y}_1^N)}{p_{\theta_0}(\mathcal{Y}_1^N)} = e^{[c(\theta)-c(\theta_0)]\sum_{i=1}^N T(y_i) - N[d(\theta)-d(\theta_0)]} \tag{4.2.35}$$

and is a monotone function of $\check{T}(\mathcal{Y}_1^N) = \sum_{i=1}^N T(y_i)$, provided that $[c(\theta) - c(\theta_0)]$ has a constant sign for all θ, θ_0 such that $\theta > \theta_0$. From theorem 4.2.2, we deduce that there exists a UMP test given by (4.2.32) and (4.2.33) for testing between hypotheses $\mathbf{H}_0 = \{\theta \leq \theta_0\}$ and $\mathbf{H}_1 = \{\theta > \theta_0\}$ when $c(\theta)$ is an increasing function. When this function is decreasing, the three inequalities in (4.2.32) and (4.2.33) should be replaced by their converse.

Up to now, we have discussed only one-sided alternative hypotheses. Another important case is that of the two-sided alternatives. In other words, we test hypothesis $\mathbf{H}_0 = \{\theta = \theta_0\}$ against $\mathbf{H}_1 = \{\theta \neq \theta_0\}$. But no UMP test exists in this case [Lehmann, 1986, Borovkov, 1984].

4.2.5 Unbiased Tests

It follows from the previous paragraph that the UMP test exists only in special cases. Let us introduce the so-called subclass $\bar{K}_\epsilon$ of unbiased tests in the class of UMP tests :

$$K_\epsilon = \{g \ : \sup_{\theta\in\Theta_0} \mathbf{E}_\theta[g(\mathcal{Y}_1^N)] \leq \epsilon\} \tag{4.2.36}$$

Definition 4.2.10 (Unbiased test). *A test $g \in K_\epsilon$ is said to be* unbiased *if the following condition holds :*

$$\inf_{\theta\in\Theta_1} \mathbf{E}_\theta[g(\mathcal{Y}_1^N)] \geq \sup_{\theta\in\Theta_0} \mathbf{E}_\theta[g(\mathcal{Y}_1^N)] \tag{4.2.37}$$

Note that this condition is very natural, because the probability of rejection of hypothesis $\mathbf{H}_0$ when $\mathbf{H}_0$ is false must be not less than the probability of rejection of $\mathbf{H}_0$ when it is true.

It turns out that for the exponential family (4.2.34), there exists a UMP unbiased test $g(\mathcal{Y}_1^N)$ in the class $\bar{K}_\epsilon$ with level $1-\epsilon$ for two-sided alternative hypotheses. Let us assume that the pdf $p_\theta(y)$ belongs to the family (4.2.34) and that we want to test hypothesis $\mathbf{H}_0 = \{\theta \in (\theta_0, \theta_1)\}$ against $\mathbf{H}_1 = \{\theta \notin (\theta_0, \theta_1)\}$, where $\theta_0 \leq \theta_1$. The UMP unbiased test is given by

$$g(\mathcal{Y}_1^N) = \begin{cases} 0 & \text{when} \quad \lambda_0 < \check{T}(\mathcal{Y}_1^N) < \lambda_1 \\ p_i & \text{when} \quad \check{T}(\mathcal{Y}_1^N) = \lambda_i \ \ (i = 0, 1) \\ 1 & \text{when} \quad \check{T}(\mathcal{Y}_1^N) \notin (\lambda_0, \lambda_1) \end{cases} \tag{4.2.38}$$

where the constants λ_i and p_i are determined by

$$\begin{cases} \mathbf{E}_{\theta_i}[g(\mathcal{Y}_1^N)] = \epsilon & \text{if} \quad \theta_0 < \theta_1 \\ \mathbf{E}_{\theta_0}[g(\mathcal{Y}_1^N)] = \epsilon \\ \mathbf{E}_{\theta_0}\left\{[g(\mathcal{Y}_1^N) - \epsilon]\check{T}(\mathcal{Y}_1^N)\right\} = 0 & \text{if} \quad \theta_0 = \theta_1 \end{cases} \tag{4.2.39}$$

4.2.6 Bayesian and Minmax Approaches for Composite Hypotheses

Let us consider the composite hypotheses $\mathbf{H}_0 = \{\theta \in \Theta_0\}$ and $\mathbf{H}_1 = \{\theta \in \Theta_1\}$. The Bayesian approach consists of introducing *a priori* probabilities for these hypotheses and *a priori* distributions $\mathbf{P}(\theta)$ for the

parameter θ on the set $\Theta = \Theta_0 \cup \Theta_1$. The distribution $\mathbf{P}$ is generated by the probabilities $q_0 = \mathbf{P}(\mathbf{H}_0)$ and $q_1 = \mathbf{P}(\mathbf{H}_1)$ such that $q_0 + q_1 = 1$, and the distributions $\mathbf{P}_0(\theta)$ for $\theta \in \Theta_0$ and $\mathbf{P}_1(\theta)$ for $\theta \in \Theta_1$. In other words, we have

$$\mathbf{P}(\theta) = q_0\mathbf{P}_0(\theta) + q_1\mathbf{P}_1(\theta) \tag{4.2.40}$$

Therefore, the observations $\mathcal{Y}_1^N$ have the pdf $p_i(\mathcal{Y}_1^N)$ under hypothesis $\mathbf{H}_i$ given by

$$p_i(\mathcal{Y}_1^N) = \int_{\Theta_i} p_\theta(\mathcal{Y}_1^N) d\mathbf{P}_i(\theta) \tag{4.2.41}$$

A *Bayesian test* $\bar{g}(\mathcal{Y}_1^N)$ for testing composite hypotheses can be written as [Lehmann, 1986, Borovkov, 1984]

$$\bar{g}(\mathcal{Y}_1^N) = \begin{cases} 1 & \text{when} \quad \frac{p_1(\mathcal{Y}_1^N)}{p_0(\mathcal{Y}_1^N)} > \lambda \\ p & \text{when} \quad \frac{p_1(\mathcal{Y}_1^N)}{p_0(\mathcal{Y}_1^N)} = \lambda \\ 0 & \text{when} \quad \frac{p_1(\mathcal{Y}_1^N)}{p_0(\mathcal{Y}_1^N)} < \lambda \end{cases} \tag{4.2.42}$$

where $\lambda = \frac{q_0}{q_1}$ and $p \in (0, 1)$ is arbitrary.

Definition 4.2.11 (Minimax test). *Consider again the class K_ϵ defined in (4.2.36) for testing composite hypotheses. The test $\tilde{g}$ is said to be* minimax *(or minmax) in the class K_ϵ if it maximizes the minimum power :*

$$\inf_{\theta \in \Theta_1} \mathbf{E}_\theta[g(\mathcal{Y}_1^N)] = \inf_{\theta \in \Theta_1} \beta(\theta) \tag{4.2.43}$$

If sets Θ_0 and Θ_1 have contact on one point, and if the power function is continuous, then the inequality $\sup_{g \in K_\epsilon} \inf_{\theta \in \Theta_1} \beta(\theta) > \epsilon$ cannot hold. In this situation, it is of interest to introduce an indifference (dead) zone as in figure 4.2, or in other words, to separate sets Θ_0 and Θ_1. From a practical point of view, this is not a major drawback, because it is well known that a value θ always exists between the hypotheses, and all choices of this point have the same likelihood.

If sets Θ_0 and Θ_1 do contact, any unbiased test g is minimax. The converse statement is true in general. Another important property is that the UMP unbiased test g in the class $\bar{K}_\epsilon$ is a minimax test in the class K_ϵ.

Now let us recall that there exists a theorem [Borovkov, 1984] that shows that the Bayes test given by (4.2.42) is minimax if there exists a pair of distributions $\mathbf{P}_0(\theta)$ and $\mathbf{P}_1(\theta)$ that is least favorable in some sense. The main difficulty in using this theorem is to guess these least favorable distributions $\mathbf{P}_0(\theta)$ and $\mathbf{P}_1(\theta)$. In many cases, for some families of distributions, it is useful to use invariance properties with respect to some transformations in order to guess least favorable distributions.

Definition 4.2.12 (Invariance). *A parametric family of distributions $\mathcal{P} = \{\mathbf{P}_\theta\}_{\theta \in \Theta}$ remains* invariant *under a group of transformation $\mathcal{G}$ if*

$$\forall g \in \mathcal{G} \ \text{ and } \forall \theta \in \Theta, \ \exists \theta_g \in \Theta \ \text{ such that : } \ \mathbf{P}_\theta(Y \in A) = \mathbf{P}_{\theta_g}(Y \in gA) \tag{4.2.44}$$

We shall note $\theta_g = \bar{g}\theta$.

Now we consider an important example that is widely used in several other chapters.

Example 4.2.4 (Gaussian vector sequence - Unit covariance matrix). *We consider an r-dimensional random vector Y, with distribution $\mathcal{L}(Y) = \mathcal{N}(\theta, I)$. Let us define the squared norm : $\|\theta\|^2 = \sum_{i=1}^r \theta_i^2$. Consider the problem of testing between the two following hypotheses :*

$$\mathbf{H}_0 = \{\theta \ : \|\theta\| \leq a\} \ \text{ and } \ \mathbf{H}_1 = \{\theta \ : \|\theta\| \geq b\} \ \text{ where } \ b > a \tag{4.2.45}$$

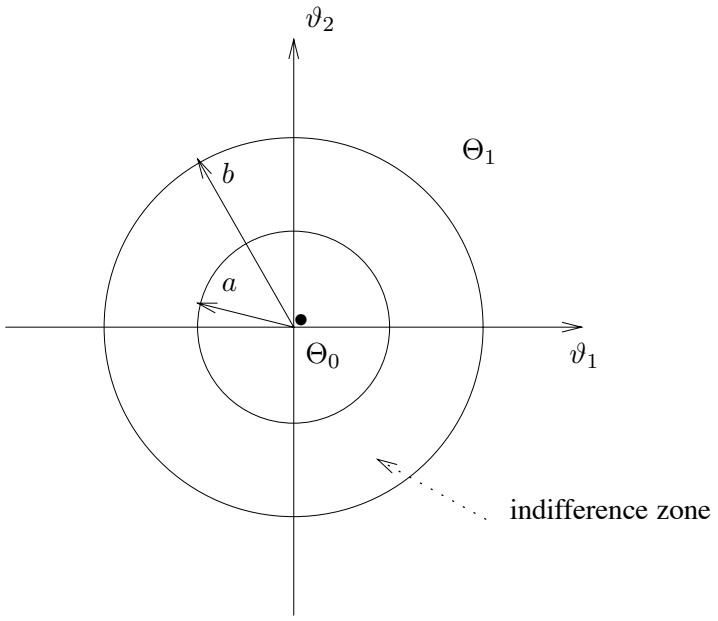

Figure 4.2 Indifference zone between the two hypotheses.

In other words, we have an indifference zone between $\mathbf{H}_0$ *and* $\mathbf{H}_1$. *These hypotheses are depicted in the figure 4.2. We use the pdf of an independent Gaussian random sequence of dimension* r *and sample size* N *:*

$$p_\theta(\mathcal{Y}_1^N) = \frac{1}{(2\pi)^{\frac{rN}{2}}} e^{-\frac{1}{2}\sum_{i=1}^N (Y_i-\theta)^T(Y_i-\theta)} \tag{4.2.46}$$

It is possible to prove that the family of normal distributions $\mathcal{N}(\theta, I)$ *remains invariant under every orthogonal transformation* $gx = Cx$, *where* C *is the matrix of the orthogonal transformation. In this case, the corresponding transformation* $\bar{g}$ *of the parameter set* Θ *can be defined by* $\bar{g}\theta = C\theta$, *and hypotheses* $\mathbf{H}_i$ *remain invariant under the transformation* $\bar{g}$.

It turns out that, first, the least favorable distribution $\mathbf{P}_i(\theta)$ *must remain invariant under the transformation* $\bar{g}$, *and second, this distribution must be concentrated on the boundary of* Θ_i. *Therefore, it follows that the least favorable distributions* $\mathbf{P}_0(\theta)$ *and* $\mathbf{P}_1(\theta)$ *are uniform distributions on the spheres* $\tilde{\Theta}_0 = \{\theta : \|\theta\| = a\}$ *and* $\tilde{\Theta}_1 = \{\theta : \|\theta\| = b\}$. *In this case,* the minmax test is a Bayes test $\tilde{g}(\mathcal{Y}_1^N) = \bar{g}(\mathcal{Y}_1^N)$, *where* $\bar{g}(\mathcal{Y}_1^N)$ *is a Bayes test which can be written as*

$$\frac{\frac{1}{V_1}\int_{\tilde{\Theta}_1} \exp\left[-\frac{1}{2}\sum_{i=1}^N (Y_i-\theta)^T(Y_i-\theta)\right] dV(\theta)}{\frac{1}{V_0}\int_{\tilde{\Theta}_0} \exp\left[-\frac{1}{2}\sum_{i=1}^N (Y_i-\theta)^T(Y_i-\theta)\right] dV(\theta)} \begin{array}{c} H_1 \\ \gtrless \\ H_0 \end{array} \lambda \tag{4.2.47}$$

where $dV(\theta)$ *is the surface element of the sphere and* $V_i = \int_{\tilde{\Theta}_i} dV(\theta)$. *After straightforward computations, we get*

$$\exp\left[\frac{-N(b^2-a^2)}{2}\right] \frac{\frac{1}{V_1}\int_{\tilde{\Theta}_1} \exp\left[\frac{1}{N}\sum_{i=1}^N (Y_i^T\theta)\right] dV(\theta)}{\frac{1}{V_0}\int_{\tilde{\Theta}_0} \exp\left[\frac{1}{N}\sum_{i=1}^N (Y_i^T\theta)\right] dV(\theta)} \begin{array}{c} H_1 \\ \gtrless \\ H_0 \end{array} \lambda \tag{4.2.48}$$

Each of the integrals in equation (4.2.48) can be written with the aid of

$$\Omega(z) = \frac{1}{V}\int_{\tilde{\Theta}} \exp\left(z\mathcal{E}^T\theta\right) dV(\theta) \tag{4.2.49}$$

where $\mathcal{E} = \frac{\bar{Y}_N}{\|\bar{Y}_N\|}$, $\bar{Y}_N = \frac{1}{N}\sum_{i=1}^N Y_i$, *and* $\tilde{\Theta}$ *is the unit sphere. More precisely, the test can be written as*

$$\frac{\Omega(\|\bar{Y}_N\|b)}{\Omega(\|\bar{Y}_N\|a)} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda_0 \tag{4.2.50}$$

where $\lambda_0 = \lambda_0(\lambda, a, b)$. *Because* $\Omega(z)$ *is a convex increasing function of* z *when* $z \in [0, \infty)$, *the minmax test can finally be written as*

$$\tilde{g}(\mathcal{Y}_1^N) = \begin{cases} 1 & when \quad \|\bar{Y}_N\|^2 \geq \lambda_\epsilon^2 \\ 0 & when \quad \|\bar{Y}_N\|^2 < \lambda_\epsilon^2 \end{cases} \tag{4.2.51}$$

Note here that $\|\bar{Y}_N\|$ *is a* sufficient statistic. *This is obvious from (4.2.50) and the Neyman-Fisher factorization theorem. The power function of the test is*

$$\beta(\|\theta\|) = \mathbf{E}_\theta[\tilde{g}(\mathcal{Y}_1^N)] = \mathbf{P}_\theta(\|\xi - \theta\|^2 \geq N\lambda_\epsilon^2) \tag{4.2.52}$$

where $\mathcal{L}(\xi) = \mathcal{N}(0, I)$. *Now the random variable* $\Xi = \|\xi - \theta\|^2$ *is distributed as a* $\chi'^2(r, c^2)$ *variable, where* $c^2 = \|\theta\|^2$ *is the noncentrality parameter. It can be easily shown that* $\beta(\|\theta\|)$ *is an increasing function of* $\|\theta\|$ *for all* λ_ϵ. *Thus, the threshold* λ_ϵ *is determined by*

$$\sup_{\theta \in \Theta_0} \mathbf{E}_\theta[\tilde{g}(\mathcal{Y}_1^N)] = \sup_{c^2 \leq a^2} \mathbf{P}_{c^2}(\Xi \geq N\lambda_\epsilon^2) = \mathbf{P}_{a^2}(\Xi \geq N\lambda_\epsilon^2) = \epsilon \tag{4.2.53}$$

Under this condition, the guaranteed power of the test is

$$\inf_{\theta \in \Theta_1} \mathbf{E}_\theta[\tilde{g}(\mathcal{Y}_1^N)] = \mathbf{P}_{b^2}(\Xi \geq N\lambda_\epsilon^2) \tag{4.2.54}$$

We now investigate the more complex case of general covariance matrix.

Example 4.2.5 (Gaussian vector sequence - General covariance matrix). *Assume that we have an* r*-dimensional random vector* Y, $\mathcal{L}(Y) = \mathcal{N}(\theta, \Sigma)$. *Consider the problem of testing between the two following hypotheses :*

$$\mathbf{H}_0 = \{\theta : \theta^T \Sigma^{-1} \theta \leq a^2\} \ \ and \ \ \mathbf{H}_1 = \{\theta : \theta^T \Sigma^{-1} \theta \geq b^2\} \ \ where \ \ b > a \tag{4.2.55}$$

Let us show that it is possible to transform this hypotheses testing problem into the previous one. It is well known that any positive definite covariance matrix Σ *can be decomposed as*

$$\begin{aligned} \Sigma &= RR^T \\ \Sigma^{-1} &= (R^{-1})^T R^{-1} \end{aligned} \tag{4.2.56}$$

We know [Borovkov, 1984] that the family $\mathcal{N}(\theta, I)$ *remains invariant under the transformation* $gx = Rx$. *Therefore, equation (4.2.44) can be written as*

$$\Phi_{\theta,I}(A) = \Phi_{\bar{g}(\theta,I)}(gA) \tag{4.2.57}$$

where

$$\begin{aligned} \bar{g}(\theta, I) &= (R\,\theta, \Sigma) \\ A &= \{X : X^T X < c^2\} \\ g(A) &= \{Y = RX : X^T X < c^2\} = \{Y : Y^T \Sigma^{-1} Y < c^2\} \end{aligned} \tag{4.2.58}$$

The problem of testing between hypotheses $\mathbf{H}_0$ *and* $\mathbf{H}_1$ *(4.2.45) of example 4.2.4 can be transformed with the aid of* $\bar{g}$ *into*

$$\begin{aligned} \tilde{\Theta}_0 &= \{\tilde{\theta} = R\,\theta : \theta^T\theta \leq a^2\} \\ \tilde{\Theta}_1 &= \{\tilde{\theta} = R\,\theta : \theta^T\theta \geq b^2\} \end{aligned} \tag{4.2.59}$$

or

$$\begin{aligned} \tilde{\Theta}_0 &= \{\tilde{\theta} : \tilde{\theta}^T\Sigma^{-1}\tilde{\theta} \leq a^2\} \\ \tilde{\Theta}_1 &= \{\tilde{\theta} : \tilde{\theta}^T\Sigma^{-1}\tilde{\theta} \geq b^2\} \end{aligned} \tag{4.2.60}$$

Therefore, the minmax test (4.2.51) with critical region $\bar{X}^T\bar{X} \geq \lambda_\epsilon^2$*, which was derived under the assumption that* $\mathcal{L}(X) = \mathcal{N}(\theta, I)$ *for testing hypotheses (4.2.45), also holds under* $\mathcal{L}(Y) = \mathcal{N}(\theta, \Sigma)$*, where* $Y = RX$*, for testing hypotheses (4.2.55). Finally, the minmax test for testing hypotheses (4.2.55) can be written as*

$$\tilde{g}(\mathcal{Y}_1^N) = \begin{cases} 1 & \textit{when} \quad \bar{Y}_N^T\Sigma^{-1}\bar{Y}_N \geq \lambda_\epsilon^2 \\ 0 & \textit{when} \quad \bar{Y}_N^T\Sigma^{-1}\bar{Y}_N < \lambda_\epsilon^2 \end{cases} \tag{4.2.61}$$

4.2.7 Generalized Likelihood Ratio Test

The generalized likelihood ratio (GLR) test is one of the most general and important methods for solving composite hypotheses testing problems.

Definition 4.2.13 (GLR test). *We say that a test* $\hat{g}$ *is a* generalized likelihood ratio test *for testing between hypotheses* $\mathbf{H}_0 = \{\theta : \theta \in \Theta_0\}$ *and* $\mathbf{H}_1 = \{\theta : \theta \in \Theta_1\}$ *when*

$$\hat{g}(\mathcal{Y}_1^N) = \begin{cases} 1 & \textit{when} \quad \hat{\Lambda}(\mathcal{Y}_1^N) \geq \lambda_\epsilon \\ 0 & \textit{when} \quad \hat{\Lambda}(\mathcal{Y}_1^N) < \lambda_\epsilon \end{cases} \tag{4.2.62}$$

where

$$\hat{\Lambda}(\mathcal{Y}_1^N) = \frac{\sup_{\theta\in\Theta_1} p_\theta(\mathcal{Y}_1^N)}{\sup_{\theta\in\Theta_0} p_\theta(\mathcal{Y}_1^N)} \tag{4.2.63}$$

and where the constant λ_ϵ *is such that*

$$\sup_{\theta\in\Theta_0} \mathbf{E}_\theta[\hat{g}(\mathcal{Y}_1^N)] = \sup_{\theta\in\Theta_0} \mathbf{P}_\theta(\hat{\Lambda}(\mathcal{Y}_1^N) \geq \lambda_\epsilon) = \epsilon \tag{4.2.64}$$

Therefore, test $\hat{g}$ is in class K_ϵ. A relevant name for test (4.2.62) is the generalized Neyman-Pearson test, as is obvious from a comparison with (4.2.10) and (4.2.11). The precise optimal properties of the GLR test in the general case are unknown, but for many special cases, the GLR test is optimal.

Example 4.2.6 (Gaussian vector sequence - contd.). *Let us show that the minmax test given by (4.2.51) and (4.2.61) is a GLR test, in the case where* $\mathcal{L}(Y) = \mathcal{N}(\theta, I)$*. Consider the problem of testing between the two following hypotheses :* $\mathbf{H}_0 = \{\theta : \|\theta\| \leq a\}$ *and* $\mathbf{H}_1 = \{\theta : \|\theta\| \geq b\}$*, where* $b > a$*. In this case, the critical region of the GLR test can be written as*

$$S(\mathcal{Y}_1^N) = \ln \sup_{\|\theta\|\geq b} e^{-\frac{1}{2}\sum_{i=1}^N (Y_i-\theta)^T(Y_i-\theta)} - \ln \sup_{\|\theta\|\leq a} e^{-\frac{1}{2}\sum_{i=1}^N (Y_i-\theta)^T(Y_i-\theta)} \geq \ln \lambda_\epsilon \tag{4.2.65}$$

After simple transformations, we get

$$S(\mathcal{Y}_1^N) = \sup_{\|\theta\|\geq b}\left\{-\frac{N}{2}\left\|\theta - \frac{1}{N}\sum_{i=1}^{N} Y_i\right\|^2\right\} - \sup_{\|\theta\|\leq a}\left\{-\frac{N}{2}\left\|\theta - \frac{1}{N}\sum_{i=1}^{N} Y_i\right\|^2\right\} \geq \ln\lambda_\epsilon \tag{4.2.66}$$

This equation can be rewritten as

$$S(\mathcal{Y}_1^N) = \begin{cases} -\frac{N}{2}(\|\bar{Y}_N\| - b)^2 & \textit{when} \quad \|\bar{Y}_N\| \leq a \\ -\frac{N}{2}(\|\bar{Y}_N\| - b)^2 + \frac{N}{2}(\|\bar{Y}_N\| - a)^2 & \textit{when} \quad a \leq \|\bar{Y}_N\| \leq b \\ +\frac{N}{2}(\|\bar{Y}_N\| - a)^2 & \textit{when} \quad \|\bar{Y}_N\| \geq b \end{cases} \tag{4.2.67}$$

Therefore, $S(\mathcal{Y}_1^N)$ is a continuous increasing function of $\|\bar{Y}_N\|$. For this reason, the LR test $\hat{g}(\mathcal{Y}_1^N)$ (4.2.62) coincides with the minmax test $\tilde{g}(\mathcal{Y}_1^N)$ (4.2.51) for a suitable constant λ_ϵ.

4.2.8 Nuisance Parameters

Hypotheses testing problems with *nuisance parameters* are a special class of statistical procedures, which will be of interest when investigating the diagnosis or isolation issue in chapter 7. Let us define a composite hypotheses testing problem in the parametric case in the following manner :

$$\mathbf{H}_i = \{\mathcal{L}(Y) = \mathbf{P}_{\theta,\xi};\ \ \theta \in \Theta_i,\ \ \xi \in \Xi_i\}\ \ (i = 0, 1) \tag{4.2.68}$$

where θ is the *informative* parameter and ξ is the *nuisance* parameter. We mean that we are interested in detecting a change in θ from set Θ_0 to set Θ_1, while considering ξ as an *unknown parameter*. In other words, the changes in ξ are not of interest, but since this parameter of the distribution $\mathbf{P}_{\theta,\xi}$ is unknown, the design of the test is a nontrivial problem.

Let us describe one possible solution which is based upon the *minmax* approach, and let us consider the particular case of a *Gaussian vector*. For simplifying our explanation, we assume that $N = 1$ and we consider an r-dimensional random vector Y, with distribution

$$\mathcal{L}(Y) = \mathcal{N}(\mu_i, \Sigma), \ \text{ where } \ \mu_i = \begin{pmatrix} \theta_i \\ \xi_i \end{pmatrix} \tag{4.2.69}$$

4.2.8.1 Simple Hypotheses - Minmax approach

Let us start with a hypotheses testing problem that is simple for the informative parameter :

$$\mathbf{H}_0 = \{\theta = \theta_0,\ \ \xi_0 \in \Xi_0\} \ \text{ and } \ \mathbf{H}_1 = \{\theta = \theta_1,\ \ \xi_1 \in \Xi_1\} \tag{4.2.70}$$

The power function β of the optimal test for these simple hypotheses is an increasing function of the Kullback information :

$$\mathbf{K}(\mu_0, \mu_1) = \frac{1}{2}(\mu_0 - \mu_1)^T\Sigma^{-1}(\mu_0 - \mu_1) \tag{4.2.71}$$

As for the minimax test for composite hypotheses, the main idea of the minmax approach for nuisance parameters is to maximize the minimum possible power over the *unknown* parameters. Therefore, the design of the minmax test consists of finding a pair of *least favorable values* ξ_0^* and ξ_1^* for which the power of the optimal test will be minimum, and in computing the LR test for these values.

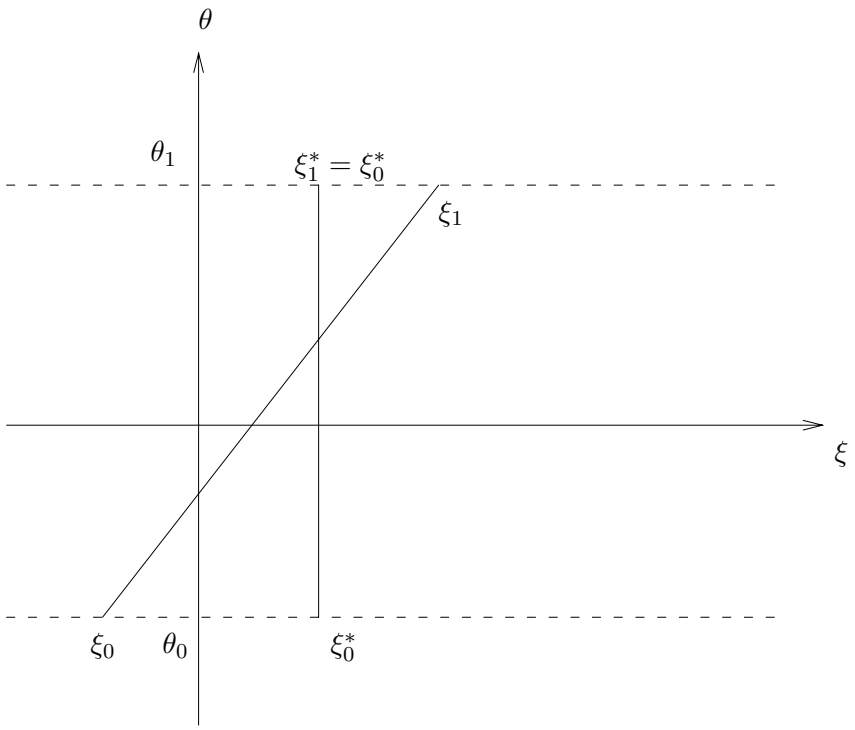

Figure 4.3 Least favorable values of nuisance parameters.

Assume first that the covariance matrix Σ of the observation Y is identity. It is obvious intuitively that, in this case, the least favorable value of the nuisance parameters ξ_0 and ξ_1 is $\xi_0^* = \xi_1^*$. A simple graphical interpretation of this idea is depicted in figure 4.3. In this case, the Kullback distance is simply the Euclidean distance, and the minimum value of the distance between two points belonging to two parallel lines is reached when they also belong to the same perpendicular to these lines.

Let us now discuss the case of a general covariance matrix Σ and find the least favorable values ξ_0^* and ξ_1^*. It is obvious that $\mathbf{K}(\mu_0, \mu_1)$ is a function of the differences $\theta = \theta_1 - \theta_0$ and $\xi = \xi_1 - \xi_0$. Therefore, we minimize $\mathbf{K}(\mu_0, \mu_1)$ with respect to the parameter ξ, and we denote it simply by $\mathbf{K}(\xi)$. Remembering the Fisher information (4.1.88) in this case, we can write

$$\mathbf{K}(\xi) = \frac{1}{2}\,(\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1) = \frac{1}{2} \begin{pmatrix} \theta \\ \xi \end{pmatrix}^T \begin{pmatrix} \mathbf{I}_{\theta\theta} & \mathbf{I}_{\theta\xi} \\ \mathbf{I}_{\xi\theta} & \mathbf{I}_{\xi\xi} \end{pmatrix} \begin{pmatrix} \theta \\ \xi \end{pmatrix} \tag{4.2.72}$$

The system of normal equations can be written in the following manner :

$$\frac{\partial \mathbf{K}(\xi)}{\partial \xi} = \mathbf{I}_{\xi\theta}\,\theta + \mathbf{I}_{\xi\xi}\,\xi = 0 \tag{4.2.73}$$

The minimum is obtained for

$$\xi^* = -\mathbf{I}_{\xi\xi}^{-1}\,\mathbf{I}_{\xi\theta}\,\theta \tag{4.2.74}$$

and is given by

$$\mathbf{K}^* = \frac{1}{2}\,\theta^T\,(\mathbf{I}_{\theta\theta} - \mathbf{I}_{\theta\xi}\,\mathbf{I}_{\xi\xi}^{-1}\,\mathbf{I}_{\xi\theta})\,\theta \tag{4.2.75}$$

On the other hand, it is known [Seber, 1977] that

$$\Sigma_{\theta\theta}^{-1} = \mathbf{I}_{\theta\theta} - \mathbf{I}_{\theta\xi}\ \mathbf{I}_{\xi\xi}^{-1}\ \mathbf{I}_{\xi\theta} \tag{4.2.76}$$

Finally, we can rewrite the minimum value of the Kullback information as

$$\mathbf{K}^* = \frac{1}{2}\ \theta^T\ \Sigma_{\theta\theta}^{-1}\ \theta \tag{4.2.77}$$

Thus, the log-likelihood ratio $S(Y)$ for the hypotheses (4.2.70) under the least favorable value of the nuisance parameter is

$$\begin{aligned} S(Y) &= (\mu_1 - \mu_0)^T \Sigma^{-1}(Y - \mu_0) - \frac{1}{2}(\mu_1 - \mu_0)^T \Sigma^{-1}(\mu_1 - \mu_0) \\ &= \begin{pmatrix} \theta_1 - \theta_0 \\ -\mathbf{I}_{\xi\xi}^{-1}\ \mathbf{I}_{\xi\theta}(\theta_1 - \theta_0) \end{pmatrix}^T \begin{pmatrix} \mathbf{I}_{\theta\theta} & \mathbf{I}_{\theta\xi} \\ \mathbf{I}_{\xi\theta} & \mathbf{I}_{\xi\xi} \end{pmatrix} \begin{pmatrix} Y_\theta - \theta_0 \\ Y_\xi - \xi_0 \end{pmatrix} - \frac{1}{2}(\theta_1 - \theta_0)^T \Sigma_{\theta\theta}^{-1}(\theta_1 - \theta_0) \\ &= (\theta_1 - \theta_0)^T (\mathbf{I}_{\theta\theta} - \mathbf{I}_{\theta\xi}\ \mathbf{I}_{\theta\theta}^{-1}\ \mathbf{I}_{\xi\theta})(Y_\theta - \theta_0) - \frac{1}{2}(\theta_1 - \theta_0)^T \Sigma_{\theta\theta}^{-1}(\theta_1 - \theta_0) \\ &= (\theta_1 - \theta_0)^T \Sigma_{\theta\theta}^{-1}(Y_\theta - \theta_0) - \frac{1}{2}(\theta_1 - \theta_0)^T \Sigma_{\theta\theta}^{-1}(\theta_1 - \theta_0) \end{aligned} \tag{4.2.78}$$

Note that the final test, which is independent of the unknown value ξ^*, turns out to be also independent of the unknown value ξ_0^*.

4.2.8.2 Equivalence Between the Minmax and GLR Approaches

In the previous paragraph, we explained the minmax approach to hypotheses testing problems in the presence of nuisance parameters. Another traditional approach to this problem is the GLR algorithm, which is based upon the maximization of the likelihood ratio with respect to the unknown nuisance parameters. Now we show the *a priori* nonobvious fact that these two approaches result in exactly the same test.

For solving the hypotheses testing problem (4.2.70), the GLR algorithm is

$$S(Y) = \ln \sup_{\xi_1} p_{\theta_1,\xi_1}(Y) - \ln \sup_{\xi_0} p_{\theta_0,\xi_0}(Y) \tag{4.2.79}$$

In this equation, the density is given by

$$2\ \ln p_{\theta,\xi}(Y) = \begin{pmatrix} Y_\theta - \theta \\ Y_\xi - \xi \end{pmatrix}^T \begin{pmatrix} \mathbf{I}_{\theta\theta} & \mathbf{I}_{\theta\xi} \\ \mathbf{I}_{\xi\theta} & \mathbf{I}_{\xi\xi} \end{pmatrix} \begin{pmatrix} Y_\theta - \theta \\ Y_\xi - \xi \end{pmatrix} \tag{4.2.80}$$

Computations similar to those made before lead to

$$Y_\xi - \xi^* = -\mathbf{I}_{\xi\xi}^{-1}\ \mathbf{I}_{\xi\theta}\ (Y_\theta - \theta) \tag{4.2.81}$$

where ξ^* is the value of ξ for which the supremum is reached. As before, the supremum can be written as

$$2\ \ln \sup_{\xi} p_{\theta,\xi}(Y) = (Y_\theta - \theta)^T \Sigma_{\theta\theta}^{-1}(Y_\theta - \theta) \tag{4.2.82}$$

from which we deduce

$$S(Y) = (\theta_1 - \theta_0)^T \Sigma_{\theta\theta}^{-1}(Y_\theta - \theta_0) - \frac{1}{2}(\theta_1 - \theta_0)^T \Sigma_{\theta\theta}^{-1}(\theta_1 - \theta_0) \tag{4.2.83}$$

which is exactly (4.2.78).

4.2.8.3 Composite Hypotheses

Let us continue our discussion about nuisance parameters with composite hypotheses testing problems. For this purpose, we start from the examples 4.2.4, 4.2.5, and 4.2.6, where we have shown that the test $\tilde{g}(Y)$, which coincides with the GLR test $\hat{g}(Y)$, is based upon the χ^2-test :

$$\tilde{g}(Y) = \left\{ \begin{array}{lll} 1 & \text{when} & Y^T\Sigma^{-1}Y \geq \lambda_\epsilon^2 \\ 0 & \text{when} & Y^T\Sigma^{-1}Y < \lambda_\epsilon^2 \end{array} \right. \tag{4.2.84}$$

We have also shown that the power of the test $\tilde{g}(Y)$ is an increasing function of $\mu^T\Sigma^{-1}\mu$. Now, let us consider the same testing problem about $\mu = \begin{pmatrix} \theta \\ \xi \end{pmatrix}$ with nuisance parameter ξ. Thus, we test between the following two hypotheses :

$$\mathbf{H}_0 = \{\theta : \theta^T\Sigma_{\theta\theta}^{-1}\theta \leq a^2, \;\; \xi_0 \in \Xi_0\} \;\text{ and }\; \mathbf{H}_1 = \{\theta : \theta^T\Sigma_{\theta\theta}^{-1}\theta \geq b^2, \;\; \xi_1 \in \Xi_1\} \tag{4.2.85}$$

where $b > a$. We again minimize the power of the test, or equivalently the quantity $\mu^T\Sigma^{-1}\mu$ with respect to ξ. Since this quantity is equal to twice the Kullback information (4.2.71), its minimum value for all θ such that $\theta^T\Sigma_{\theta\theta}^{-1}\theta \geq b^2$, under the least favorable ξ^* is equal to

$$2\,\mathbf{K}^* = \theta^T\,\Sigma_{\theta\theta}^{-1}\,\theta = b^2 \tag{4.2.86}$$

Now let us derive the GLR test (4.2.62) for hypotheses (4.2.85) under the least favorable values ξ_0^* and ξ_1^* :

$$\begin{aligned} \hat{\Lambda}(Y) &= \frac{\sup_{\theta^T\Sigma_{\theta\theta}^{-1}\theta \geq b^2} \; p_{\theta,\xi_1^*}(Y)}{\sup_{\theta^T\Sigma_{\theta\theta}^{-1}\theta \leq a^2} \; p_{\theta,\xi_0^*}(Y)} \\ &= \frac{\sup_{\theta^T\Sigma_{\theta\theta}^{-1}\theta \geq b^2} \; p_{\theta,\xi_1^*}(Y)}{p_{0,\xi_0^*}(Y)} \;\; \frac{p_{0,\xi_0^*}(Y)}{\sup_{\theta^T\Sigma_{\theta\theta}^{-1}\theta \leq a^2} \; p_{\theta,\xi_0^*}(Y)} \end{aligned} \tag{4.2.87}$$

The logarithm $\hat{S}_1(Y)$ of the first term on the right side of this equation can be rewritten as

$$\begin{aligned} \hat{S}_1(Y) &= \ln \sup_{\theta^T\Sigma_{\theta\theta}^{-1}\theta \geq b^2} p_{\theta,\xi_1^*}(Y) - \ln p_{0,\xi_0^*}(Y) \\ &= \sup_{\theta^T\Sigma_{\theta\theta}^{-1}\theta \geq b^2} \left[\ln p_{\theta,\xi_1^*}(Y) - \ln p_{0,\xi_0^*}(Y)\right] \end{aligned}$$

Now we can apply to the right side of the last equation the result we obtained before for the likelihood ratio for simple hypotheses with nuisance parameter :

$$\hat{S}_1(Y) = \sup_{\theta^T\Sigma_{\theta\theta}^{-1}\theta \geq b^2} \left[-\frac{1}{2}(Y_\theta - \theta)^T\Sigma_{\theta\theta}^{-1}(Y_\theta - \theta) + Y_\theta\Sigma_{\theta\theta}^{-1}Y_\theta\right]$$

Using a similar computation for the second term $\hat{S}_2(Y)$ on the right side of (4.2.87), we finally obtain $\hat{S}(Y)$ for hypotheses (4.2.85) under the least favorable values ξ_0^* and ξ_1^* of the nuisance parameter

$$\begin{aligned} \hat{S}(Y) &= \sup_{\theta^T\Sigma_{\theta\theta}^{-1}\theta \geq b^2} \left[-\frac{1}{2}(Y_\theta - \theta)^T\Sigma_{\theta\theta}^{-1}(Y_\theta - \theta)\right] \\ &- \sup_{\theta^T\Sigma_{\theta\theta}^{-1}\theta \leq a^2} \left[-\frac{1}{2}(Y_\theta - \theta)^T\Sigma_{\theta\theta}^{-1}(Y_\theta - \theta)\right] \end{aligned} \tag{4.2.88}$$

From this equation, it is obvious that we get the GLR test (4.2.65), which is equivalent to the minmax test (4.2.61) that we discussed in example 4.2.6. Hence, the test (4.2.84) can finally be rewritten as

$$\tilde{g}(Y) = \begin{cases} 1 & \text{when} \quad Y_\theta^T \Sigma_{\theta\theta}^{-1} Y_\theta \geq \lambda_\epsilon^2 \\ 0 & \text{when} \quad Y_\theta^T \Sigma_{\theta\theta}^{-1} Y_\theta < \lambda_\epsilon^2 \end{cases} \tag{4.2.89}$$

Note that the final test, which is independent of the unknown value ξ^*, also turns out to be independent of the unknown value ξ_0^*.

4.2.9 Asymptotic Local Approach for Composite Hypotheses

We have shown in the previous paragraphs that optimal tests exist only for special cases under essential restrictions. However, one case exists in which the design and analysis of optimal tests is basically simpler than in the general case. This case is encountered in the so-called local hypotheses approach. We discussed this approach in the simple hypotheses case in subsection 4.2.3. Now we consider the asymptotic local approach for composite hypotheses. This approach is based upon an asymptotic local expansion of the likelihood ratio, which was informally presented in subsection 4.2.3. Let us briefly discuss the main ideas of this expansion.

4.2.9.1 Local Asymptotic Expansion of the Likelihood Ratio

The LR has been shown to be a central tool in hypotheses testing problems. Thus, the investigation of its properties under local asymptotic conditions, namely close hypotheses, is of interest. On the other hand, a theory of contiguity or closeness of probability measures was developed in [Le Cam, 1960, Roussas, 1972, Davies, 1973, Ibragimov and Khasminskii, 1981]. Let us introduce its main features.

We consider a parametric family of distributions $\mathcal{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta}$, $\Theta \subset R^\ell$, satisfying some regularity assumptions [Roussas, 1972, Davies, 1973, Ibragimov and Khasminskii, 1981] and a sample of size N. Let $(\nu_N \Upsilon)_N$, where $\|\Upsilon\| = 1$, be a convergent sequence of points in the space R^ℓ such that $\nu_N \to \nu \in R$. Let $\theta_N = \theta + \frac{\nu_N}{\sqrt{N}}\Upsilon$. Therefore, the distance between the hypotheses

$$\mathbf{H}_0 = \{\mathcal{L}(Y) = \mathbf{P}_\theta\} \text{ and } \mathbf{H}_1 = \left\{\mathcal{L}(Y) = \mathbf{P}_{\theta+\frac{\nu_N}{\sqrt{N}}\Upsilon}\right\} \tag{4.2.90}$$

depends upon N in such a way that the two probability measures get closer to each other when N grows to infinity. The logarithm of the LR for the sample $\mathcal{Y}_1^N$ can be written as

$$S(\theta, \theta_N) = \ln \frac{p_{\theta_N}(\mathcal{Y}_1^N)}{p_\theta(\mathcal{Y}_1^N)} \tag{4.2.91}$$

Definition 4.2.14 (LAN family of distributions). *The parametric family of distributions $\mathcal{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta}$ is called* locally asymptotic normal *(LAN) if the logarithm of the LR for hypotheses $\mathbf{H}_0$ and $\mathbf{H}_1$ can be written as*

$$S(\theta, \theta_N) = \nu \Upsilon^T \Delta_N(\theta) - \frac{\nu^2}{2} \Upsilon^T \mathbf{I}_N(\theta) \Upsilon + \alpha_N(\mathcal{Y}_1^N, \theta, \nu\Upsilon) \tag{4.2.92}$$

where, in a similar manner as in the subsection 4.2.3,

$$\Delta_N(\theta) = \frac{1}{\sqrt{N}} \frac{\partial \ln p_\theta(\mathcal{Y}_1^N)}{\partial\theta} = \frac{1}{\sqrt{N}} \mathcal{Z}_N, \tag{4.2.93}$$

$\mathbf{I}_N(\theta)$ is the Fisher information matrix for the sample $\mathcal{Y}_1^N$, and where the following asymptotic normality holds :

$$\mathcal{L}[\Delta_N(\theta)] \rightsquigarrow \mathcal{N}[0, \mathbf{I}(\theta)] \tag{4.2.94}$$

In expansion (4.2.92), the random variable α_N is such that $\alpha_N \to 0$ almost surely under the probability measure $\mathbf{P}_\theta$.

Let $\mathcal{P}$ be a LAN family of distributions. Let $\Upsilon_\nu = \nu\Upsilon$, and denote $\hat{\Upsilon}_\nu = \sqrt{N}(\hat{\theta} - \theta)$ the value of the parameter Υ_ν for which $S(\theta, \theta_N)$ is maximum. This value is

$$\hat{\Upsilon}_\nu = \mathbf{I}_N^{-1}(\theta)\ \Delta_N(\theta) + \alpha_N(\mathcal{Y}_1^N, \theta, \Upsilon_\nu) \tag{4.2.95}$$

and the logarithm $S\left(\theta, \theta + \frac{1}{\sqrt{N}}\hat{\Upsilon}_\nu\right)$ of the LR can be rewritten as

$$2S\left(\theta, \theta + \frac{1}{\sqrt{N}}\hat{\Upsilon}_\nu\right) = \Delta_N^T(\theta)\ \mathbf{I}_N^{-1}(\theta)\ \Delta_N(\theta) + \alpha_N(\mathcal{Y}_1^N, \theta, \Upsilon_\nu) \tag{4.2.96}$$

As proven in [Roussas, 1972, Davies, 1973, Ibragimov and Khasminskii, 1981], LAN properties exist for some important special cases. More precisely, the asymptotic local expansion (4.2.92) can be derived if

- $(Y_n)_n$ is a sequence of *independent identically distributed* (i.i.d.) random variables;
- $(Y_n)_n$ is a stationary Markov process of order p;
- $(Y_n)_n$ is a stationary Gaussian random process, in particular, $(Y_n)_n$ is an *autoregressive moving average* (ARMA) process.

Furthermore, we have the following asymptotic normality of $S(\theta, \theta_N)$, $\Delta_N(\theta)$, and $\hat{\Upsilon}_\nu$:

$$\begin{aligned}
\mathcal{L}(S(\theta,\theta_N)) \quad &\rightsquigarrow \quad \begin{cases} \mathcal{N}\left[-\frac{\nu^2}{2}\ \Upsilon^T\ \mathbf{I}(\theta)\ \Upsilon, \nu^2\ \Upsilon^T\ \mathbf{I}(\theta)\ \Upsilon\right] & \text{when} \quad \mathcal{L}(Y) = \mathbf{P}_\theta \\ \mathcal{N}\left[+\frac{\nu^2}{2}\ \Upsilon^T\ \mathbf{I}(\theta)\ \Upsilon, \nu^2\ \Upsilon^T\ \mathbf{I}(\theta)\ \Upsilon\right] & \text{when} \quad \mathcal{L}(Y) = \mathbf{P}_{\theta+\frac{\nu}{\sqrt{N}}\Upsilon} \end{cases} \\
\mathcal{L}(\Delta_N(\theta)) \quad &\rightsquigarrow \quad \mathcal{N}\left[\nu\ \mathbf{I}(\theta)\ \Upsilon, \mathbf{I}(\theta)\right] \quad \text{when} \quad \mathcal{L}(Y) = \mathbf{P}_{\theta+\frac{\nu}{\sqrt{N}}\Upsilon} \\
\mathcal{L}(\hat{\Upsilon}_\nu) \quad &\rightsquigarrow \quad \mathcal{N}[0, \mathbf{I}^{-1}(\theta)] \quad \text{when} \quad \mathcal{L}(Y) = \mathbf{P}_\theta
\end{aligned} \tag{4.2.97}$$

We also have the following convergence :

$$\mathcal{L}\left[2S\left(\theta, \theta + \frac{1}{\sqrt{N}}\hat{\Upsilon}_\nu\right)\right] \rightsquigarrow \chi^2(\ell) \ \text{ when } \ \mathcal{L}(Y) = \mathbf{P}_\theta \tag{4.2.98}$$

Note here that the random variable $\alpha_N(\mathcal{Y}_1^N, \theta, \Upsilon_\nu)$ converges to zero almost surely under the probability measure $\mathbf{P}_{\theta+\frac{\nu}{\sqrt{N}}\Upsilon}$.

The important corollary of the LAN properties for a parametric family $\mathcal{P}$ satisfying the regularity conditions is that the LR $e^{S(\theta,\theta_N)}$ behaves approximately as if the family were exponential. Thus, the vector of the *efficient score* $\Delta_N(\theta)$ is an *asymptotic sufficient statistic*. Moreover, from the above asymptotic normality it is possible to transform the asymptotic local hypotheses testing problem $\mathbf{H}_0 = \{\mathcal{L}(Y) = \mathbf{P}_\theta\}$ against $\mathbf{H}_1 = \left\{\mathcal{L}(Y) = \mathbf{P}_{\theta+\frac{\nu}{\sqrt{N}}\Upsilon}\right\}$ into a much simpler hypotheses testing problem for the mean of a Gaussian law. We continue to investigate this result in the next paragraphs.

Finally, let us add one comment about the reason for which the speed of convergence of hypotheses $\mathbf{H}_0$ and $\mathbf{H}_1$ was chosen of the order of $\frac{1}{\sqrt{N}}$. For a hypotheses testing problem, the contiguity of the probability measures must be compensated by the growth of the sample size N. If the varying parameter is $\theta_N = \theta_0 + \frac{\nu}{\sqrt{N}}\Upsilon$, then the quantity of information for distinguishing between hypotheses $\mathbf{H}_0$ and $\mathbf{H}_1$ remains constant when N goes to infinity. For this reason, the probabilities of errors of the first and second types tend to fixed values, as was discussed in subsection 4.2.3 for the case of a scalar parameter.

4.2.9.2 Asymptotic Optimal Tests for Composite Hypotheses

Here we follow [Borovkov, 1984, Roussas, 1972]. Let $\mathbf{H}_0 = \{\theta \ : \theta \in \Theta_0\}$ and $\mathbf{H}_1 = \{\theta \ : \theta \in \Theta_1\}$ be given composite hypotheses. Assume that each set Θ_i can be written as

$$\Theta_i = \theta^* + \frac{\Gamma_i}{\sqrt{N}}, \ \ i = 0, 1 \tag{4.2.99}$$

where the vector $\theta^* \in \Theta$ and the sets Γ_i are independent of N. According to the previous discussion about the local case, it is possible to transform hypotheses $\mathbf{H}_0$ and $\mathbf{H}_1$ in the following manner :

$$\mathbf{H}_0 = \{\Upsilon_\nu \ : \Upsilon_\nu \in \Gamma_0\} \ \text{ and } \ \mathbf{H}_1 = \{\Upsilon_\nu \ : \Upsilon_\nu \in \Gamma_1\} \tag{4.2.100}$$

The main idea of the asymptotic optimal tests Let us consider two hypotheses testing problems.

- **First problem.** Assume that $\mathcal{P}$ is a LAN family, and that we want to test the local hypotheses (4.2.99) by using $\mathcal{Y}_1^N$ when $N \to \infty$.
- **Second problem.** Assume that the family $\mathcal{P}$ is such that $\mathcal{L}(Y) = \mathcal{N}(\Upsilon_\nu, \Sigma)$, where $\Sigma = \mathbf{I}^{-1}(\theta^*)$, and that we want to test between hypotheses $\mathbf{H}_0 = \{\Upsilon_\nu \in \Gamma_0\}$ and $\mathbf{H}_1 = \{\Upsilon_\nu \in \Gamma_1\}$ about the mean of a Gaussian law by using one sample point Y_1.

Now, assume that the second problem can be solved by an optimal (UMP, Bayes, minmax,...) test $g(Y_1)$. Let us denote the maximum likelihood estimate as $\hat{\theta}$, and let $\hat{\Upsilon}_\nu = \sqrt{N}(\hat{\theta} - \theta^*)$. Then the test $g(\hat{\Upsilon}_\nu)$ for the first problem will have *asymptotically* the same properties as the optimal test $g(Y_1)$ for the second problem.

This idea can be explained as follows. Let $\theta = \theta^* + \frac{1}{\sqrt{N}}\Upsilon_\nu$. From formulae (4.2.95) and (4.2.98) we have

$$\mathcal{L}[\sqrt{N}(\hat{\theta} - \theta^*)] \rightsquigarrow \mathcal{N}[\Upsilon_\nu, \mathbf{I}^{-1}(\theta^*)] \tag{4.2.101}$$

This normal distribution with mean vector Υ_ν and covariance matrix $\mathbf{I}^{-1}(\theta^*)$ is precisely the distribution in the second problem.

Asymptotic optimal properties of the GLR tests Let us consider some examples of asymptotic optimal tests.

Definition 4.2.15 *The test $g(y)$ which satisfies $\lim_{n\to\infty} \alpha_0(g) = \epsilon$ is called test with asymptotic level $1 - \epsilon$ or asymptotic size ϵ.*

Example 4.2.7 (Scalar parameter). *Let $\mathcal{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta}$, $\theta \subset R$ be a LAN family with scalar parameter θ. Consider the problem of testing between the following two hypotheses :*

$$\mathbf{H}_0 = \{\theta \ : \theta \in \Theta_0\} \ \textit{ against } \ \mathbf{H}_1 = \{\theta \ : \theta \in \Theta_1\} \tag{4.2.102}$$

where

$$\Theta_0 = \left\{\theta \ : \theta \leq \theta_0 = \theta^* + \frac{\nu_0}{\sqrt{N}}\right\} \ \textit{and} \ \ \Theta_1 = \left\{\theta : \theta > \theta_1 = \theta^* + \frac{\nu_1}{\sqrt{N}}\right\} \tag{4.2.103}$$

with $\nu_0 \leq \nu_1$. This problem was discussed before for monotone likelihood ratio *density (see test g^* (4.2.32 - 4.2.33)). Let us first introduce a class K_ϵ^a of tests with* asymptotic *level $1-\epsilon$ in the following manner :*

$$K_\epsilon^a = \{g : \limsup_{N\to\infty} \sup_{\theta\in\Theta_0} \mathbf{E}_\theta[g(\mathcal{Y}_1^N)] \leq \epsilon\} \tag{4.2.104}$$

and then give one theorem [Borovkov, 1984] that defines the asymptotic optimal test for hypotheses $\mathbf{H}_0$ and $\mathbf{H}_1$.

Theorem 4.2.3 *The GLR test $\hat{g}_1(\mathcal{Y}_1^N)$ (4.2.62) with critical region*

$$\frac{\sup_{\theta\in\Theta_1} p_\theta(\mathcal{Y}_1^N)}{\sup_{\theta\in\Theta_0} p_\theta(\mathcal{Y}_1^N)} \geq \lambda_{1_\epsilon} \tag{4.2.105}$$

is asymptotically equivalent *to the test $\hat{g}_2(\mathcal{Y}_1^N)$ with critical region :*

$$\hat{\Upsilon}_\nu = (\hat{\theta} - \theta^*)\sqrt{N} \geq \lambda_{2_\epsilon} \tag{4.2.106}$$

or to the test $\hat{g}_3(\mathcal{Y}_1^N)$ with critical region :

$$\Delta_N(\theta^*) = \frac{1}{\sqrt{N}} \frac{\partial \ln p_\theta(\mathcal{Y}_1^N)}{\partial\theta} \geq \lambda_{3_\epsilon} \tag{4.2.107}$$

Tests (4.2.105)-(4.2.107) are asymptotic UMP *in the class K_ϵ^a.*

Example 4.2.8 (Vector parameter). *Let us consider again the case of an ℓ-dimensional parameter. Assume that $\mathcal{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta}$, $\Theta \subset R^\ell$ is a LAN family, and consider the composite hypotheses $\mathbf{H}_0 = \{\theta : \theta \in \Theta_0\}$ and $\mathbf{H}_1 = \{\theta : \theta \in \Theta_1\}$, where*

$$\begin{aligned} \Theta_0 &= \left\{\theta : (\theta - \theta^*)^T \ \mathbf{I}(\theta^*) \ (\theta - \theta^*) \leq \frac{a^2}{N}\right\} \\ \Theta_1 &= \left\{\theta : (\theta - \theta^*)^T \ \mathbf{I}(\theta^*) \ (\theta - \theta^*) \geq \frac{b^2}{N}\right\} \end{aligned} \tag{4.2.108}$$

Theorem 4.2.4 *[Borovkov, 1984] The GLR test $\hat{g}(\mathcal{Y})$ (4.2.105) for testing between the hypotheses (4.2.108) is asymptotically equivalent to the test with critical region*

$$\frac{p_{\hat{\theta}}(\mathcal{Y}_1^N)}{p_{\theta^*}(\mathcal{Y}_1^N)} \geq \lambda_{1_\epsilon} \tag{4.2.109}$$

or to the test with critical region

$$N(\hat{\theta} - \theta^*)^T \ \mathbf{I}_N(\theta^*) \ (\hat{\theta} - \theta^*) \geq \lambda_{2_\epsilon} \tag{4.2.110}$$

or to the test with critical region

$$\Delta_N^T(\theta^*) \ \mathbf{I}_N^{-1}(\theta^*) \ \Delta_N(\theta^*) \geq \lambda_{3_\epsilon} \tag{4.2.111}$$

4.3 Sequential Analysis

In section 4.2, we discussed statistical inference and hypotheses testing problems about a parametric family of distributions $\mathcal{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta}$, using raw data $\mathcal{Y}_1^N$ with *fixed sample size*. We have shown that optimal tests exist in this situation. More precisely, we have shown that it is possible to minimize the error probabilities for a given sample size. Now the problem of interest is the following : For given error probabilities, try to minimize the sample size or equivalently to make the decision with as few observations as possible. This problem arises in practice when the cost of each observation has to be taken into account.

Sequential analysis is the theory of solving hypotheses testing problems when the sample size is not fixed *a priori* but depends upon the data that have been already observed. The theory of sequential analysis was formulated by A. Wald in his famous book [Wald, 1947]. Let us describe in this section the main ideas and methods of sequential analysis. We follow here [Wald, 1947, Ghosh, 1970, Borovkov, 1984, Siegmund, 1985b].

4.3.1 Notation and Main Criteria

Let us first define the main types of sequential tests and criteria.

Definition 4.3.1 (Sequential test). *A* sequential statistical test *for testing between hypotheses* $\mathbf{H}_0$ *and* $\mathbf{H}_1$ *is defined to be a pair* (g, T)*, where* T *is a stopping time and* $g(\mathcal{Y}_1^T)$ *is a decision function*.

The performance index of a sequential test is usually defined with the aid of the following criterion.

Definition 4.3.2 (ASN). *The* average sample number (ASN) *is the mean number of sample points* $\mathbf{E}_\theta(T)$ *necessary for testing the hypotheses with acceptable probabilities of errors of first and second types*.

The fact that a sequential test does not indefinitely continue is ensured by the following termination property. This is important in practice because we should take the decision in finite time.

Definition 4.3.3 (Closed test). *We say that a sequential test* (g, T) *is* closed *if*

$$\forall\theta \in \Theta,\ \mathbf{P}_\theta(T < \infty) = 1 \tag{4.3.1}$$

A test that is not closed is said to be open.

Definition 4.3.4 (Valid test). *We say that a class* K_{α_0,α_1} *of tests* (g, T) *for testing between hypotheses* $\mathbf{H}_0 = \{\theta\ :\ \theta \in \Theta_0\}$ *and* $\mathbf{H}_1 = \{\theta\ :\ \theta \in \Theta_1\}$ *is* valid *if each* $(g, T) \in K_{\alpha_0,\alpha_1}$ *satisfies the following :*

- (g, T) *is closed*
- *when* $\theta \in \Theta_0$, $0 \le \alpha_0(g) \le \alpha_0 \le 1$
- *when* $\theta \in \Theta_1$, $0 \le \alpha_1(g) \le \alpha_1 \le 1$

Now let (g, T) and $(\tilde{g}, \tilde{T})$ be two valid tests.

Definition 4.3.5 (UME test). *We say that the test* $(g, T) \in K_{\alpha_0,\alpha_1}$ *is* more efficient *than* $(\tilde{g}, \tilde{T}) \in K_{\alpha_0,\alpha_1}$ *at* θ *if*

$$\mathbf{E}_\theta(T) < \mathbf{E}_\theta(\tilde{T}) \tag{4.3.2}$$

for some $\theta \in \Theta$. *We say that* (g^*, T^*) *is* uniformly most efficient (UME) *if*

$$\forall\theta \in \Theta,\quad \inf_{(g,T)\in K_{\alpha_0,\alpha_1}} \mathbf{E}_\theta(T) = \mathbf{E}_\theta(T^*) \tag{4.3.3}$$

4.3.2 Sequential Testing Between Two Simple Hypotheses

As in section 4.2, let us start from the important case of two simple hypotheses. In this case, the amount of available theoretical results is the highest - proof of optimality of sequential tests does exist - and the practical implementation of sequential algorithms is the simplest. However, it should be clear that simple hypotheses scarcely arise in practice, and that these tests are *not* optimal when the assumption of simple hypotheses is not valid. This provides us with motivation for further investigation of more complex cases after.

4.3.2.1 Sequential Probability Ratio Test

Definition 4.3.6 (SPRT). *We say that the test (g, T) is a* sequential probability ratio test (SPRT) *for testing between simple hypotheses $\mathbf{H}_0 = \{\theta \ : \theta = \theta_0\}$ and $\mathbf{H}_1 = \{\theta \ : \theta = \theta_1\}$ if we sequentially observe data $(Y_n)_{n\geq 1}$ and if, at time n, we make one of the following decisions :*

- *accept $\mathbf{H}_0$ when $S_n \leq -a$;*
- *accept $\mathbf{H}_1$ when $S_n \geq h$;*
- *continue to observe and to test when $-a < S_n < h$*

where

$$S_n = \ln \frac{p_{\theta_1}(\mathcal{Y}_1^n)}{p_{\theta_0}(\mathcal{Y}_1^n)} \tag{4.3.4}$$

and $-a, h$ are boundaries (thresholds) such that $-\infty < -a < h < \infty$. Sometimes two types of limit cases may be of interest : the first is when one *of the thresholds is infinite; the second is the asymptotic case where both thresholds go to infinity.*

This definition can be rewritten as follows :

$$g(\mathcal{Y}_1^T) = \begin{cases} 1 & \text{when} \quad S_T \geq h \\ 0 & \text{when} \quad S_T \leq -a \end{cases} \tag{4.3.5}$$

where T is the exit time

$$T = T_{-a,h} = \min\{n \geq 1 : (S_n \geq h) \cup (S_n \leq -a)\} \tag{4.3.6}$$

Example 4.3.1 (Mean in a Gaussian sequence - contd.). *We consider again a scalar Gaussian sequence $(y_k)_{k\geq 1}$ with distribution $\mathcal{L}(y_k) = \mathcal{N}(\theta, 1)$, and the two following hypotheses :*

$$\mathbf{H}_0 = \{\theta \ : \theta = 0\} \ \text{ and } \ \mathbf{H}_1 = \{\theta \ : \theta = 2\} \tag{4.3.7}$$

By definition, the log-likelihood ratio of the SPRT can be written as

$$S_n = \sum_{i=1}^{n} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} = \sum_{i=1}^{n} \ln \frac{\varphi\left(y_i - 2\right)}{\varphi\left(y_i\right)} \tag{4.3.8}$$

or

$$S_n = \sum_{i=1}^{n} 2(y_i - 1) \tag{4.3.9}$$

The graphical representation of SPRT (4.3.5) is shown in figure 4.4, where the typical behavior of the cumulative sum S_n and the acceptance zones of hypotheses $\mathbf{H}_0$ and $\mathbf{H}_1$ are depicted.

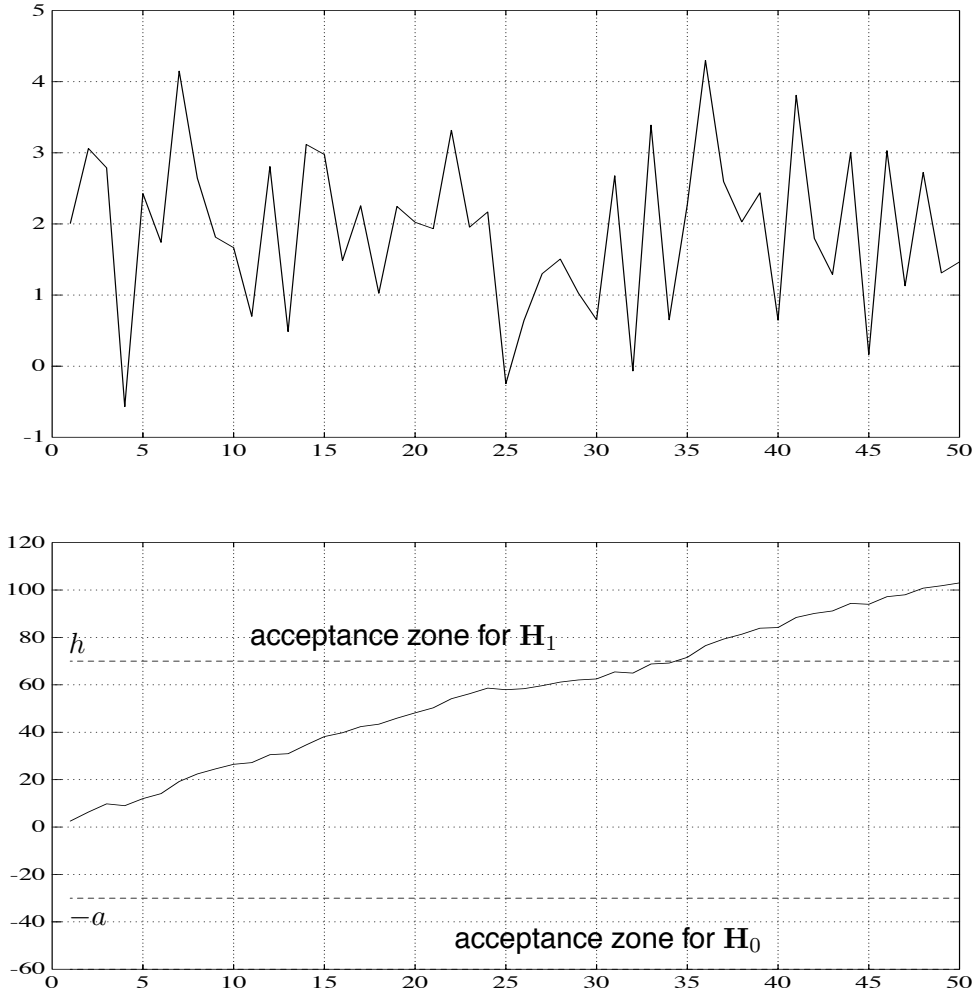

Figure 4.4 Typical behavior of a SPRT test : signal $(y_k)_k$ with distribution $\mathcal{N}(2,1)$ (first row) and decision function S_k (second row) for testing between $\mathbf{H}_0 \ : \theta = 0$ and $\mathbf{H}_1 \ : \theta = 2$.

4.3.2.2 Optimal and Termination Properties of the Sequential Probability Ratio Test

It has been proven [Wald, 1947, Ghosh, 1970, Siegmund, 1985b] that there exists a UME sequential test in the class of all valid SPRT, and moreover that the SPRT is the UME test among all sequential and nonsequential (fixed sample size) tests.

Theorem 4.3.1 *Let $(Y_i)_{1\leq i\leq n}$ be an i.i.d. sequence. Assume that (g,T) is a SPRT with boundaries $-a < 0 < h$ for testing between hypotheses $\mathbf{H}_0 = \{\theta \ : \theta = \theta_0\}$ and $\mathbf{H}_1 = \{\theta \ : \theta = \theta_1\}$, with error probabilities $\alpha_0(g)$ and $\alpha_1(g)$. Let the ASN be $\mathbf{E}_{\theta_0}(T)$ and $\mathbf{E}_{\theta_1}(T)$. Denote by K_{α_0,α_1} the class of all (sequential and nonsequential) tests $\tilde{g}$ such that*

$$\begin{gathered}\alpha_0(\tilde{g}) \leq \alpha_0(g), \quad \alpha_1(\tilde{g}) \leq \alpha_1(g) \\ \mathbf{E}_{\theta_0}(\tilde{T}) < \infty, \quad \mathbf{E}_{\theta_1}(\tilde{T}) < \infty\end{gathered} \tag{4.3.10}$$

Then, for every $(\tilde{g},\tilde{T}) \in K_{\alpha_0,\alpha_1}$, we have

$$\mathbf{E}_{\theta_0}(\tilde{T}) \geq \mathbf{E}_{\theta_0}(T), \qquad \mathbf{E}_{\theta_1}(\tilde{T}) \geq \mathbf{E}_{\theta_1}(T) \tag{4.3.11}$$

Now we discuss another important property for a sequential test, namely the termination property (when the test is closed), and we give sufficient conditions under which a SPRT (g,T) is closed, namely $\mathbf{P}_\theta(T < \infty) = 1$ for all $\theta \in \Theta$. There are two possible situations :

- *The random variables $(s_n)_{n\geq 1}$ are i.i.d.* Suppose that $\mathbf{P}_\theta(s_n = 0) < 1$ for each given $\theta \in \Theta$. Then, the SPRT is closed. It is obvious that this assumption about s_n can be replaced by $\mathrm{var}_\theta(s_n) > 0$ for any given $\theta \in \Theta$.
- *The random variables $(s_n)_{n\geq 1}$ are not i.i.d.* Under this assumption the probability $\mathbf{P}_\theta(T > n)$ can be rewritten as

$$\begin{aligned}\mathbf{P}_\theta(T > n) &= \mathbf{P}_\theta(-a < S_1 < h, \ldots, -a < S_n < h) \\ &\leq \mathbf{P}_\theta\left[-a < S_n = \ln\frac{p_{\theta_1}(\mathcal{Y}_1^n)}{p_{\theta_0}(\mathcal{Y}_1^n)} < h\right]\end{aligned} \tag{4.3.12}$$

Now the following result holds. A sufficient condition under which a SPRT *with not i.i.d. cumulative sum increments* s_i is closed is

$$\lim_{n\to\infty} \mathbf{P}_\theta(-a < S_n < h) = 0 \ \text{ for every } \ \theta \in \Theta \tag{4.3.13}$$

An alternative sufficient condition is the following. Assume that there exists a number $N > 0$ such that, for each $n > N$, this condition can be rewritten with variable boundaries :

$$\lim_{n\to\infty} \mathbf{P}_\theta(-\tilde{a}_n < \tilde{S}_n < \tilde{h}_n) = 0 \ \text{ for every } \ \theta \in \Theta \tag{4.3.14}$$

where $\tilde{S}_n = \tilde{S}_n(\mathcal{Y}_1^n)$, the boundaries $-\tilde{a}_n < \tilde{h}_n$ are nonrandom values, and $-\tilde{a}_n, \tilde{h}_n, \tilde{S}_n$ can be functions of θ. Moreover, assume also one of the two following conditions :

- The random sequence $(\tilde{S}_n)_n$ converges in probability to f (which can be a function of θ) and $\lim_{n\to\infty} -\tilde{a}_n = \lim_{n\to\infty} \tilde{h}_n \neq f$;

- $\mathcal{L}(\tilde{S}_n) \rightsquigarrow \mathcal{L}(\tilde{S})$, where $\tilde{S}$ is a random variable that has a continuous distribution function, and $\lim_{n\to\infty} -\tilde{a}_n = \lim_{n\to\infty} \tilde{h}_n$.

Now we add some comments for these results and one example. Let us emphasize that the optimality and other properties of the SPRT do hold even if the *data* $(Y_n)_n$ *are not i.i.d.*, provided that the *increments* $(s_n)_n$ *of the cumulative sum* S_n *are i.i.d.* for every $\theta \in \Theta$. This is an important fact for subsequent chapters because, as shown in section 3.1, the increments of the log-likelihood of the observations $\mathcal{Y}_1^n$ are nothing but the likelihood of the innovations. It turns out that in some special cases (e.g., for additive changes) the innovation (residual) process is an independent random sequence under both hypotheses $\mathbf{H}_0$ and $\mathbf{H}_1$. Thus, in these cases, the increments $(s_n)_n$ are i.i.d., and the termination and optimal properties of the SPRT hold.

The second termination property (for non-i.i.d. increments) is very useful and important for the generalized SPRT for composite hypotheses testing, which we define later, because in this case the decision statistic $(\tilde{S}_n)_n$ is not a cumulative sum with independent increments even if $(Y_n)_n$ is i.i.d.

Example 4.3.2 (Change in the mean of an AR process). *We consider again the example of change in the mean value θ of a Gaussian sequence $(y_n)_n$, but now we no longer assume the sequence to be independent; rather we assume it to be a stable* autoregressive sequence *:*

$$y_k = \sum_{i=1}^{p} a_i y_{k-i} + v_k + \left(1 - \sum_{i=1}^{p} a_i\right)\ \theta \tag{4.3.15}$$

where $a_1, \ldots, a_p$ are autoregressive coefficients and $(v_k)_k$ is a white noise sequence with zero mean and variance σ^2. From the transformation lemma (3.1.44) (see also example 3.1.8) we deduce that the likelihood function of such a sequence can be written as

$$\ln p_\theta(\mathcal{Y}_1^n) = \sum_{k=1}^{n} p_\theta(\varepsilon_k) \tag{4.3.16}$$

where

$$\begin{aligned} \varepsilon_k &= y_k - \mathbf{E}_\theta(y_k|\mathcal{Y}_{k-p}^{k-1}) \\ &= y_k - \sum_{i=1}^{p} a_i y_{k-i} - \left(1 - \sum_{i=1}^{p} a_i\right)\ \theta \end{aligned} \tag{4.3.17}$$

is the innovation.

The likelihood ratio corresponding to the two hypotheses $\mathbf{H}_0 = \{\theta\ : \theta = \theta_0\}$ against $\mathbf{H}_1 = \{\theta\ : \theta = \theta_1\}$ is thus

$$\begin{aligned} S_n &= \sum_{k=1}^{n} \ln \frac{p_{\theta_1}(\varepsilon_k)}{p_{\theta_0}(\varepsilon_k)} \\ &= \sum_{k=1}^{n} \ln \frac{\varphi\left\{\frac{1}{\sigma}[y_k - \sum_{i=1}^{p} a_i y_{k-i} - (1 - \sum_{i=1}^{p} a_i)\ \theta_1]\right\}}{\varphi\left\{\frac{1}{\sigma}[y_k - \sum_{i=1}^{p} a_i y_{k-i} - (1 - \sum_{i=1}^{p} a_i)\ \theta_0]\right\}} \end{aligned} \tag{4.3.18}$$

Therefore,

$$S_n = \sum_{k=1}^{n} \frac{\tilde{\theta}_1 - \tilde{\theta}_0}{\sigma^2} \left(y_k - \sum_{i=1}^{p} a_i y_{k-i} - \frac{\tilde{\theta}_0 + \tilde{\theta}_1}{2}\right) \tag{4.3.19}$$

where $\tilde{\theta}_j = (1 - \sum_{i=1}^{p} a_i)\ \theta_j$ for $j = 0, 1$. Thus, the increments of the cumulative sum are i.i.d. for every value of $\theta \in \mathbf{R}$.

4.3.2.3 Operating Characteristic and Wald's Identity

We now give some general and important results concerning the properties of the SPRT.

Operating Characteristic

Definition 4.3.7 (OC). *The probability $Q(\theta)$ of accepting hypothesis $\mathbf{H}_0$, treated as a function of $\theta \in \Theta$, is called the* operating characteristic (OC).

It is obvious that, for a *closed* sequential test, we can write $\beta(\theta) = 1 - Q(\theta)$ for every $\theta \in \Theta$, but in general (for the class of open tests), we have $\beta(\theta) + Q(\theta) \leq 1$.

Now, the thresholds $-a$ and h and the error probabilities $\alpha_0(g)$ and $\alpha_1(g)$ of the SPRT (4.3.5) corresponding to hypotheses $\mathbf{H}_0 = \{\theta = \theta_0\}$ and $\mathbf{H}_1 = \{\theta = \theta_1\}$ satisfy the two following *Wald's inequalities* :

$$\begin{array}{rcl} \ln \frac{\alpha_1(g)}{1-\alpha_0(g)} & \leq & \min(0, -a) \\ \ln \frac{1-\alpha_1(g)}{\alpha_0(g)} & \geq & \max(0, h) \end{array} \tag{4.3.20}$$

where $-\infty < -a < h < \infty$. The equalities hold in (4.3.20) if and only if $-a < 0 < h$, the test is closed, and

$$\mathbf{P}_{\theta_0}(S_T = -a \,|\mathbf{H}_0 \text{ is accepted}) = \mathbf{P}_{\theta_1}(S_T = h \,|\mathbf{H}_0 \text{ is rejected}) = 1 \tag{4.3.21}$$

Note here that the result (4.3.20) is *very general*. It can be applied to *closed or open* SPRT, to an *independent* or a *dependent* sequence $(Y_n)_n$.

Example 4.3.3 (Open-ended test). *Let $(Y_i)_{1 \leq i \leq n}$ be a sequence of random variables with joint density of distribution p_θ. The likelihood ratio LR is defined by*

$$\Lambda_n = e^{S_n} = \frac{p_{\theta_1}(\mathcal{Y}_1^n)}{p_{\theta_0}(\mathcal{Y}_1^n)} \tag{4.3.22}$$

Define a stopping time T_h by

$$T_h = \begin{cases} \min\{n : \Lambda_n \geq e^h\} \\ \infty \ \textit{if no such } n \textit{ exists} \end{cases} \tag{4.3.23}$$

This test is called open-ended test because $\mathbf{P}_{\theta_0}(T_h < \infty) < 1$. It is obvious that T_h is the stopping time corresponding to the one-sided SPRT with lower threshold $-a = -\infty$.

Let us prove the Wald's inequality :

$$\mathbf{P}_{\theta_0}(\Lambda_n \geq e^h \textit{ for some } n \geq 1) = \mathbf{P}_{\theta_0}(T_h < \infty) \leq e^{-h} \tag{4.3.24}$$

which is a special case of the second inequality in (4.3.20) for $\alpha_0(g)$ when $\alpha_1(g) = 0$. This result is important to the discussion about the properties of CUSUM type algorithms in chapters 5 and 7.

Define the critical region at time n for this one-sided SPRT :

$$\Omega_1^n = \{\mathcal{Y}_1^n : \Lambda_k < e^h \textit{ for } k = 1, 2, \ldots, n-1 \textit{ and } \Lambda_n \geq e^h\} \tag{4.3.25}$$

It is obvious that these regions Ω_1^n do not intersect for different n. Therefore, the critical region corresponding to the stopping time T_h is

$$\Omega_1 = \cup_{n=1}^{\infty} \Omega_1^n \tag{4.3.26}$$

and we have

$$\mathbf{P}_{\theta_0}(T_h < \infty) = \sum_{n=1}^{\infty} \mathbf{P}_{\theta_0}(\Omega_1^n) = \sum_{n=1}^{\infty} \int_{\Omega_1^n} p_{\theta_0}(\mathcal{Y}_1^n)\, d\mathcal{Y}_1^n \tag{4.3.27}$$

If Λ_n reaches the threshold e^h at $T_h = n$, then by definition

$$p_{\theta_1}(\mathcal{Y}_1^n) \geq e^h\, p_{\theta_0}(\mathcal{Y}_1^n) \;\; \textit{or} \;\; p_{\theta_0}(\mathcal{Y}_1^n) \leq e^{-h}\, p_{\theta_1}(\mathcal{Y}_1^n) \tag{4.3.28}$$

Therefore,

$$\sum_{n=1}^{\infty} \int_{\Omega_1^n} p_{\theta_0}(\mathcal{Y}_1^n)\, d\mathcal{Y}_1^n \leq e^{-h} \sum_{n=1}^{\infty} \int_{\Omega_1^n} p_{\theta_1}(\mathcal{Y}_1^n)\, d\mathcal{Y}_1^n = e^{-h} \mathbf{P}_{\theta_1}(T_h < \infty) \leq e^{-h} \tag{4.3.29}$$

which ends the proof of Wald's inequality (4.3.24) [Wald, 1947, Borovkov, 1984]. This inequality holds for any nonnegative supermartingale sequence $\Lambda_1, \ldots, \Lambda_n$:

$$\mathbf{P}(T_h < \infty) \leq \frac{\mathbf{E}(\Lambda_1)}{e^h} \tag{4.3.30}$$

when $e^h > \mathbf{E}(\Lambda_1)$ (see [Robbins, 1970, Pollak and Siegmund, 1975]).

Wald's Approximation for the Error Probabilities Let us briefly discuss the case where the equalities in (4.3.20) hold approximately. Assume that there exists a small value $\epsilon > 0$ such that

$$\begin{aligned} \mathbf{P}_{\theta_0}(-a - \epsilon \leq S_n \leq -a \,|\mathbf{H}_0 \;\text{ is accepted}) &\approx 1 \\ \mathbf{P}_{\theta_1}(h \leq S_n \leq h + \epsilon \,|\mathbf{H}_0 \;\text{ is rejected}) &\approx 1 \end{aligned} \tag{4.3.31}$$

In this case, we can assume that

$$\begin{aligned} \ln \tfrac{\alpha_1(g)}{1-\alpha_0(g)} &\approx -a \\ \ln \tfrac{1-\alpha_1(g)}{\alpha_0(g)} &\approx h \end{aligned} \tag{4.3.32}$$

In practice, the *Wald's approximations* (4.3.32) are widely used when the excess of S_n over the boundary h and the excess of S_n below the boundary $-a$ are small. We call *excess over the boundary* the random variables

$$\bar{R}_T = S_T - h \geq 0 \;\text{ and }\; \underline{R}_T = S_T + a \leq 0 \tag{4.3.33}$$

We say that the excess is small if, for small $\epsilon > 0$, $\mathbf{P}(0 \leq R_T \leq \epsilon) \approx 1$, where $R_T = \bar{R}_T$ or $-\underline{R}_T$. If the test is closed and both excesses $\bar{R}_T$ and $\underline{R}_T$ are equal to zero with probability 1, then approximations (4.3.32) become true equalities and the SPRT with the thresholds

$$-a = \ln \frac{\alpha_1}{1-\alpha_0} \;\text{ and }\; h = \ln \frac{1-\alpha_1}{\alpha_0} \tag{4.3.34}$$

achieves the strength (α_0, α_1).

But, in general, this is not the case and the following question arises. What is the matter with true errors $\alpha_0(g), \alpha_1(g)$ if we choose the thresholds for the SPRT as in (4.3.34), where α_0 and α_1 are *preassigned errors*? From the formula (4.3.20), it can be deduced that, in this case,

$$\begin{aligned} \alpha_0(g) + \alpha_1(g) &\leq \alpha_0 + \alpha_1 \\ \alpha_0(g) &\leq [1 - \alpha_1(g)] \frac{\alpha_0}{1-\alpha_1} \\ \alpha_1(g) &\leq [1 - \alpha_0(g)] \frac{\alpha_1}{1-\alpha_0} \end{aligned} \tag{4.3.35}$$

Let us assume that

- the SPRT is closed;
- the excesses $\bar{R}_T$ and $\underline{R}_T$ are small;
- both preassigned risks α_0 and α_1 are small.

Under these assumptions, we get $\alpha_0(g) \approx \alpha_0$ and $\alpha_1(g) \approx \alpha_1$ for the SPRT (4.3.5) with thresholds as in (4.3.34).

Wald's Identity We now add two important results for estimating the properties of the sequential tests, which are known as *Wald's identity*.

Theorem 4.3.2 (Wald's identity). *We distinguish the cases of finite first or second moment.*

- *Assume that, for any $\theta \in \Theta$, the increments $(s_i)_{1\leq i\leq n}$ of the cumulative sum S_n are* independent *with the same mean $\mathbf{E}_\theta(s)$ and that*

$$\mathbf{E}_\theta\left(\sum_{i=1}^{n} |s_i|\right) < \infty \tag{4.3.36}$$

In this case, for any integrable stopping time T, we have

$$\mathbf{E}_\theta(S_T) = \mathbf{E}_\theta(T)\ \ \mathbf{E}_\theta(s) \tag{4.3.37}$$

- *Assume that, for any $\theta \in \Theta$, the increments $(s_i)_{1\leq i\leq n}$ are* independent *with the same mean $\mathbf{E}_\theta(s)$ and variance $\mathrm{var}_\theta(s) = \mathbf{E}_\theta[s - \mathbf{E}_\theta(s)]^2$, and that*

$$\mathbf{E}_\theta\left(\sum_{i=1}^{n} |s_i - \mathbf{E}_\theta(s)|\right)^2 < \infty \tag{4.3.38}$$

Then, for any integrable stopping time T, we have

$$\mathbf{E}_\theta[S_T - T\ \mathbf{E}_\theta(s)]^2 = \mathbf{E}_\theta(T)\ \ \mathrm{var}_\theta(s) \tag{4.3.39}$$

These two results are useful for computing the ASN of a sequential test. It results from this theorem that

$$\begin{aligned} \mathbf{E}_\theta(T) &= \frac{\mathbf{E}_\theta(S_T)}{\mathbf{E}_\theta(s)} \quad \text{when} \quad \mathbf{E}_\theta(s) \neq 0 \\ \mathbf{E}_\theta(T) &= \frac{\mathbf{E}_\theta(S_T^2)}{\mathbf{E}_\theta(s^2)} \quad \text{when} \quad \mathbf{E}_\theta(s) = 0 \end{aligned} \tag{4.3.40}$$

4.3.2.4 OC and ASN of the Sequential Probability Ratio Test

Two functions of the parameter θ characterize the SPRT. These are the OC $Q(\theta)$, which is the probability of accepting hypothesis $\mathbf{H}_0$ treated as a function of θ, and the ASN, which is the average sample number. Recall that when the SPRT is closed, we have $\beta(\theta) = 1 - Q(\theta)$, which provides us with the link between the power function, which is the criterion for fixed sample size and the operating characteristics, which is the criterion for sequential analysis. The first function describes the error probabilities of the sequential test. The second defines the mean number of sample points necessary for the acceptance of one hypothesis. We begin the discussion about these two functions OC and ASN by an "exact" computation based upon the solution of the Fredholm integral equations. Then we continue with *Wald's approximations* and bounds.

"Exact" computation For a given θ, let $\mathbf{P}_\theta(-a|z) = \mathbf{P}(z)$ be the probability that the cumulative sum of the SPRT starting from z reaches the lower boundary $-a$, and let $\mathbf{E}_\theta(T|z) = N(z)$. It should be clear that the OC is nothing but $\mathbf{P}(0)$ and the ASN is $N(0)$. Let us assume that the increments of the cumulative sum $(s_n)_{n\geq 1}$ are i.i.d. with density $f_\theta(s)$ and distribution function $F_\theta(s)$. Then, the OC function $Q(\theta)$ and ASN function $\mathbf{E}_\theta(T)$ can be computed by solving a system of linear equations that approximates the Fredholm integral equations of the second kind. Let us explain this now.

It is known [Page, 1954b, Kemp, 1958, Cox and Miller, 1965] that $\mathbf{P}$ and N are solutions of the following equations :

$$\mathbf{P}(z) = \int_{-\infty}^{-a-z} f_\theta(x)dx + \int_{-a}^{h} \mathbf{P}(x) f_\theta(x-z)dx, \quad -a \leq z \leq h \tag{4.3.41}$$

and

$$N(z) = 1 + \int_{-a}^{h} N(x) f_\theta(x-z)dx, \quad -a \leq z \leq h \tag{4.3.42}$$

The derivation of the formula for $\mathbf{P}(z)$ and $N(z)$ is based upon the theory of random walk with absorbing and reflecting boundaries (barriers). Here the random walk is nothing but the position of S_n in the system of coordinates (n, S_n).

Let us first consider the derivation of the formula for $\mathbf{P}(z)$. We follow the idea from chapter 2 of [Cox and Miller, 1965]. Let us assume without loss of generality that the symmetrical barriers $-a$ and $h = a$ are absorbing and that the random walk starts from z. Let $\mathbf{P}_n(z)$ be the probability of the absorption at the lower barrier $-a$ at or before the time n. In other words, $\mathbf{P}_n(z)$ is the following probability :

$$\mathbf{P}_n(z) = \mathbf{P}\left(\cup_{i=1}^{n} \{S_i \leq -a | \forall k \leq i-1, S_k < a\}\right) \tag{4.3.43}$$

The event on the right side of this equation can occur in two mutually exclusive ways :

- at the first time instant : $\{S_1 \leq -a\}$;
- at one of the time instants $2, \ldots, n$:

$$\{-a < S_1 < a\} \cap \left(\cup_{i=2}^{n} \{S_i \leq -a | \forall k \leq i-1, S_k < a\}\right) \tag{4.3.44}$$

Therefore, $\mathbf{P}_n(z)$ is the sum of the probabilities of these two events, which can be written as

$$\mathbf{P}_n(z) = \int_{-\infty}^{-a-z} f_\theta(x)dx + \int_{-a}^{a} \mathbf{P}_{n-1}(x) f_\theta(x-z)dx \tag{4.3.45}$$

where

$$\mathbf{P}_0(z) = \begin{cases} 1 & \text{when } z = -a \\ 0 & \text{when } z > -a \end{cases} \tag{4.3.46}$$

On the other hand, let $H_n(z)$ be the probability distribution function of the cumulative sum at time n when the barriers $-a$ and a are reflecting. It is known [Cox and Miller, 1965] that, for all n, $H_n(z)$ and $\mathbf{P}_n(-z)$ satisfy the same initial conditions and the same recurrence relation. Thus,

$$H_n(z) = \mathbf{P}_n(-z) \tag{4.3.47}$$

Therefore, taking the limit as $n \to \infty$ results in the existence of a mathematical equivalence between the equilibrium (limit) probability distribution $H(z)$ and the probability $\mathbf{P}(-z)$:

$$H(z) = \mathbf{P}(-z) \tag{4.3.48}$$

Furthermore, the equilibrium distribution function $H(z)$ is known to satisfy the Fredholm integral equation [Cox and Miller, 1965] :

$$H(z) = F_\theta(z-a) + \int_{-a}^{a} H(x) f_\theta(z-x)dx \tag{4.3.49}$$

Therefore, we finally get

$$\mathbf{P}(z) = \int_{-\infty}^{-a-z} f_\theta(x)dx + \int_{-a}^{a} \mathbf{P}(x) f_\theta(x-z)dx \tag{4.3.50}$$

Let us now consider the derivation of the formula for the ASN $N(z)$. We first consider the event

$$\Omega_1 = \{S_1 \leq -a\} \cup \{S_1 \geq h\} \tag{4.3.51}$$

If the first observation Y_1 is such that the event Ω_1 occurs, then the run length is equal to 1. Otherwise, if $-a < S_1 < h$, the SPRT continues with new starting point S_1 and average sample number $N(S_1)$. The average number $N(z)$ of SPRT steps until absorption - on either barrier - is the weighted sum of these two run lengths, the weights being the probabilities $\mathbf{P}(\Omega_1)$ and $1 - \mathbf{P}(\Omega_1)$:

$$N(z) = \mathbf{P}(\Omega_1) \cdot 1 + [1 - \mathbf{P}(\Omega_1)] \cdot \left[1 + \frac{\int_{-a}^{h} f_\theta(x-z)N(x)dx}{1 - \mathbf{P}(\Omega_1)}\right] \tag{4.3.52}$$

Furthermore, it results from the definition of Ω_1 that

$$1 - \mathbf{P}(\Omega_1) = \int_{-a}^{h} f_\theta(x-z)dx \tag{4.3.53}$$

Finally,

$$N(z) = 1 + \int_{-a}^{h} N(x) f_\theta(x-z)dx \tag{4.3.54}$$

For solving these integral equations numerically, we can replace them by systems of linear algebraic equations. This is discussed in section 5.2.

Wald's approximation (contd.) We assume that the random variable $s = \ln \frac{p_{\theta_1}(y)}{p_{\theta_0}(y)}$ satisfies the two following conditions :

- The moment generating function (mgf) $\psi_\theta(\varsigma)$:

$$\psi_\theta(\varsigma) = \int_{-\infty}^{\infty} e^{\varsigma x} dF_\theta(x) \tag{4.3.55}$$

  where F_θ is the cdf of s under $\mathbf{P}_\theta$, exists for all real ς. For the derivation of the local SPRT, we assume additionally that the mgf exists for any ς in the complex plane and for any θ, and is a continuous function of θ;
- There exists $\delta > 0$ such that

$$\mathbf{P}_\theta(e^s > 1+\delta) > 0 \text{ and } \mathbf{P}_\theta(e^s < 1-\delta) > 0 \tag{4.3.56}$$

Then the equation

$$\mathbf{E}_\theta(e^{-\omega_0 s}) = 1 \tag{4.3.57}$$

has only one nonzero real root $\omega_0 > 0$ if $\mathbf{E}_\theta(s) > 0$, $\omega_0 < 0$ if $\mathbf{E}_\theta(s) < 0$, and no nonzero real root if $\mathbf{E}_\theta(s) = 0$.

Let us begin with a useful and important result called *fundamental identity of sequential analysis*.

Theorem 4.3.3 (Fundamental identity). *Assume that (g, T) is a SPRT with boundaries $-a$ and h for testing between hypotheses $\mathbf{H}_0 = \{\theta : \theta = \theta_0\}$ and $\mathbf{H}_1 = \{\theta : \theta = \theta_1\}$. Let the increments $(s_i)_{1\leq i\leq n}$ be i.i.d. and assume that conditions (4.3.55)-(4.3.56) hold. Then, for every real ω, we have*

$$\mathbf{E}_\theta\{e^{-\omega S_n}[\psi_\theta(-\omega)]^{-n}\} = 1 \tag{4.3.58}$$

If we replace ω in (4.3.58) by the solution $\omega_0(\theta) \neq 0$ of the equation $\mathbf{E}_\theta(e^{-\omega_0 s}) = 1$, then we get the OC function :

$$Q(\theta) = \frac{\mathbf{E}_\theta(e^{-\omega_0 S_T}|S_T \geq h) - 1}{\mathbf{E}_\theta(e^{-\omega_0 S_T}|S_T \geq h) - \mathbf{E}_\theta(e^{-\omega_0 S_T}|S_T \leq -a)} \tag{4.3.59}$$

Then, under the assumption that *both excesses* of S_n over the boundaries $-a$ and h are *small*, we have

$$\mathbf{E}_\theta(e-\omega_0 S_T|S_T \leq -a) \approx e^{\omega_0 a} \quad \text{and} \quad \mathbf{E}_\theta(e-\omega_0 S_T|S_T \geq h) \approx e^{-\omega_0 h} \tag{4.3.60}$$

The Wald's approximation of the OC is thus

$$Q(\theta) \approx \tilde{Q}(\theta) = \frac{e^{-\omega_0(\theta)h} - 1}{e^{-\omega_0(\theta)h} - e^{\omega_0(\theta)a}} \quad \text{when} \quad \mathbf{E}_\theta(s) \neq 0 \tag{4.3.61}$$

The approximation $\tilde{Q}(\theta^*)$ when $\mathbf{E}_{\theta^*}(s) = 0$ can be obtained by taking the limit, when $\theta \to \theta^*$. Note that, when $\omega_0 \to 0$, we have $e^{-\omega_0 h} \approx 1 - \omega_0 h$, and thus

$$\tilde{Q}(\theta^*) = \frac{h}{h + a} \tag{4.3.62}$$

Note here that for $\theta = \theta_0$ or $\theta = \theta_1$, we have $\omega_0(\theta_0) = -1$ and $\omega_0(\theta_1) = 1$, and thus the expression of $Q(\theta)$ becomes very simple :

$$\begin{aligned} Q(\theta_0) &= \frac{e^h - 1}{e^h - e^{-a}} \\ Q(\theta_1) &= \frac{e^{-h} - 1}{e^{-h} - e^a} \end{aligned} \tag{4.3.63}$$

When the SPRT is closed, we have by definition

$$Q(\theta_0) = 1 - \alpha_0 \ \text{ and } \ Q(\theta_1) = \alpha_1 \tag{4.3.64}$$

Therefore, we get that (4.3.63) is nothing but the *Wald's approximations* (4.3.32) which we discussed before.

Under the assumptions that both excesses of S_n over the boundaries $-a$ and h are small, we have

$$\mathbf{E}_\theta[(S_T)^k|S_T \leq -a] \approx (-a)^k \quad \text{and} \quad \mathbf{E}_\theta[(S_T)^k|S_T \geq h] \approx h^k, \ \ k = 1, 2 \tag{4.3.65}$$

It results from Wald's identity (4.3.40) that

$$\begin{aligned} \mathbf{E}_\theta(T) \approx \tilde{E}_\theta(T) &= \frac{-aQ(\theta) + h(1 - Q(\theta))}{\mathbf{E}_\theta(s)} \quad \text{when} \quad \mathbf{E}_\theta(s) \neq 0 \\ \mathbf{E}_\theta(T) \approx \tilde{E}_\theta(T) &= \frac{a^2 Q(\theta) + h^2(1 - Q(\theta))}{\mathbf{E}_\theta(s^2)} \quad \text{when} \quad \mathbf{E}_\theta(s) = 0 \end{aligned} \tag{4.3.66}$$

From (4.3.66), we deduce that the approximation of the ASN for θ_0 or θ_1 can be written as a function of the error probabilities α_0 and α_1 :

$$\begin{aligned}\tilde{E}_{\theta_0}(T) &= \frac{(1-\alpha_0)\ln\frac{1-\alpha_0}{\alpha_1} - \alpha_0 \ln\frac{1-\alpha_1}{\alpha_0}}{-\mathbf{E}_{\theta_0}(s)} \\ \tilde{E}_{\theta_1}(T) &= \frac{(1-\alpha_1)\ln\frac{1-\alpha_1}{\alpha_0} - \alpha_1 \ln\frac{1-\alpha_0}{\alpha_1}}{\mathbf{E}_{\theta_1}(s)}\end{aligned}$$

Furthermore, recalling definition (4.1.42) of the Kullback information between two probability densities, we can rewrite these formulae as follows :

$$\tilde{E}_{\theta_0}(T) = \frac{(1-\alpha_0)\ln\frac{1-\alpha_0}{\alpha_1} - \alpha_0 \ln\frac{1-\alpha_1}{\alpha_0}}{\mathbf{K}(\theta_0,\theta_1)} \tag{4.3.67}$$

$$\tilde{E}_{\theta_1}(T) = \frac{(1-\alpha_1)\ln\frac{1-\alpha_1}{\alpha_0} - \alpha_1 \ln\frac{1-\alpha_0}{\alpha_1}}{\mathbf{K}(\theta_1,\theta_0)} \tag{4.3.68}$$

From these results, we deduce the two following important facts. First, for given error probabilities α_0 and α_1, the ASN of the SPRT is a function of Kullback information, and this leads us to use this information as a weak performance index for detection algorithms, as we explain in chapter 6. Second, in general, the ASN of the SPRT depends upon the true hypothesis. In other words, the problem of sequential hypotheses testing is not symmetric with respect to $\mathbf{H}_0$ and $\mathbf{H}_1$, and this fact plays an important role when detecting spectral changes, for example, as we explain in chapter 8.

Bounds We now discuss bounds for the OC and ASN functions. From the fundamental identity and (4.3.59) we find that, when $\mathbf{E}_\theta(s) \neq 0$, the bounds for the OC function are

$$\text{when } \omega_0 > 0, \quad \frac{e^{-\omega_0 h} - 1}{e^{-\omega_0 h} - \delta(\theta) e^{\omega_0 a}} \leq Q(\theta) \leq \frac{\eta(\theta) e^{-\omega_0 h} - 1}{\eta(\theta) e^{-\omega_0 h} - e^{\omega_0 a}} \tag{4.3.69}$$

and

$$\text{when } \omega_0 < 0, \quad \frac{e^{-\omega_0 h} - 1}{e^{-\omega_0 h} - \eta(\theta) e^{\omega_0 a}} \leq Q(\theta) \leq \frac{\delta(\theta) e^{-\omega_0 h} - 1}{\delta(\theta) e^{-\omega_0 h} - e^{\omega_0 a}} \tag{4.3.70}$$

where

$$\begin{aligned}\eta(\theta) &= \inf_{\xi>1} \xi \, \mathbf{E}_\theta\left(e^{-\omega_0 S_T} | e^{-\omega_0 S_T} \leq \tfrac{1}{\xi}\right) \\ \delta(\theta) &= \sup_{0<\rho<1} \rho \, \mathbf{E}_\theta\left(e^{-\omega_0 S_T} | e^{-\omega_0 S_T} \geq \tfrac{1}{\rho}\right)\end{aligned} \tag{4.3.71}$$

For $\theta = \theta^*$ such that $\mathbf{E}_{\theta^*}(s) = 0$, bounds for $Q(\theta^*)$ can be obtained by taking the limit, when $\theta \to \theta^*$, of the formula (4.3.69) or (4.3.70).

Let us consider the ASN function. The expectation $\mathbf{E}_\theta(S_T)$ obviously consists of two conditional expectations :

$$\mathbf{E}_\theta(S_T | S_T \leq -a) \text{ and } \mathbf{E}_\theta(S_T | S_T \geq h) \tag{4.3.72}$$

weighted by the probabilities $Q(\theta)$ and $1 - Q(\theta)$. Therefore, it results directly from Wald's identity that the ASN of the SPRT can be written as

$$\mathbf{E}_\theta(T) = \frac{\mathbf{E}_\theta(S_T | S_T \leq -a) Q(\theta) + \mathbf{E}_\theta(S_T | S_T \geq h)[1 - Q(\theta)]}{\mathbf{E}_\theta(s)} \tag{4.3.73}$$

when $\mathbf{E}_\theta(s) \neq 0$. Consequently, bounds for the ASN function can be written as follows :

$$\frac{[-a+\gamma_1(\theta)]Q(\theta)+h[1-Q(\theta)]}{\mathbf{E}_\theta(s)} \leq \mathbf{E}_\theta(T) \leq \frac{-aQ(\theta)+[h+\gamma_2(\theta)][1-Q(\theta)]}{\mathbf{E}_\theta(s)} \tag{4.3.74}$$

when $\mathbf{E}_\theta(s) > 0$, and for $\mathbf{E}_\theta(s) < 0$ these inequalities must be written conversely.

For $\theta = \theta^*$, the bounds for the ASN function can be written as

$$\begin{aligned} \frac{a^2Q(\theta^*)+h^2[1-Q(\theta^*)]}{\mathbf{E}_{\theta^*}(s^2)} &\leq \mathbf{E}_{\theta^*}(T) \\ \leq \frac{[a^2-2a\gamma_0(\theta^*)+\gamma_3(\theta^*)]Q(\theta^*)+[h^2+2h\gamma_1(\theta^*)+\gamma_4(\theta^*)][1-Q(\theta^*)]}{\mathbf{E}_{\theta^*}(s^2)} & \end{aligned} \tag{4.3.75}$$

where

$$\begin{aligned} \gamma_1(\theta) &= \inf_{r>0} \mathbf{E}_\theta(s+r|s \leq -r < 0) \\ \gamma_2(\theta) &= \sup_{r>0} \mathbf{E}_\theta(s-r|s \geq r > 0) \\ \gamma_3(\theta) &= \sup_{r>0} \mathbf{E}_\theta[(s+r)^2|s \leq -r < 0] \\ \gamma_4(\theta) &= \sup_{r>0} \mathbf{E}_\theta[(s-r)^2|s \geq r > 0] \end{aligned} \tag{4.3.76}$$

We now continue to discuss our main examples.

Example 4.3.4 (Mean in a Gaussian sequence - contd.). *The OC and ASN of the SPRT corresponding to a change in the mean θ of an independent Gaussian sequence can be computed as follows. The increment of the cumulative sum is*

$$s_k = \frac{\theta_1-\theta_0}{\sigma^2}\left(y_k - \frac{\theta_0+\theta_1}{2}\right) \tag{4.3.77}$$

and is thus distributed as a Gaussian random variable with mean

$$\mu_s = \frac{\theta_1-\theta_0}{\sigma^2}\left(\theta - \frac{\theta_0+\theta_1}{2}\right) \tag{4.3.78}$$

and variance

$$\sigma_s^2 = \frac{(\theta_1-\theta_0)^2}{\sigma^2} \tag{4.3.79}$$

Therefore the solution to the equation

$$\mathbf{E}(e^{-\omega_0 s}) = \int_{-\infty}^{\infty} e^{-\omega_0 s} \frac{1}{\sigma_s\sqrt{2\pi}} e^{-\frac{(s-\mu_s)^2}{2\sigma_s^2}} ds = 1 \tag{4.3.80}$$

is given by

$$\omega_0 = \frac{2\mu_s}{\sigma_s^2} = \frac{2}{\theta_1-\theta_0}\left(\theta - \frac{\theta_0+\theta_1}{2}\right) \tag{4.3.81}$$

From (4.3.61) the approximation of the OC function can be written as

$$\begin{aligned} \tilde{Q}(\theta) &= \frac{e^{-\frac{2\mu_s h}{\sigma_s^2}}-1}{e^{-\frac{2\mu_s h}{\sigma_s^2}} - e^{\frac{2\mu_s a}{\sigma_s^2}}} \quad \textit{when} \quad \mu_s \neq 0 \\ \tilde{Q}(\theta) &= \frac{h}{h+a} \qquad \textit{when} \quad \mu_s = 0 \end{aligned} \tag{4.3.82}$$

Similarly, from (4.3.66) the approximation of the ASN function is

$$\begin{aligned}\tilde{E}_{\mu_s}(T) &= \frac{1}{\mu_s}\left[\frac{1-e^{\frac{2\mu_s a}{\sigma_s^2}}}{e^{-\frac{2\mu_s h}{\sigma_s^2}}-e^{\frac{2\mu_s a}{\sigma_s^2}}}\,h-\frac{e^{-\frac{2\mu_s h}{\sigma_s^2}}-1}{e^{-\frac{2\mu_s h}{\sigma_s^2}}-e^{\frac{2\mu_s a}{\sigma_s^2}}}\,a\right] \quad \textit{when} \quad \mu_s \neq 0 \\ \tilde{E}_{\mu_s}(T) &= \frac{ah}{\sigma_s^2} \quad \textit{when} \quad \mu_s = 0 \end{aligned} \tag{4.3.83}$$

The comparison between these two approximations for the OC and ASN and the formulas (3.1.92), (3.1.98), and (3.1.99) shows that Wald's approximations are equivalent to the approximation of the cumulative sum by a Brownian motion.

Example 4.3.5 (Mean in a Gaussian autoregressive sequence - contd.). *We now relax the assumption of independence and test the mean value θ in an autoregressive Gaussian sequence $(y_k)_k$ with variance σ_y^2. For simplicity, we compute the OC and ASN in the AR(1) case :*

$$y_k = a_1 y_{k-1} + v_k + (1-a_1)\theta \tag{4.3.84}$$

where $(v_k)_k$ is a white noise sequence with variance $\sigma^2 = (1-a_1^2)\sigma_y^2$ and where we assume $|a_1| < 1$. The increment of the cumulative sum is then

$$s_k = \frac{\theta_1-\theta_0}{(1+a_1)\sigma_y^2}\left[y_k - a_1 y_{k-1} - \frac{(1-a_1)(\theta_0+\theta_1)}{2}\right] \tag{4.3.85}$$

and thus

$$\mathbf{E}_\theta(s) = \frac{(1-a_1)(\theta_1-\theta_0)}{(1+a_1)\sigma_y^2}\left(\theta - \frac{\theta_0+\theta_1}{2}\right) \tag{4.3.86}$$

$$\mathrm{var}(s) = \frac{(1-a_1)(\theta_1-\theta_0)^2}{(1+a_1)\sigma_y^2} \tag{4.3.87}$$

Thus, from (4.3.83),

$$\begin{aligned}\tilde{E}_{\mu_s}(T) &= \frac{-a\tilde{Q}(\mu_s)+h[1-\tilde{Q}(\mu_s)]}{\frac{(1-a_1)(\theta_1-\theta_0)}{(1+a_1)\sigma_y^2}\left(\theta-\frac{\theta_0+\theta_1}{2}\right)} \quad \textit{when} \quad \theta \neq \frac{\theta_0+\theta_1}{2} \\ \tilde{E}_{\mu_s}(T) &= \frac{ah}{\frac{(1-a_1)(\theta_1-\theta_0)}{(1+a_1)\sigma_y^2}} \quad \textit{when} \quad \theta = \frac{\theta_0+\theta_1}{2}\end{aligned} \tag{4.3.88}$$

Let us discuss this formula with respect to the corresponding formula (4.3.83) in the independent case. It results from (4.3.88) that, for fixed error probabilities α_0 and α_1, the ASN of the SPRT is a function of the autoregressive coefficient a_1 (or equivalently of the serial correlation in the sequence $(y_k)_k$. For positive correlation $(a_1 > 0)$, the ASN is $\frac{1+a_1}{1-a_1}$ times greater than the ASN of the i.i.d. case. For negative correlation $(a_1 < 0)$, the ASN is $\frac{1+a_1}{1-a_1}$ times less than the ASN of the i.i.d. case.

4.3.3 Local Hypotheses Approach in Sequential Analysis

Let us now consider again a *local hypotheses approach*, where the distance $\theta_1 - \theta_0$ is small. In this case, we first continue to use the SPRT as defined in (4.3.5) and we use the local approach for the investigation of

its properties (ASN and OC). Second, we use the local approach for the design of the algorithm through the efficient score. The criteria then are slightly different, as we explain below.

In the case of sequential analysis, the local approach has some specific features. From section 4.2, we know that, for investigating the probability properties of fixed sample size tests (namely, the power function $\beta(\theta)$) and for computing the threshold λ, it is necessary to know the distribution of the raw data $\mathcal{Y}_1^n$. In general, this investigation is a complex problem, because the distribution of the LR is non-Gaussian. But, in some sense, the situation in sequential analysis is a little better. Actually, from formulas (4.3.61) and (4.3.66) we find that if the parameter is exactly equal to θ_0 or θ_1, then the OC and the ASN do *not* depend upon the distribution of the increment s of the cumulative sum S_n. Another problem of interest from the robustness point of view is the computation of the OC and ASN for some θ that is in the neighborhood of θ_0 or θ_1. In this case, the OC and the ASN can be computed using the local approach, as we explain now.

4.3.3.1 ASN function

Let $\mathcal{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta},\ (\Theta \subset \mathbf{R})$ be a family with scalar parameter θ. We consider the following simple *local* hypotheses :

$$\mathbf{H}_0 = \{\theta = \theta_0\} \ \text{ against } \ \mathbf{H}_1 = \{\theta = \theta_1\} \ \text{ when } \ \theta_1 = \theta_0 + \nu, \ \ \nu \to 0 \tag{4.3.89}$$

Recall that, under some general regularity assumptions about the pdf of the family $\mathcal{P}$ (see subsection 4.1.2), we have the following approximation for the increment $s(Y) = \ln \frac{p_{\theta_1}(Y)}{p_{\theta_0}(Y)}$ of the LLR :

$$\begin{array}{rcrcl} \mathbf{K}(\theta_0,\theta_1) & = & -\mathbf{E}_{\theta_0}(s) & = & \frac{1}{2}\nu^2\ \mathbf{I}(\theta_0) + o(\nu^2) \\ \mathbf{K}(\theta_1,\theta_0) & = & \mathbf{E}_{\theta_1}(s) & = & \frac{1}{2}\nu^2\ \mathbf{I}(\theta_1) + o(\nu^2) \\ & & \mathbf{E}_{\theta}(s) & = & \nu\left(\tilde{\nu} - \frac{1}{2}\nu\right)\ \mathbf{I}(\theta) + o(\nu^2) \end{array} \tag{4.3.90}$$

where $\tilde{\nu} = \theta - \theta_0$. From the approximations (4.3.67), the ASN for θ_0 or θ_1 can be approximated as

$$\begin{array}{rcl} \mathbf{E}_{\theta_0}(T) \approx \tilde{E}_{\theta_0}(T) & = & \dfrac{(1-\alpha_0)\ln\frac{1-\alpha_0}{\alpha_1} - \alpha_0 \ln\frac{1-\alpha_1}{\alpha_0}}{\frac{1}{2}\nu^2\ \mathbf{I}(\theta_0) + o(\nu^2)} \\ \mathbf{E}_{\theta_1}(T) \approx \tilde{E}_{\theta_1}(T) & = & \dfrac{(1-\alpha_1)\ln\frac{1-\alpha_1}{\alpha_0} - \alpha_1 \ln\frac{1-\alpha_0}{\alpha_1}}{\frac{1}{2}\nu^2\ \mathbf{I}(\theta_1) + o(\nu^2)} \end{array}$$

In section 4.2, we explain that the speed of convergence in local situations should be of the order of magnitude of $\frac{\text{constant}}{\sqrt{N}}$. For the SPRT, the sample size is a random variable and the above approximations for the ASN provide us with the relevant sequential counterpart for the order of magnitude for the deviation ν between the two hypotheses :

$$\nu \approx \frac{\text{constant}(\alpha_0,\alpha_1)}{\sqrt{\text{ASN}}} \tag{4.3.91}$$

For other values of θ, the approximation of the ASN is the following :

$$\mathbf{E}_\theta(T) \approx \tilde{E}_\theta(T) = \frac{-aQ(\theta) + h(1-Q(\theta))}{\nu(\tilde{\nu} - \frac{1}{2}\nu)\ \mathbf{I}(\theta) + o(\nu^2)} \tag{4.3.92}$$

4.3.3.2 OC function

Let us now discuss the problem of the computation of the OC in local case. Assume again that, in the neighborhood of θ_0 or θ_1, the first three derivatives of the pdf $p_\theta(Y)$ with respect to θ exist and that the

following integrals also exist :

$$\begin{aligned}\mathbf{E}_\theta(|s|^k) &= \int \left|\ln \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)}\right|^k p_\theta(x)\, dx, \quad k=2,3 \\ I_k(\theta) &= \int \left|\frac{\partial \ln p_\theta(x)}{\partial \theta}\right|^k p_\theta(x)\, dx > 0, \quad k=2,3\end{aligned}$$

Further details can be found in [Wald, 1947, Basharinov and Fleishman, 1962]. Under these assumptions, and using Taylor's expansion exactly as we did in subsection 4.1.2, we have

$$e^s \approx 1+\nu \left.\frac{\partial \ln p_\theta(Y)}{\partial \theta}\right|_{\theta=\theta_0} + \frac{\nu^2}{2}\left\{\left.\frac{\partial^2 \ln p_\theta(Y)}{\partial \theta^2}\right|_{\theta=\theta_0} + \left[\left.\frac{\partial \ln p_\theta(Y)}{\partial \theta}\right|_{\theta=\theta_0}\right]^2\right\} \tag{4.3.93}$$

$$\begin{aligned}e^{-\omega_0 s} \approx\ & 1-\omega_0\nu \left.\frac{\partial \ln p_\theta(Y)}{\partial \theta}\right|_{\theta=\theta_0} - \frac{\omega_0\nu^2}{2}\left\{\left.\frac{\partial^2 \ln p_\theta(Y)}{\partial \theta^2}\right|_{\theta=\theta_0} + \left[\left.\frac{\partial \ln p_\theta(Y)}{\partial \theta}\right|_{\theta=\theta_0}\right]^2\right\} \\ & + \frac{\omega_0(1+\omega_0)\nu^2}{2}\left[\left.\frac{\partial \ln p_\theta(Y)}{\partial \theta}\right|_{\theta=\theta_0}\right]^2\end{aligned} \tag{4.3.94}$$

Recall that $\frac{\partial \ln p_\theta(Y)}{\partial \theta}$ is the *efficient score*. Taking the expectation $\mathbf{E}_\theta$ of both sides of (4.3.94), we get

$$\mathbf{E}_\theta(e^{-\omega_0 s}) = \int e^{-\omega_0 s} p_\theta(x)dx \approx 1-\omega_0\left(\tilde{\nu}-\nu\,\frac{\omega_0+1}{2}\right)\nu\,\mathbf{I}(\theta_0) \tag{4.3.95}$$

Finally, we get the nonzero root ω_0 of the equation $\mathbf{E}_\theta(e^{-\omega_0 s})=1$ for some θ in the neighborhood of θ_0 or θ_1 :

$$\omega_0(\theta) \approx \frac{2}{\theta_1-\theta_0}\left(\theta - \frac{\theta_1+\theta_0}{2}\right) \tag{4.3.96}$$

The OC can then be estimated using (4.3.61).

Let us continue the discussion about the comparison between the local hypotheses approach in nonsequential hypotheses testing and in sequential analysis. From the comparison between (4.3.90) and (4.3.78) on one hand, and between (4.3.96) and (4.3.81) on the other, it results that, in the case of local hypotheses, the ASN and OC can be computed exactly as in the Gaussian independent case, replacing μ_s by the values given in (4.3.90) and ω_0 by (4.3.96). In other words, in the case of sequential analysis, the use of the local approach for the investigation of the properties OC and ASN follows exactly the same lines as for nonsequential hypotheses testing.

4.3.3.3 Locally most powerful sequential test

In subsection 4.2.9 it is shown that optimal nonsequential tests for local one-sided hypotheses are based upon the *efficient score* :

$$z = \frac{\partial \ln p_\theta(Y)}{\partial \theta} \tag{4.3.97}$$

One particular optimal property for local sequential test is proven in [Berk, 1975]. Let us consider the following composite hypotheses :

$$\mathbf{H}_0 = \{\theta\ : \theta \le \theta^*\} \text{ and } \mathbf{H}_1 = \{\theta\ : \theta > \theta^*\} \tag{4.3.98}$$

where θ is the scalar parameter of a family $\mathcal{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta}$. Let (g,T) be a sequential test, and note $\dot\beta(\theta^*) = \frac{\partial\beta(\theta)}{\partial\theta}\Big|_{\theta=\theta^*}$. The *local characteristic* of the sequential test (g,T) is the triplet $(\alpha_0(g), \mathbf{E}(T), \dot\beta(\theta^*))$. The interpretation of this local characteristic is natural and obvious. Let us consider the class K_{α_0} of sequential tests with fixed size α_0. We want to find the test (g,T) with maximum *power* $\beta(\theta)$ for local alternative $\theta > \theta^*$. In other words, we want to have maximum *local* slope $\dot\beta(\theta^*)$ of the power function $\beta(\theta)$ at θ^*.

Let us define the following sequential test :

$$g(\mathcal{Y}_1^T) = \begin{cases} 1 & \text{when} \quad \mathcal{Z}_T \geq h \\ 0 & \text{when} \quad \mathcal{Z}_T \leq -a \end{cases} \tag{4.3.99}$$

where

$$\begin{aligned} T &= T_{-a,h} = \min\{n \geq 1 : (\mathcal{Z}_n \geq h) \cup (\mathcal{Z}_n \leq -a)\} \\ \mathcal{Z}_n &= \sum_{i=1}^n z_i^* = \sum_{i=1}^n \frac{\partial \ln p_\theta(Y_i)}{\partial\theta}\Big|_{\theta=\theta^*} \end{aligned}$$

Under some regularity conditions, test (4.3.99) is *locally most powerful* (LMP) [Berk, 1975]. This means that, for another sequential test $(\tilde g, \tilde T)$ of $\mathbf{H}_0$ against $\mathbf{H}_1$, which has local triplet $(\alpha_0(\tilde g), \mathbf{E}(\tilde T), \dot{\tilde\beta}(\theta^*))$ and $\mathbf{E}(\tilde T) \leq \mathbf{E}(T)$, the following inequality holds true :

$$\dot{\tilde\beta}(\theta^*) \leq \dot\beta(\theta^*) \tag{4.3.100}$$

Thus, the sequential test (4.3.99) is LMP among all sequential tests of size $\tilde\alpha_0 \leq \alpha_0$ and for which the ASN satisfies $\mathbf{E}(\tilde T) \leq \mathbf{E}(T)$.

Let us continue the discussion about this local sequential test and show that, under some conditions, it coincides with the usual SPRT. Assume that the family $\mathcal{P}$ is exponential and has a density of the form

$$p_\theta(y) = h(y)\ e^{\theta y - d(\theta)} \tag{4.3.101}$$

where $d(\theta)$ is infinitely differentiable and strictly convex. The logarithm of the likelihood ratio (4.2.35) in this case is

$$s(Y) = \ln\frac{p_{\theta_1}(Y)}{p_{\theta_0}(Y)} = (\theta_1 - \theta_0)Y + d(\theta_0) - d(\theta_1) \tag{4.3.102}$$

and the efficient score is

$$z^* = \frac{\partial \ln p_\theta(Y)}{\partial\theta}\Big|_{\theta=\theta^*} = Y - \dot d(\theta^*) \tag{4.3.103}$$

Let us define the simple hypotheses $\mathbf{H}_0'$ and $\mathbf{H}_1'$:

$$\mathbf{H}_0' = \{\theta = \theta_0\} \ \text{ and } \ \mathbf{H}_1' = \{\theta = \theta_1\} \tag{4.3.104}$$

where $\theta_0 < \theta^* < \theta_1$. The SPRT (g', T') for testing $\mathbf{H}_0'$ against $\mathbf{H}_1'$ is

$$g'(\mathcal{Y}_1^{T'}) = \begin{cases} 1 & \text{when} \quad S_{T'} \geq h' \\ 0 & \text{when} \quad S_{T'} \leq -a' \end{cases} \tag{4.3.105}$$

where

$$\begin{aligned} T' &= T'_{-a',h'} = \min\{n \geq 1 : (S_n \geq h') \cup (S_n \leq -a')\} \\ S_n &= \sum_{i=1}^n (\theta_1 - \theta_0)Y_i - n[d(\theta_0) - d(\theta_1)] \end{aligned}$$

Suppose that there is a continuum of pairs (θ_0, θ_1) of points in the neighborhood of θ^* for which the following equation holds true :

$$\dot{d}(\theta^*) = \frac{d(\theta_1) - d(\theta_0)}{\theta_1 - \theta_0} \tag{4.3.106}$$

The existence of such a continuum is proved in [Berk, 1975]. Then, the SPRT for the simple hypotheses $\mathbf{H}'_0$ and $\mathbf{H}'_1$ (4.3.104) is *simultaneously LMP sequential test* for the local hypotheses $\mathbf{H}_0$ and $\mathbf{H}_1$ (4.3.98). Note here that the point θ^* is a special point for the SPRT for hypotheses $\mathbf{H}'_0$ and $\mathbf{H}'_1$, because it satisfies

$$\mathbf{E}_{\theta^*}(s) = 0 \tag{4.3.107}$$

4.3.4 Sequential Testing Between Two Composite Hypotheses

In the previous subsection, we described the sequential testing between two simple hypotheses. The case of composite hypotheses is more useful and important from a practical point of view. Unfortunately, this case is much more complicated and the amount of available theoretical results is lower than in the case of simple hypotheses.

As mentioned in the subsection 2.4.1, two possible solutions for the case of composite hypotheses are suggested in [Wald, 1947] : the method of weighting function and the generalized likelihood ratio algorithm. We now discuss the method of weighting function because this approach will be widely used in chapters 7, 8, and 9.

4.3.4.1 Method of Weighting Functions

Let $\mathcal{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta},\ \Theta \subset \mathbf{R}^\ell$, be a parametric family, where θ is an ℓ-dimensional parameter, and consider the composite hypotheses $\mathbf{H}_0 = \{\theta :\ \theta \in \Theta_0\}$ and $\mathbf{H}_1 = \{\theta :\ \theta \in \Theta_1\}$. In the case of simple hypotheses $\mathbf{H}'_0 = \{\theta :\ \theta = \theta_0\}$ and $\mathbf{H}'_1 = \{\theta :\ \theta = \theta_1\}$, the SPRT is based upon the likelihood ratio Λ_n :

$$\Lambda_n = \frac{p_{\theta_1}(\mathcal{Y}_1^n)}{p_{\theta_0}(\mathcal{Y}_1^n)} \tag{4.3.108}$$

But, in the case of composite hypotheses, the LR cannot be defined, because θ_0 and θ_1 are unknown. The *weighting function approach* consists of introducing two weighting functions $f_i(\theta) \geq 0$ such that $\int f_i(\theta)d\theta = 1,\ \ (i = 0, 1)$. In other words, we assume that, under hypothesis $\mathbf{H}_i$, $\mathcal{Y}_1^n$ has the pdf $p_i(\mathcal{Y}_1^n)$, given by

$$p_i(\mathcal{Y}_1^n) = \int p_{\theta_i}(\mathcal{Y}_1^n) f_i(\theta_i) d\theta_i \tag{4.3.109}$$

Consequently, the idea is to replace Λ_n (4.3.108) by the *weighted likelihood ratio* (WLR) :

$$\tilde{\Lambda}_n = \frac{\int p_{\theta_1}(\mathcal{Y}_1^n) f_1(\theta_1) d\theta_1}{\int p_{\theta_0}(\mathcal{Y}_1^n) f_0(\theta_0) d\theta_0} \tag{4.3.110}$$

In other words, we use the LR for testing between two simple "weighted" hypotheses $\widetilde{\mathbf{H}}_0$ and $\widetilde{\mathbf{H}}_1$, given by

$$\widetilde{\mathbf{H}}_0 = \{\mathcal{L}(\mathcal{Y}_1^n) = \mathbf{P}_0\} \ \text{ and } \ \widetilde{\mathbf{H}}_1 = \{\mathcal{L}(\mathcal{Y}_1^n) = \mathbf{P}_1\} \tag{4.3.111}$$

where $F_i(y) = \int_{-\infty}^{y} p_i(x)dx$ is the cdf of $\mathbf{P}_i$.

Now, the corresponding sequential test based upon the observations $(Y_i)_{i\geq 1}$, consists, at time n, of making one of the following decisions :

- accept $\widetilde{\mathbf{H}}_0$ when $\tilde{S}_n = \ln \tilde{\Lambda}_n \leq -\epsilon$;
- accept $\widetilde{\mathbf{H}}_1$ when $\tilde{S}_n \geq h$;
- continue to observe when $-\epsilon < \tilde{S}_n < h$.

Note here that the weighting function $f(\theta)$ may be interpreted as pdf of the *a priori distribution*. Therefore, there exists a correspondence between the method of weighting function and the Bayes and the minmax tests (see subsection 4.2.6). The key issue in this approach is the choice of a convenient weighting function $f(\theta)$ for the parameter θ in the sets Θ_0 and Θ_1. In many cases, it is useful to use the invariance properties with respect to some transformation in order to guess the convenient weighting function.

4.3.4.2 Invariant Sequential Test

Let us now consider the special case of the Gaussian distribution, which is very important for the subsequent chapters of this book. The general theory of the sequential invariant tests can be found in [Jackson and Bradley, 1961, W.Hall *et al*., 1965, Ghosh, 1970, Laï, 1981].

Derivation of the sequential χ^2-test We start from the simplest case of unit covariance matrix and then continue our discussion with the case of general covariance matrix.

Unit covariance matrix Assume that we have an r-dimensional random vector Y, with distribution $\mathcal{L}(Y) = \mathcal{N}(\theta, I)$. Consider the problem of testing between the two following hypotheses :

$$\mathbf{H}_0 = \{\theta : \ \|\theta\| \leq a\} \ \text{ and } \ \mathbf{H}_1 = \{\theta : \ \|\theta\| \geq b\}, \ \text{ where } \ b > a \tag{4.3.112}$$

The nonsequential version of this problem is discussed in example 4.2.4. By analogy with this example, we start from the invariant properties of the normal distribution $\mathcal{N}(\theta, I)$ and get a convenient choice of the weighting functions $f_i(\theta)$. These are constant functions concentrated on the spheres :

$$\tilde{\Theta}_0 = \{\theta : \ \|\theta\| = a\} \ \text{ and } \ \tilde{\Theta}_1 = \{\theta : \ \|\theta\| = b\} \tag{4.3.113}$$

Note here that $f_i(\theta)$ is analogous to the pdf of the least favorable distributions $\mathbf{P}_i$ in example 4.2.4. The WLR (4.3.110) is equal to the left side of inequality (4.2.48). It is shown in example 4.2.4 that $\|\bar{Y}_n\|^2$, where

$$\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{4.3.114}$$

is a sufficient statistic for testing between hypotheses $\mathbf{H}_0$ and $\mathbf{H}_1$. Since $\mathcal{L}(\sqrt{n}\bar{Y}_n) = \mathcal{N}(\sqrt{n}\theta, I)$, it results that

$$\mathcal{L}(n\|\bar{Y}_n\|^2) = \chi^2(r, \lambda_n) \tag{4.3.115}$$

where

$$\lambda_n = n \sum_{i=1}^{r} \theta_i^2 \tag{4.3.116}$$

is the noncentrality parameter. Therefore, the initial problem (4.3.112) can be replaced by the following hypotheses testing problem concerning the noncentrality parameter λ of the χ^2-distribution :

$$\mathbf{H}_0 = \{\lambda : \ \lambda = a^2\} \ \text{ and } \ \mathbf{H}_1 = \{\lambda : \ \lambda = b^2\}, \ \text{ where } \ b > a \tag{4.3.117}$$

In other words, for solving the initial problem, the following must be done. First, we have to transform the initial data $(Y_1, \ldots, Y_n)$ into the sequence of sufficient statistics $(\|\bar{Y}_1\|^2, 2\|\bar{Y}_2\|^2, \ldots, n\|\bar{Y}_n\|^2)$. Second, we have to compute the log-likelihood ratio :

$$\tilde{S}_n = \ln \frac{p_{b^2}(\|\bar{Y}_1\|^2, 2\|\bar{Y}_2\|^2, \ldots, n\|\bar{Y}_n\|^2)}{p_{a^2}(\|\bar{Y}_1\|^2, 2\|\bar{Y}_2\|^2, \ldots, n\|\bar{Y}_n\|^2)} \tag{4.3.118}$$

It turns out that the direct computation of this ratio using the joint densities is very difficult, and that it is of key interest to use the following factorization :

$$p_\lambda(\|\bar{Y}_1\|^2, 2\|\bar{Y}_2\|^2, \ldots, n\|\bar{Y}_n\|^2) = p_\lambda(n\|\bar{Y}_n\|^2)\ \beta(\|\bar{Y}_1\|^2, 2\|\bar{Y}_2\|^2, \ldots, n\|\bar{Y}_n\|^2) \tag{4.3.119}$$

which results from Cox's theorem [Jackson and Bradley, 1961]. In this formula, β is a function of the sufficient statistics. Using this factorization, it results from example 3.1.4 that the logarithm of the LR of hypotheses (4.3.117) can be written as

$$\tilde{S}_n = -n\,\frac{b^2 - a^2}{2} + \ln G\left(\frac{r}{2}, \frac{b^2 n^2 \chi_n^2}{4}\right) - \ln G\left(\frac{r}{2}, \frac{a^2 n^2 \chi_n^2}{4}\right) \tag{4.3.120}$$

$$\chi_n^2 = \bar{Y}_n^T \bar{Y}_n = \|\bar{Y}_n\|^2 \tag{4.3.121}$$

where

$$G(m, x) = 1 + \frac{x}{m} + \frac{x^2}{m(m+1)2!} + \ldots + \frac{x^n}{m(m+1)\ldots(m+n-1)n!} + \ldots \tag{4.3.122}$$

is the generalized hypergeometric function. Note here that the LR of the χ^2-distribution is a monotone function of the sufficient statistic $\|\bar{Y}_n\|^2$. This sequential test is known as the *sequential χ^2-test*.

General covariance matrix We now investigate the more complex case of a general covariance matrix. Assume that we have an r-dimensional random vector Y with distribution $\mathcal{L}(Y) = \mathcal{N}(\theta, \Sigma)$. Consider the problem of testing between the two following hypotheses :

$$\mathbf{H}_0 = \{\theta : \theta^T \Sigma^{-1} \theta \leq a^2\} \text{ and } \mathbf{H}_1 = \{\theta : \theta^T \Sigma^{-1} \theta \geq b^2\} \text{ where } b > a \tag{4.3.123}$$

It is shown in example 4.2.5 that it is possible to transform the hypotheses testing problem (4.3.123) into the previous one (4.3.112). The formula for the logarithm of the LR (4.3.120) holds true, but the χ^2 statistic (4.3.121) must be replaced by

$$\chi_n^2 = \bar{Y}_n^T \Sigma^{-1} \bar{Y}_n \tag{4.3.124}$$

Properties of the sequential χ^2-test From the previous sections we know that a sequential test must be closed. The sufficient conditions under which a SPRT (g, T) is closed is discussed in subsection 4.3.2. Let us show one important result about a termination property of the sequential χ^2-test.

Theorem 4.3.4 (Ghosh). *The sequential χ^2-test defined by*

- *accept $\widetilde{\mathbf{H}}_0$ when $\tilde{S}_n = \ln \tilde{\Lambda}_n \leq -\epsilon$;*
- *accept $\widetilde{\mathbf{H}}_1$ when $\tilde{S}_n \geq h$;*
- *continue to observe when $-\epsilon < \tilde{S}_n < h$*

where

$$\tilde{S}_n = -n\frac{b^2-a^2}{2} + \ln G\left(\frac{r}{2}, \frac{b^2 n^2 \chi_n^2}{4}\right) - \ln G\left(\frac{r}{2}, \frac{a^2 n^2 \chi_n^2}{4}\right) \quad (4.3.125)$$

$$\chi_n^2 = \bar{Y}_n^T \Sigma^{-1} \bar{Y}_n \quad (4.3.126)$$

is closed for any $r \geq 1,\ \ 0 \leq a < b$ *and thresholds* $-\epsilon < h$.

The Wald's inequalities (4.3.20) for the error probabilities $\alpha_0(g)$ and $\alpha_1(g)$ remain valid. But, because the logarithm of the LR (4.3.120) is *not* a sum of independent random variables, the Wald's approximation of the ASN is not valid for the invariant sequential tests. However, the following asymptotic result about the general SPRT

$$T = \min\{n \geq 1\ : (S_n \geq h) \cup (S_n \leq -\epsilon)\} \quad (4.3.127)$$

is useful for the calculation of the ASN of the invariant sequential tests. We refer the reader to [Berk, 1973] for details and proof.

Theorem 4.3.5 (Berk). *Suppose that with probability* 1 *(w.p.1) :*

$$\lim_{n\to\infty} \frac{S_n}{n} = \varrho, \quad \textit{where} \quad \varrho \in (0, \infty] \quad (4.3.128)$$

Then, w.p.1,

$$\lim_{(\epsilon,h)\to\infty} \mathbf{P}(S_n > h) = 1 \quad (4.3.129)$$

$$\lim_{(\epsilon,h)\to\infty} \frac{T}{h} = \frac{1}{\varrho} \quad (4.3.130)$$

Moreover, if for some $\tilde{\varrho} \in (0, \varrho)$*, the "large deviation" probability* $p_n = \mathbf{P}(\frac{S_n}{n} < \tilde{\varrho})$ *satisfies*

$$\lim_{n\to\infty} np_n = 0 \quad (4.3.131)$$

$$\sum_{i=1}^{\infty} p_n < \infty \quad (4.3.132)$$

then also

$$\lim_{(\epsilon,h)\to\infty} \frac{\mathbf{E}(T)}{h} = \frac{1}{\varrho} \quad (4.3.133)$$

If $\varrho < 0$*, then the analogous statements hold true*.

The practical interpretation of this theorem is simple. If the log-likelihood ratio of some general SPRT satisfies conditions (4.3.128), (4.3.131), and (4.3.132), then, when the thresholds are large, the average sample number $\mathbf{E}(T)$ of this SPRT can be approximated by

$$\mathbf{E}(T) \sim \frac{h}{\varrho} \quad (4.3.134)$$

4.4 Formal Definition of Criteria

In this section, we define several criteria for the performance evaluation of change detection algorithms. These criteria were informally introduced in section 1.4. Since the main topic of the book is on-line change detection algorithms, we mainly concentrate on the formal definition of the criterion for on-line situations.

It is of key interest to understand that on-line change detection problems are very close to off-line change detection problems. In many practical cases, it is useful to know that methods other than on-line methods exist which can extend and improve the possibilities of on-line algorithms. For example, in geophysics the accurate estimation of arrival times of seismic waves is an important problem. This problem is typically for off-line algorithms. But, when this problem has to be solved either for very long samples of signals or for samples collected in on-line mode, off-line algorithms are too complex and it is of crucial interest to have preliminary estimation of arrival times before using off-line methods. These first estimates can typically be obtained with the aid of on-line change detection algorithms. Then off-line change time estimation algorithms can be applied inside time windows centered around these first estimates. This necessity comes from the fact that off-line algorithms are difficult to use when many changes occur, as is likely to be the case in long samples; off-line algorithms are also generally very time-consuming.

Another possible connection between on-line and off-line change detection algorithms can be illustrated by the following example. Very often after on-line detection of changes, it is necessary to increase the reliability of the decision by testing between the two hypotheses of "no change occurred" or "a change occurred." This hypotheses testing problem can be efficiently solved off-line inside a time-window which ends at the alarm time given by the on-line algorithm.

4.4.1 On-line Detection of a Change

Let $(Y_k)_{k\geq 1}$ be a random sequence with conditional density $p_\theta(Y_k|\mathcal{Y}_1^{k-1})$. Until the unknown time t_0, the parameter θ is $\theta = \theta_0$ and from t_0 becomes $\theta = \theta_1$. The problem is to detect the change as soon as possible. Several examples of decision rules are introduced in chapter 2. Let t_a be the alarm time at which a detection occurs; t_a is a stopping time, as defined in subsection 3.1.3. As we explain in chapter 1, for estimating the efficiency of the detection, it is convenient to use the mean delay for detection and the mean time between false alarms. In this book, we always consider the mean time between false alarms and the mean time before the first false alarm to be the same. This is not true, for example, when an alarm is followed by inspection and repair. These inspection and repair times are not of interest here and thus are assumed to be zero. Therefore, we assume that after each false alarm the decision function is immediately restarted as at the beginning.

In the subsequent discussion, we follow early investigators [Page, 1954a, Shiryaev, 1961, Kemp, 1961, Lorden, 1971] who introduced these criteria and investigated optimal properties of change detection algorithms. We distinguish two cases of the unknown change time : nonrandom t_0 and random t_0.

4.4.1.1 Nonrandom Change Time t_0

We first investigate the case of simple hypotheses. The definition concerning false alarms is straightforward.

Definition 4.4.1 (Mean time between false alarms). *We define* mean time between false alarms *as the following expectation :*

$$\bar{T} = \mathbf{E}_{\theta_0}(t_a) \tag{4.4.1}$$

Because the delay for detection depends upon the behavior of the process both before and after the change time, the issue of a convenient definition for this delay is more difficult. Let us begin by the definition of the

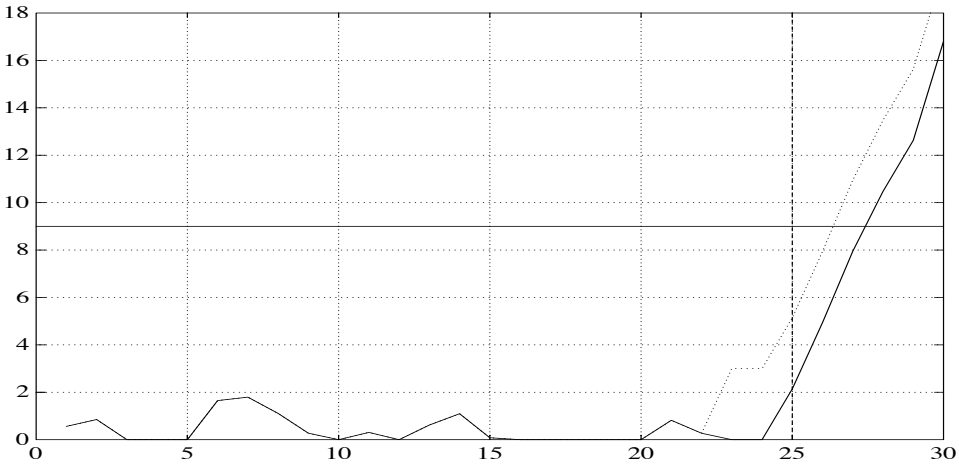

Figure 4.5 Worst mean delay. The smaller the value of the decision function at time $t_0 - 1 = 24$, the greater is the delay.

conditional mean delay for detection, which takes into account the trajectory of the observed process before the change time.

Definition 4.4.2 (Conditional mean delay). *We define* conditional mean delay for detection *as the following expectation :*

$$\mathbf{E}_{\theta_1}(t_a - t_0 + 1 | t_a \geq t_0, \mathcal{Y}_1^{t_0-1}) \tag{4.4.2}$$

The conditional mean delay can be used in several different ways, as we explain now. We emphasize that in this conditional delay, the conditioning is with respect to *two* quantities : the change time t_0 and the sample path of the past observations $\mathcal{Y}_1^{t_0-1}$. Thus, it should be clear that we can define other delays using expectations with respect to the distributions of these two quantities or using supremum. Let us consider first the most pessimistic point of view and define the worst mean delay. In other words, the mean delay that corresponds to the situation where the change time and the sample path of past observations are such that the value of the decision function at the change time is the least favorable with respect to speed of detection; see the solid line leading to value zero in figure 4.5, for example. This worst delay is formally defined as follows.

Definition 4.4.3 (Worst mean delay). *We define* worst mean delay *as the following quantity :*

$$\bar{\tau}^* = \sup_{t_0 \geq 1} \operatorname{ess\,sup} \ \mathbf{E}_{\theta_1}(t_a - t_0 + 1 | t_a \geq t_0, \mathcal{Y}_1^{t_0-1}) \tag{4.4.3}$$

where $\operatorname{ess\,sup}$ *is defined as in section 3.1.*

This criteria was introduced in [Lorden, 1971]. On the right side of this formula two suprema are computed, one for the change time and the other for the behavior of the process before change.

A second and slightly different criterion can be obtained by replacing the second (essential) supremum by the expectation with respect to the distribution before the change time :

$$\bar{\tau}^0 = \sup_{t_0 \geq 1} \mathbf{E}_{\theta_0} \left[\mathbf{E}_{\theta_1}(t_a - t_0 + 1 | t_a \geq t_0, \mathcal{Y}_1^{t_0-1}) \right] \tag{4.4.4}$$

The third use of the conditional mean delay (4.4.2) consists of assuming that the behavior of the process before the change time is nonrandom and arbitrarily fixed, and consequently assuming that the change time t_0 is equal to 1. This point of view results in the following definition.

Definition 4.4.4 (Mean delay). *We define* mean delay *as the following quantity :*

$$\bar{\tau} = \mathbf{E}_{\theta_1}(t_a) \tag{4.4.5}$$

Let us now introduce a particular function that contains all the information related to the performances. As discussed in subsections 4.2.2 and 4.2.4, the power function for hypotheses testing contains the entire information about the statistical properties of the test. In on-line change detection problems, the analog of this function is the *average run length function*, which was introduced in [Aroian and Levene, 1950].

Definition 4.4.5 (ARL function). *We define the* ARL function *as the following function of the parameter* θ *:*

$$L(\theta) = \mathbf{E}_{\theta}(t_a) \tag{4.4.6}$$

Sometimes it is useful to make explicit the dependence upon the starting value z of the decision function using the notation $L_z(\theta)$. Also, as for the power, we sometimes consider simply the ARL L for a given θ. In this case, the dependence upon z is denoted by L_z.

It is obvious from the two equations (4.4.1) and (4.4.5) that the ARL function defines, at θ_0, the mean time between false alarms, and at θ_1 the mean delay for detection, as depicted in figure 4.6.

Actually, the ARL function contains much more information than these two values, and this additional information is useful in practice, because it is related to the behavior of the change detection algorithm for different parameter values before and after the change.

Moreover, for the CUSUM algorithm, which plays a central role in this book, there exists a strong connection between the ARL function and the expectation μ of the increment s_i of the decision function, as depicted in figure 4.6. As we explain at the beginning of chapter 2, the mean value of this increment is negative before the change and positive after the change. *In the next chapters, we often use the fact that, for $\mu < 0$, $L(\mu)$ is the mean time between false alarms $\bar{T}$ and, for $\mu > 0$, $L(\mu)$ is the mean delay for detection $\bar{\tau}$.* Therefore, in many subsequent discussions about the properties of the CUSUM-type algorithms, we often omit the initial parameter θ and consider only the mean value μ of the increment of the decision function.

For the CUSUM and GLR algorithms described in chapter 2, the two definitions of delay (4.4.3) and (4.4.5) are equivalent, as will be discussed in chapter 5. For elementary algorithms, the main criterion used in the literature is the ARL function, and in view of its practical interest, will be the most widely investigated criterion in chapter 5 for these algorithms.

Definition 4.4.6 (Optimal algorithm). *An* optimal on-line algorithm *for change detection is any algorithm that minimizes the mean delay for detection (4.4.2), (4.4.3), or (4.4.5) for a fixed mean time between false alarms $\bar{T}$ (4.4.1).*

We say that an algorithm is asymptotically optimal *when it reaches this optimal property asymptotically when $\bar{T} \to \infty$.*

Up to now, we have assumed that the vector parameters θ_0 and θ_1 correspond to simple hypotheses before and after change. Let us consider now the more difficult case of composite hypotheses, namely when $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$. This situation, in the scalar case, is depicted in figure 4.7. Let K_T be the class of

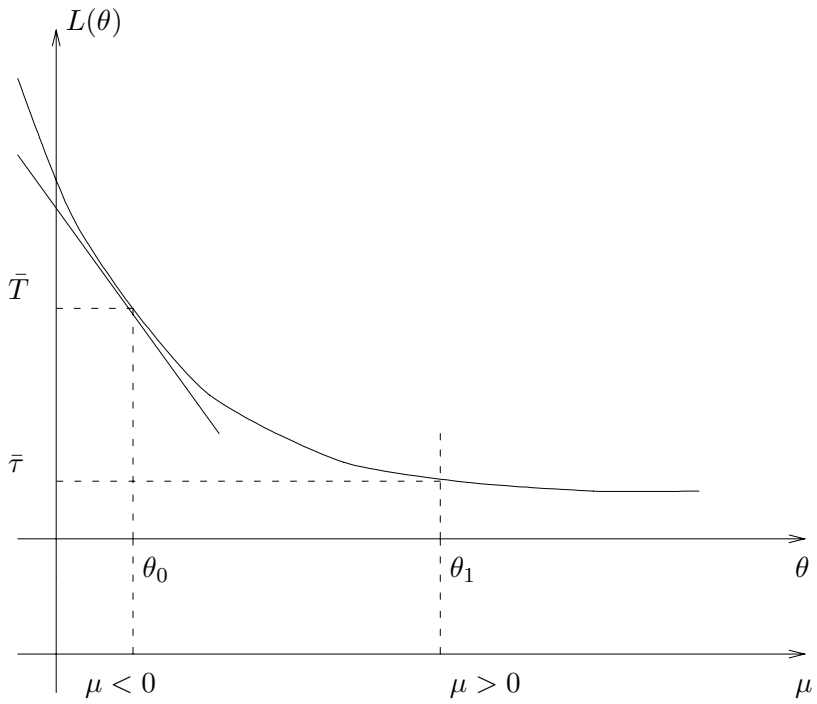

Figure 4.6 The ARL function, as a function of the parameter θ, provides us with the mean time between false alarms $\bar{T}$ and with the mean delay for detection $\bar{\tau}$, and with much other information regarding the robustness of the algorithm. The ARL function can also be viewed as a function of the mean value μ of the increment s of the log-likelihood ratio : for $\mu < 0$, $L(\mu)$ is $\bar{T}$, while for $\mu > 0$, $L(\mu)$ is $\bar{\tau}$. The slope of L provides us with the performance under a local assumption.

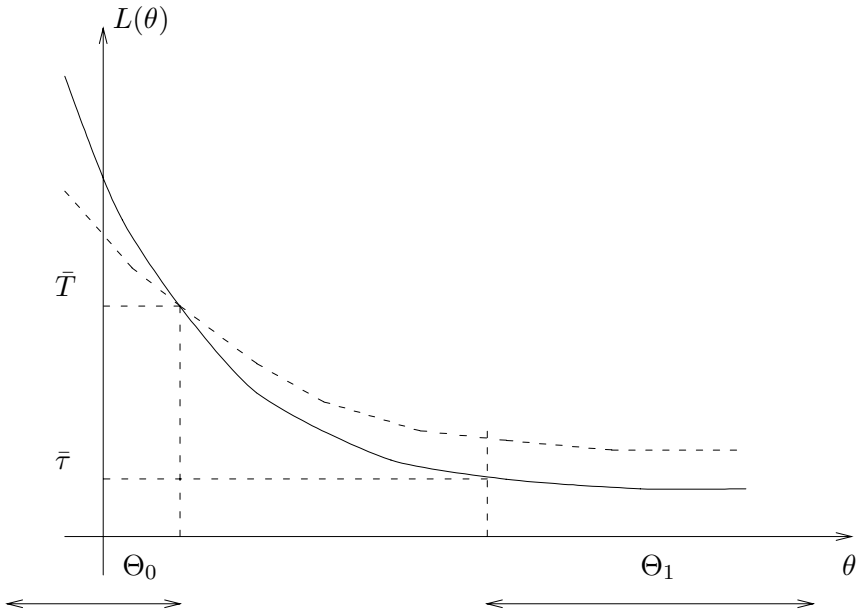

Figure 4.7 The ARL function of a UMP algorithm is uniformly minimum over Θ_1 (solid line); it is compared with another algorithm (dashed line). Both algorithms are unbiased.

algorithms t_a with mean time between false alarms at least equal to $\bar{T}$. A UMP on-line algorithm t_a^* in the class K_T can be defined if, for any other algorithm $t_a \in K_T$, the following inequality holds :

$$\forall \theta \in \Theta_1, \quad \bar{\tau}_{t_a^*}(\theta) \leq \bar{\tau}_{t_a}(\theta) \tag{4.4.7}$$

The property of this ideal UMP on-line algorithm is depicted in figure 4.7, where the ARL function of the UMP algorithm is actually uniformly minimum in the domain Θ_1. But in the present case of composite hypotheses, the proof of the existence of this UMP on-line algorithm has not been obtained yet.

However, as discussed in section 4.2, there exists another useful property of statistical tests, namely the unbiasedness. Since the ARL function is analogous to the power function, it is useful to define the unbiasedness of an on-line change detection algorithm in terms of the ARL function.

Definition 4.4.7 (Unbiased on-line algorithm). *An on-line change detection algorithm is said to be* unbiased *if the following condition holds :*

$$\inf_{\theta \in \Theta_0} L(\theta) \geq \sup_{\theta \in \Theta_1} L(\theta) \tag{4.4.8}$$

The ARL functions depicted in figures 4.6 and 4.7 both correspond to unbiased on-line change detection algorithms. This requirement is intuitively straightforward, because an algorithm with mean time between false alarms *less* than delay for detection is obviously of no interest.

Finally, in the case of *local* hypotheses, another useful criterion for investigating the properties of on-line change detection algorithms is the *slope* $\left.\frac{\partial L(\theta)}{\partial \theta}\right|_{\theta=\theta_0}$ of the ARL function at θ_0, as depicted in figure 4.6. For fixed $L(\theta_0)$, this (negative) slope should be as great as possible, because for local hypotheses the maximization of this slope is equivalent to the minimization of the delay for fixed mean time between false alarms.

4.4.1.2 Random Change Time t_0

In this case, we follow [Shiryaev, 1961, Shiryaev, 1963]. Let us assume that there exists a distribution $\mathbf{P}_\pi$ of the integer random value t_0. Assume that for all $\pi, 0 < \pi < 1$, the distribution $\mathbf{P}_\pi$ has initial probability of change before the first observation

$$\mathbf{P}_\pi(t_0 = 0) = \pi \tag{4.4.9}$$

and is geometric

$$\forall n > 0, \quad \mathbf{P}_\pi(t_0 = n | t_0 > 0) = (1 - \varrho)^{n-1}\varrho \tag{4.4.10}$$

where $0 < \varrho < 1$.

Definition 4.4.8 (Mean delay). *We define a* mean delay *for detecting a change that occurs at a random time t_0 by the following quantity :*

$$\bar{\tau} = \mathbf{E}_\pi(t_a - t_0 + 1 | t_a \geq t_0) \tag{4.4.11}$$

Definition 4.4.9 (Optimal algorithm). *We say that an algorithm is* optimal *if it has minimum delay $\bar{\tau}$ in the class K_α, which is defined by*

$$\mathbf{P}_\pi(t_a < t_0) \leq \alpha \tag{4.4.12}$$

where $0 \leq \alpha < 1$. The probability on the left side of this inequality is the probability of false alarm, and is assumed to be bounded.

Let us compare this criterion and the criterion introduced earlier for nonrandom change time. From a mathematical point of view, the Bayesian criterion is simpler and more formal and does not have the drawback of being heuristic. This superiority holds with respect to two aspects. First, for nonrandom t_0 we cannot use averaging of all possible change situations because we have no distribution. For this reason it is necessary to add some heuristic assumptions about the behavior of the trajectory of the random sequence before time t_0. A typical assumption is to assume the worst case with respect to mean delay, as discussed before.

Second, when the change time is nonrandom, the mean time between false alarms and the mean delay for detection are computed each with respect to a different probability distribution, and the problem arises of how to manage with changes that could have occurred before the first observation. In the case of a random change time, this problem is naturally inserted into the algorithm because of the introduction of the *a priori* probability π of a change before first observation. From a practical point of view the second criterion is not the most useful, because it assumes that the *a priori* distribution of t_0 is known, and little is known about the behavior of the resulting optimal algorithm when the true distribution of t_0 is not exactly as assumed.

[Shiryaev, 1961] described another possible problem statement that is a kind of limit case with respect to the previous one. Let us assume that, before the change time t_0, the detection process reached a steady state after infinitely many observations; under such an assumption, it is possible to compute the steady-state distribution of the decision function at time t_0, which is obviously useful for computing the delay for detection. The criterion is then to minimize the mean delay for detection :

$$\bar{\tau} = \mathbf{E}(t_a - t_0 + 1 | t_a \geq t_0) \tag{4.4.13}$$

for fixed mean time between false alarms :

$$\bar{T} = \mathbf{E}(t_a | t_a < t_0) \tag{4.4.14}$$

4.4.2 Off-line Algorithms

We now introduce the criteria for investigating the properties of off-line change detection algorithms, considering the two problem statements introduced in chapter 1 and discussed in chapter 2.

4.4.2.1 Off-line Hypotheses Testing

This type of change detection problem is a typical testing problem between a simple hypothesis $\mathbf{H}_0$ and a composite hypothesis $\mathbf{H}_1$. In section 4.2, we introduced several optimality criteria for such types of problems. The typical criterion consists of maximizing the power of the test for a fixed level. It should be clear, however, that the present change detection problem statement results in a hypotheses testing problem with a *nuisance* parameter, namely the change time t_0. It is known [Deshayes and Picard, 1986] that no optimal test exists for this problem if no *a priori* information is available about t_0. Moreover, in [Deshayes and Picard, 1986] it is also proved that there exists an optimality property in a nonlocal asymptotic sense, namely the large deviations approach. In this case, the performance index is the exponential rate of convergence to zero of the error probabilities α_0 and α_1. In some cases, it is relevant to use the minimum power $\beta_{\text{min}} = \min_{1 \leq t_0 \leq N} \beta(t_0)$ or the mean power $\bar{\beta} = \sum_{i=1}^{N} \gamma_i \beta(i)$ instead of $\beta(k)$ for some fixed value $t_0 = k$.

4.4.2.2 Off-line Estimation of the Change Time

Several approaches exist for estimating the properties of change detection algorithms using this estimation problem statement. Let us discuss two of them. The first approach consists of using, as the criterion, the probability distribution of the change time estimation error [Jurgutis, 1984, Kligiene, 1980, Bojdecki and Hosza, 1984] :

$$\mathbf{P}(\hat{t}_0 = t_0 \pm n) \tag{4.4.15}$$

for $n = 0, 1, 2, \ldots$, where $\hat{t}_0$ is the estimated change time. Another criterion is

$$\mathbf{P}(|\hat{t}_0 - t_0| \leq n) \tag{4.4.16}$$

for fixed n.

The second approach is common in estimation theory and includes investigations of the properties of the estimate $\hat{t}_0$, such as consistency and efficiency. It follows from [Hinkley, 1970] that the maximum likelihood estimate $\hat{t}_0$ is *not* a consistent estimate of t_0 even in the independent Gaussian case. For this reason, it is useful to discuss the local asymptotic point of view as in [Deshayes and Picard, 1986]. Let us assume that there exists a sequence of change time estimation problems with growing sample size $n \to \infty$ such that $t_0(n) \to \infty$ and $n - t_0(n) \to \infty$, and such that $|\theta_0(n) - \theta_1(n)| \to 0$. Under these conditions, Deshayes-Picard describe the distribution of $\frac{\hat{t}_0(n) - t_0(n)}{n}$ and the consistency of $\frac{\hat{t}_0(n)}{n}$.

4.5 Notes and References

Section 4.1

The main textbooks that are useful for statistical inference and information are [Kullback, 1959, Anderson, 1971, Borovkov, 1984, Gray and Davisson, 1986, Cox and Hinkley, 1986, Blahut, 1987]. A more advanced treatment can be found in [Pinsker, 1964].

Section 4.2

The main textbooks concerning hypotheses testing problems are [Lehmann, 1986, Borovkov, 1984]. For the asymptotic local approach, more specialized books are [Roussas, 1972, Ibragimov and Khasminskii, 1981, Le Cam, 1986] and papers are [Le Cam, 1960, Davies, 1973].

Section 4.3

The main textbooks about sequential analysis are [Wald, 1947, Basharinov and Fleishman, 1962, Ghosh, 1970, Shiryaev, 1978, Borovkov, 1984, Siegmund, 1985b]. Additional specialized papers are [Jackson and Bradley, 1961, W.Hall *et al*., 1965, Robbins, 1970, Y.Chow *et al*., 1971, Berk, 1973, Berk, 1975].

Section 4.4

The ARL function was introduced in [Aroian and Levene, 1950] and has been widely investigated since then [Van Dobben De Bruyn, 1968, Lorden, 1971]. The Bayesian problem statement and criteria were introduced in [Shiryaev, 1961, Shiryaev, 1963]. The non-Bayesian approach and criteria are discussed in [Lorden, 1971].

5

Properties of On-line Algorithms

In this chapter, we investigate the properties of the on-line change detection algorithms described in chapter 2. Recall that these algorithms are devoted to the on-line detection of a change in the scalar parameter θ of an independent random sequence which jumps from θ_0 to θ_1. We first discuss the analytical formulas for investigating the properties of on-line detectors, namely Shewhart control charts, geometric moving average charts, finite moving average charts, CUSUM-type algorithms, GLR detector, and, very briefly, Bayes-type algorithms. Together with these formulas, we also discuss numerical methods for solving some equations that are useful for estimating the properties of change detection algorithms. We compare different algorithms using analytical methods, numerical results, and statistical simulations. We also discuss robustness for some algorithms, with respect to *a priori* information.

Note that the theoretical results we derive here for CUSUM-type and GLR algorithms are used also for investigating the properties of change detection algorithms in part II.

5.1 Elementary Algorithms

In this section, we derive some analytical properties of the on-line algorithms described in section 2.1 and discuss numerical approximations for them. As discussed in section 4.4, we investigate the properties of the elementary algorithms mainly with the aid of the ARL function.

5.1.1 Shewhart Control Charts

A Shewhart control chart is a repeated Neyman-Pearson test applied to samples of fixed size N. In other words, let $S_1^N(K)$ be the log-likelihood ratio corresponding to the Kth such sample. The stopping time of this chart is

$$t_a = NK^* = N \min\{K \geq 1 : d_K = 1\} \tag{5.1.1}$$

where the decision rule d_K is defined by

$$d_K = \begin{cases} 0 & \text{if} \quad S_1^N(K) < \lambda \\ 1 & \text{if} \quad S_1^N(K) \geq \lambda \end{cases} \tag{5.1.2}$$

and

$$S_1^N(K) = \sum_{i=(K-1)N+1}^{KN} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \tag{5.1.3}$$

Let us compute the ARL function of this chart. It is well known that the number of samples K^* has a geometrical distribution $\mathbf{P}(K^* = k) = (1-\alpha_0)^k \alpha_0$ where α_0 is the probability of false alarms of this

Neyman-Pearson test. Therefore, the expectation of K^* is $\frac{1}{\alpha_0}$. Finally, the ARL function of the Shewhart control chart at θ_0 can be written as

$$L(\theta_0) = \mathbf{E}_{\theta_0}(t_a) = \frac{N}{\alpha_0} \tag{5.1.4}$$

Similarly, the ARL function at θ_1 is

$$L(\theta_1) = \mathbf{E}_{\theta_1}(t_a) = \frac{N}{1-\alpha_1} \tag{5.1.5}$$

Finally, the entire ARL function is

$$L(\theta) = \mathbf{E}_{\theta}(t_a) = \frac{N}{\beta(\theta)} \tag{5.1.6}$$

where β is the power function of the Neyman-Pearson test.

Therefore, the ARL function of a Shewhart control chart can be computed directly from the properties of a Neyman-Pearson test which are described in detail in subsection 4.2.2. Moreover, equation (5.1.6) shows that change detection algorithms should satisfy the same requirements as hypotheses testing problems. Actually, the power function must be close to zero before the change, leading to a large mean time between false alarms, and close to one after the change, leading to a small delay for detection. When the hypotheses before and after the change are composite, the formula (5.1.6) for the ARL function is still valid, and in this case the use of the results of subsection 4.2.4 is relevant.

Let us add some comments about the corresponding two-sided change detection problem, namely the change from θ_0 to $\underline{\theta}_1$ or $\bar{\theta}_1$, such that $\underline{\theta}_1 < \theta_0 < \bar{\theta}_1$. As discussed in subsection 4.2.5, for an exponential family of distributions, the optimal test in this situation is

$$d_K = \begin{cases} 0 & \text{if} \quad \lambda_1 < S_1^N(K) < \lambda_2 \\ 1 & \text{if} \quad S_1^N(K) \notin (\lambda_1, \lambda_2) \end{cases} \tag{5.1.7}$$

Other possible solutions to this problem are described in section 4.2. In all cases, the ARL function is computed as in (5.1.6) with the aid of the power function of the statistical test that is used in each case.

Before proceeding to the investigation of the main example in this part, let us further discuss the issue of *criterion* which is used for performance evaluation of a Shewhart chart. In this subsection, we basically discuss the computation of the *ARL function*. For computing the mean delay for detection, this criterion can be used provided that the change time t_0 is nonrandom and equal to a multiple of the sample size N. When this assumption fails to be true, namely when $(K-1)N + 1 \leq t_0 < KN$ (which is practically the most relevant situation), the computation of the mean delay with the aid of the ARL function introduces an error. The source of this error lies in the nonstationary distribution of the observations in the sample in which the change occurs. On the other hand, for computing the worst mean delay in this case, the conditional mean delay with respect to the unknown change time t_0 has to be computed and maximized with respect to t_0 and to $y_{(K-1)N+1}, \ldots, y_{t_0-1}$.

Example 5.1.1 (Change in mean - contd.). *Let us now apply these results to the case of a change, from μ_0 to $\mu_1 > \mu_0$, in the mean of an independent Gaussian sequence with variance σ^2. As shown in section 2.1, in this case for each sample the alarm is set the first time at which*

$$\bar{y}(K) - \mu_0 \geq \kappa \sigma_{\bar{y}} = \kappa \frac{\sigma}{\sqrt{N}} \tag{5.1.8}$$

Thus, the false alarm probability is

$$\alpha_0 = \mathbf{P}_{\mu_0}[\bar{y}(K) - \mu_0 \geq \kappa \sigma_{\bar{y}}] = 1 - \phi(\kappa) \tag{5.1.9}$$

and the ARL function at μ_0 is

$$L(\mu_0) = \frac{N}{1-\phi(\kappa)} \tag{5.1.10}$$

Similarly, we have

$$L(\mu_1) = \frac{N}{1-\phi(\kappa + \frac{\mu_1-\mu_0}{\sigma})} \tag{5.1.11}$$

and finally

$$L(\mu) = \frac{N}{1-\phi(\kappa - \frac{\mu-\mu_0}{\sigma})} \tag{5.1.12}$$

In the two-sided case, namely for a change from μ_0 to either $\mu_1^+ = \mu_0 + \nu$ or $\mu_1^- = \mu_0 - \nu$, the alarm is set when

$$|\bar{y}(K) - \mu_0| \geq \kappa\sigma_{\bar{y}} = \kappa\frac{\sigma}{\sqrt{N}} \tag{5.1.13}$$

and thus the ARL function at μ_0 is

$$L(\mu_0) = \frac{N}{2[1-\phi(\kappa)]} \tag{5.1.14}$$

and, more generally

$$L(\mu) = \frac{N}{1-\phi(\kappa - \frac{\mu-\mu_0}{\sigma}) + \phi(-\kappa - \frac{\mu-\mu_0}{\sigma})} \tag{5.1.15}$$

Let us finally comment upon the choice of the tuning parameters N and κ. The numerical optimization of Shewhart's algorithm with respect to the two criteria for change detection algorithms is discussed in [Page, 1954c], for values of the signal-to-noise ratio $\frac{\mu_1-\mu_0}{\sigma}$ between 0.2 and 1.8. The first criterion is to minimize the delay for detection $\bar{\tau}$ for fixed false alarms rate with respect to the parameters N and κ. The second criterion is to maximize the mean time between false alarms for a fixed delay. In these optimizations, the change time t_0 is taken to be equal to a multiple of the sample size N.

5.1.2 Geometric Moving Average Control Charts

The decision function of a geometric moving average control chart (GMA) (2.1.18) is

$$g_k = (1-\alpha)g_{k-1} + \alpha s_k, \quad \text{with } g_0 = 0 \tag{5.1.16}$$

where s_k is some function of observations, as we explain in chapter 2. The two-sided stopping rule is as usual

$$t_a = \min\{k \ : |g_k| \geq \lambda\} \tag{5.1.17}$$

where λ is a conveniently chosen threshold. Different methods exist for computing the mean delay for detection $\bar{\tau}$ and the mean time between false alarms $\bar{T}$, and more generally, for estimating the ARL function of the GMA algorithm.

A simple and efficient numerical method for computing the ARL function is suggested in [Crowder, 1987], where the derivation of the formula for the ARL function of GMA is very close to the computation of the average sample number $N(z)$ (4.3.52) of the SPRT with symmetric thresholds (absorbing boundaries) $(-\lambda, \lambda)$ and starting point $g_0 = z$. In this subsection, the ARL is considered as a function of the initial value z of the decision function, because θ is fixed. But we continue to use the notation L_z for this quantity. Let $\mathbf{P}(\Omega_1)$ be the probability of the event :

$$\Omega_1 = \{|g_1| \geq \lambda\} \tag{5.1.18}$$

It results from the transformation lemma that the density of $g = (1-\alpha)z + \alpha s$ can be written as

$$f(g) = \frac{f(s)}{\dot{g}} \quad \text{where} \quad \dot{g} = \frac{\partial g}{\partial s} = \alpha \tag{5.1.19}$$

If the first observation y_1 is such that the event Ω_1 occurs and the run length is equal to 1. Otherwise, if $|g_1| < \lambda$, the run of the GMA continues with new starting point $g_1 = (1-\alpha)z + \alpha s_1$ and ARL L_{g_1}. The ARL L_z of the GMA is thus equal to the conveniently weighted sum of these two run lengths :

$$L_z = \mathbf{P}(\Omega_1) \cdot 1 + [1 - \mathbf{P}(\Omega_1)] \cdot \left\{ 1 + \frac{\frac{1}{\alpha} \int_{-\lambda}^{\lambda} L_y \, f_\theta \left[\frac{y-(1-\alpha)z)}{\alpha} \right] dy}{1 - \mathbf{P}(\Omega_1)} \right\} \tag{5.1.20}$$

$$= 1 + \frac{1}{\alpha} \int_{-\lambda}^{\lambda} L_y \, f_\theta \left[\frac{y-(1-\alpha)z}{\alpha} \right] dy \tag{5.1.21}$$

where f_θ *is the density* $f(s)$ *of the increment of the decision function*. This integral equation for L_z is a Fredholm integral equation. The numerical solution of this equation is discussed in section 5.2.

Another numerical method for the approximation of the ARL function is described in [Robinson and Ho, 1978], where this problem was addressed for the first time. The observations $(y_k)_{k\geq 1}$ are assumed to form an independent Gaussian sequence, and the decision function can be rewritten as

$$\tilde{g}_k = (1-\alpha)\tilde{g}_{k-1} + \alpha(y_k - \mu_0), \quad \text{with} \quad \tilde{g}_0 = z \tag{5.1.22}$$

where μ_0 is the mean value before change. If the mean value after change is $\mu_1 = \mu_0 + \nu$, the stopping rule is

$$t_a = \min\{k \ : \tilde{g}_k \geq \lambda\} \tag{5.1.23}$$

If the mean value after change is either $\mu_1^+ = \mu_0 + \nu$ or $\mu_1^- = \mu_0 - \nu$, the stopping rule for this two-sided situation is

$$t_a = \min\{k \ : |\tilde{g}_k| \geq \lambda\} \tag{5.1.24}$$

The computation of the ARL function is based upon the following idea. Let us define the probability

$$p_k = \mathbf{P}_\theta(t_a > k | \tilde{g}_0 = z) \tag{5.1.25}$$

By definition, the ARL function is

$$L_z = \sum_{k=0}^{\infty} p_k \tag{5.1.26}$$

On the other hand, we have

$$p_k = p_{k-1} \, \mathbf{P}_\theta[\tilde{g}_k \in (-\lambda, \lambda) | \tilde{g}_{k-1} \in (-\lambda, \lambda)] \tag{5.1.27}$$

Let

$$p = \lim_{k\to\infty} \mathbf{P}_\theta[\tilde{g}_k \in (-\lambda, \lambda) | \tilde{g}_{k-1} \in (-\lambda, \lambda)] \tag{5.1.28}$$

which can be approximated by

$$\tilde{p} = \mathbf{P}_\theta[\tilde{g}_K \in (-\lambda, \lambda) | \tilde{g}_{K-1} \in (-\lambda, \lambda)] \tag{5.1.29}$$

Then the following approximation holds

$$L_z \approx \sum_{k=0}^{K-1} p_k + p_{K-1}\frac{\tilde{p}}{1-\tilde{p}} \tag{5.1.30}$$

For computing the ARL function, [Robinson and Ho, 1978] assumed that the change arises after an infinitely long period. Therefore, the mean delay is computed with the aid of a weighting of all the possible values of the random variable g_{t_0-1}, as opposed to Lorden's idea of worst mean delay which considers only the worst value of this variable. Tabulations of the ARL function are given for different values of the change magnitude ν, the threshold λ, and the autoregressive coefficient $1-\alpha$ for the one-sided and two-sided GMA.

Analytical approaches to the problem of the investigation of the GMA properties are suggested in [Fishman, 1988, Novikov and Ergashev, 1988]. In [Fishman, 1988], formulas for the mean delay for detection $\bar{\tau}$ and the mean time between false alarms $\bar{T}$ are obtained for a continuous time stochastic process, namely the Ornstein-Uhlenbeck process. The asymptotic optimal value of the parameter α is computed, and it is shown that the GMA's mean delay $\bar{\tau}$ is asymptotically, when $\bar{T}$ goes to infinity, greater than the delay of the optimal Shiryaev's algorithm by 23%. In [Novikov and Ergashev, 1988], martingale techniques are used for deriving asymptotic formulas for $\bar{\tau}$ and $\bar{T}$ for the one-sided and two-sided GMA (5.1.22)-(5.1.24).

5.1.3 Finite Moving Average Charts

A finite moving average control chart (FMA) (2.1.26) decision function is

$$g_k = \sum_{i=0}^{N-1} \gamma_i \ln \frac{p_{\theta_1}(y_{k-i})}{p_{\theta_0}(y_{k-i})} \tag{5.1.31}$$

where γ_i are the weights. The stopping rule is as usual

$$t_a = \min\{k : g_k \geq h\} \tag{5.1.32}$$

where h is a conveniently chosen threshold. The estimation of the ARL function of the FMA algorithm is addressed in [Laï, 1974] and [Böhm and Hackl, 1990]. Suppose that a stationary sequence of independent random variables $(y_k)_{k\geq 1}$ has the density $p_\mu(y)$ with finite second moment, $\mathrm{var}(y) = \sigma^2 < \infty$, where μ is the expectation of the observation y. Let us rewrite the formula (2.1.28) and construct the random sequence $(g_k)_{k\geq 1}$:

$$g_k = \sum_{i=1}^{k} \gamma_{k-i}(y_i - \mu_0), \quad \text{when} \quad k \geq N \tag{5.1.33}$$

where the weights γ_i satisfy

$$\gamma_i = \begin{cases} 0 < \gamma_i < \infty & \text{when} \quad i = 0, \ldots, N-1 \\ 0 & \text{when} \quad i \geq N \end{cases} \tag{5.1.34}$$

Under the assumption of stationarity of the sequence $(y_k)_{k\geq 1}$, the new random sequence $(g_k)_{k\geq N}$ is stationary also with mean

$$\mathbf{E}_\mu(g) = \sum_{i=0}^{N-1} \gamma_i(\mu - \mu_0) \tag{5.1.35}$$

and covariance function

$$R_l^g = \begin{cases} \sigma^2 \sum_{i=0}^{N-l-1} \gamma_i \gamma_{i+l} & \text{when} \quad l = 0, \ldots, N-1 \\ 0 & \text{when} \quad i \geq N \end{cases} \tag{5.1.36}$$

Since the y_k are i.i.d. random variables and the function $g = g(y)$ is a nondecreasing function, then the g_k are *associated random variables* [Esary *et al*., 1967]. Such variables turn out to be useful for proving a key inequality when computing bounds for the ARL function $L(\mu) = \mathbf{E}_\mu(t_a)$ of the FMA algorithm, as we show now. Let p_n be the probability :

$$\begin{aligned} p_0(h) &= 1 \\ p_n(h) &= \mathbf{P}_\mu(g_N < h, \ldots, g_{N+n} < h) = \mathbf{P}_\mu(t_a > n+N) \text{ when } n > 0 \end{aligned} \tag{5.1.37}$$

The ARL is obviously

$$L(\mu) = N + \sum_{i=1}^{\infty} p_i(h) \tag{5.1.38}$$

Unfortunately, no analytical expression of the ARL function is available. For this reason, *upper and lower bounds* for the ARL are given in [Laï, 1974, Böhm and Hackl, 1990] and are derived in the following manner.

Let q_n be the probability that the decision function g_k exceeds the threshold h for the first time at instant $k = n + N$, in other words the probability of the event $\Omega = \{\mathcal{Y}_1^{n+N} : \ t_a = n + N\}$. It is obvious that

$$q_n(h) = p_{n-1}(h) - p_n(h) \tag{5.1.39}$$

and that $q_n(h)$ is a nonincreasing function of n for any h. The probabilities have the following property, which is proved in [Laï, 1974] for the Gaussian case, and then generalized in [Böhm and Hackl, 1990] for any probability distribution $p_\theta(y)$ with finite second moment.

Lemma 5.1.1 (Böhm - Hackl). *Assume that the random sequence $(y_k)_k$ is i.i.d. and let $(g_k)_{k \geq N}$ be defined as in (5.1.33). Then the following inequality holds true for $k \geq N$:*

$$q_k(h) \geq p_{k-N}(h) q_{k-1}(h) \tag{5.1.40}$$

This lemma and other results in [Laï, 1974] lead to the following lower and upper bounds for the ARL function :

$$1 + \frac{q_N(h)}{p_N(h)} \leq L(\mu) \leq N + \frac{q_N(h)}{p_N(h)} \tag{5.1.41}$$

Now, let us conclude this section by adding a comment concerning the filtered derivative algorithms. Even though filtered derivative algorithms are also finite moving average charts, the above-mentioned results cannot be applied to these detectors. Actually, these results assume that the coefficients of the linear combination of observations are all *positive*, which is obviously not the case for filtered derivative algorithms. Moreover, to our knowledge, no similar result exists for these algorithms.

5.2 CUSUM-type Algorithms

In this section, we describe the statistical properties of CUSUM-type algorithms in the independent case. We begin with Lorden's important result about optimal solution of change detection problems. We introduce a class of one-sided tests and investigate the properties of these tests and the connection between this class and CUSUM change detection algorithms. We then discuss approximations and bounds for the ARL function, together with numerical solutions of the integral equations for computing this ARL function. Then we describe an extension of these results to weighted CUSUM algorithms. The extension of these results to the GLR algorithm is discussed in section 5.3.

5.2.1 Optimal Properties and the CUSUM Algorithm

We now describe Lorden's result about optimal solutions of change detection problems. We consider a sequence of independent random variables $(y_k)_k$ with density $p_\theta(y)$. Before the unknown change time t_0, the parameter θ is equal to θ_0, and after the change it is equal to θ_1. In chapter 2, we describe several types of change detection algorithms for solving this problem, among which are CUSUM-type algorithms. As discussed in subsection 2.2.2, CUSUM-type algorithms can be viewed as extended stopping times associated with open-ended SPRT's. We first give some useful results concerning this class of extended stopping times, which can be applied not only to CUSUM-type algorithms (namely CUSUM, weighted CUSUM, ...), but also to other algorithms, such as the GLR test. The key result of [Lorden, 1971] is that the properties of an on-line change detection algorithm can be deduced from the properties of a set of parallel open-ended SPRT. This is formally stated as follows :

Theorem 5.2.1 *Let T be a stopping time with respect to $y_1, y_2, \ldots$ such that*

$$\mathbf{P}_{\theta_0}(T < \infty) \leq \alpha \tag{5.2.1}$$

For $k = 1, 2, \ldots$, let $\tilde{T}_k$ be the stopping time obtained by applying T to $y_k, y_{k+1}, \ldots$ and let $T_k = \tilde{T}_k + k - 1$. Define the extended stopping time by

$$T^* = \min\{T_k | k = 1, 2, \ldots\} \tag{5.2.2}$$

Then T^* is such that

$$\begin{array}{rcl} \mathbf{E}_{\theta_0}(T^*) & \geq & \frac{1}{\alpha} \\ \bar{\mathbf{E}}_{\theta_1}(T^*) & \leq & \mathbf{E}_{\theta_1}(T) \end{array} \tag{5.2.3}$$

for alternative distribution $\mathbf{P}_{\theta_1}$, where

$$\bar{\mathbf{E}}_{\theta_1}(T^*) = \sup_{k \geq 1} \operatorname{ess\,sup} \mathbf{E}_k[(T^* - k + 1)^+ | y_1, \ldots, y_{k-1}] \tag{5.2.4}$$

and where $\mathbf{E}_k$ is the expectation under the distribution of the observations when the change time is k.

Remember that, according to the section 4.4, $\mathbf{E}_{\theta_0}(T^*) = \bar{T}$ is the mean time between false alarms and $\bar{\mathbf{E}}_{\theta_1}(T^*) = \bar{\tau}^*$ is the worst mean delay for detection. Therefore, this theorem states the relation between the lower bound for the mean time between false alarms and the upper bound for the worst mean delay for detection.

Now let us explain the consequence of this theorem when applied to the case where T_k corresponds to an open-ended SPRT with upper threshold h :

$$T_k = \begin{cases} \min\{n \geq k \ : \sum_{i=k}^{n} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \geq h\} \\ \infty \ \text{ if no such } n \text{ exists} \end{cases} \tag{5.2.5}$$

Then the extended stopping time T^* is Page's CUSUM stopping time t_a, as discussed in sections 2.2.2 and 4.3.2 :

$$t_a = T^* = \min\{T_k | k = 1, 2, \ldots\} \tag{5.2.6}$$

In this case, it follows from Wald's identity that when h goes to infinity

$$\mathbf{E}_{\theta_1}(T) \sim \frac{h}{\mathbf{K}(\theta_1, \theta_0)} \tag{5.2.7}$$

where

$$\mathbf{K}(\theta_1, \theta_0) = \mathbf{E}_{\theta_1}\left[\ln \frac{p_{\theta_1}(y)}{p_{\theta_0}(y)}\right] \tag{5.2.8}$$

is the Kullback information. Second, from the Wald's inequality given in the example of subsection 4.3.2, we have

$$\mathbf{P}_{\theta_0}(T < \infty) \leq e^{-h} = \alpha \tag{5.2.9}$$

Thus, using (5.2.7), (5.2.9), and formulas (5.2.3) of the previous theorem, we deduce that

$$\bar{\tau}^* = \bar{\mathbf{E}}_{\theta_1}(T^*) \sim \frac{\ln \mathbf{E}_{\theta_0}(T^*)}{\mathbf{K}(\theta_1, \theta_0)} = \frac{\ln \bar{T}}{\mathbf{K}(\theta_1, \theta_0)} \tag{5.2.10}$$

when h goes to infinity, which gives the basic relation between the delay for detection and the mean time for false alarms for the CUSUM algorithm.

Let us now give the main result of Lorden concerning the optimal solution of change detection problems. Before proceeding, we briefly explain the main lines of the proof. The previous discussion provides us with the asymptotic relation between the mean time between false alarms and the worst mean delay for CUSUM algorithm in (5.2.10). On the other hand, Lorden proved that the infimum of the worst mean delay among a class of extended stopping times is precisely given by this relation. From these two facts results the next theorem.

Theorem 5.2.2 (Lorden). *Let $\{T(\alpha)|0 < \alpha < 1\}$ be a class of open-ended SPRT such that*

$$\mathbf{P}_{\theta_0}[T(\alpha) < \infty] \leq \alpha \tag{5.2.11}$$

and for all real θ_1

$$\mathbf{E}_{\theta_1}[T(\alpha)] \sim \frac{\ln(\alpha^{-1})}{\mathbf{K}(\theta_1, \theta_0)} \tag{5.2.12}$$

For $\gamma > 1$, let $\alpha = \gamma^{-1}$, and let $T^*(\gamma)$ be the associated extended stopping time defined by

$$T^*(\gamma) = \min\{T_k(\alpha)|k = 1, 2, \ldots\} \tag{5.2.13}$$

Then

$$\mathbf{E}_{\theta_0}[T^*(\gamma)] \geq \gamma \tag{5.2.14}$$

and, for all real θ_1, $T^*(\gamma)$ minimizes $\bar{\mathbf{E}}_{\theta_1}[\bar{T}(\gamma)]$ among all stopping times $\bar{T}(\gamma)$ satisfying (5.2.14). Furthermore,

$$\bar{\mathbf{E}}_{\theta_1}[T^*(\gamma)] \sim \frac{\ln \gamma}{\mathbf{K}(\theta_1, \theta_0)} \quad \textit{when} \quad \gamma \to \infty \tag{5.2.15}$$

This theorem shows the optimality of the CUSUM algorithm from an asymptotic point of view, what is often called *first-order optimality* [Pollak, 1987]. More precisely, CUSUM is optimal, with respect to the worst mean delay, when the mean time between false alarms goes to infinity. This asymptotic point of view is convenient in practice because a low rate of false alarms is always desirable. Based upon the same criterion of worst mean delay, another optimality result for CUSUM is proven in [Moustakides, 1986, Ritov, 1990], in a nonasymptotic framework : The CUSUM algorithm minimizes the worst mean delay for all $\bar{T} \geq \bar{T}_0$, where $\bar{T}_0$ is small for most cases of practical interest.

On the other hand, as emphasized before, this theorem gives the infimum of the worst mean delay for a class of stopping times with preassigned rate of false alarms. This result is important not only for the CUSUM algorithm, but also for other types of algorithms, because it makes it possible to compare the

performances of these other tests to the optimal one. In some sense, *equation (5.2.10) plays the same role in the change detection theory as the Cramer-Rao lower bound in estimation theory*.

Next, let us add some comments about the practically important issue of *robustness* for the CUSUM algorithm and its connection with the optimal property. The previous theorem states that the CUSUM algorithm is optimal when it is tuned with the true values of the parameters before and after change. When the algorithm is used in situations where the actual parameter values are different from the preassigned values, this optimal property is *lost*. For this reason, it is of key interest to compute the ARL function for *other* parameter values, which we do in the next subsection.

Finally, let us examine the *detectability* issue. As discussed in chapter 2, the Kullback information can be used as a relevant detectability index. Now, from the equation (5.2.15) of theorem 5.2.2, there exists an intrinsic feature of a given change detection problem in terms of the Kullback information. This theorem states that the properties of *all* optimal algorithms are defined in terms of this information.

5.2.2 The ARL Function of the CUSUM Algorithm

In this subsection, we investigate the ARL function of the CUSUM algorithm in the independent case, for which we describe "exact" ARL, approximations, and bounds. We assume that the stopping time and the decision function can be written in the following particular manner :

$$\begin{aligned} t_a &= \min\{k \ : g_k \geq h\} && (5.2.16)\\ g_k &= (g_{k-1} + s_k)^+ && (5.2.17)\\ g_0 &= z \geq 0 \end{aligned}$$

where $(s_k)_k$ is an i.i.d. sequence with density f_θ. This form of decision function is used throughout this subsection. Note that the weighted CUSUM algorithm cannot be written in that form.

5.2.2.1 "Exact" Function

We first describe an exact computation based upon the solution of some Fredholm integral equations. We begin by explaining how to compute the ARL function $L_z(\theta)$ in the case $z = 0$. This case is appropriate for both the mean time between false alarms and the mean delay for detection. First, it is relevant to use $g_0 = 0$ because we start the detection procedure from normal condition. Therefore, the mean time between false alarms is

$$\bar{T} = L_0(\theta_0) \tag{5.2.18}$$

Second, using the criterion of worst mean delay, the relevant value of the decision function at the change time is $g_{t_0-1} = 0$. Therefore, the convenient criterion is $L_0(\theta_1)$, as shown in the following formula :

$$\bar{\tau}^* = \sup_{t_0 \geq 1} \operatorname{ess\,sup} \ \mathbf{E}_{\theta_1}\left(t_a - t_0 + 1 | t_a \geq t_0, \mathcal{Y}_1^{t_0-1}\right) = L_0(\theta_1) \tag{5.2.19}$$

Remembering that the CUSUM algorithm can be derived with the aid of a SPRT formulation, as done in subsection 2.2.2, we derive Page's formula which links the ARL and the statistical properties of the SPRT with lower threshold $-\epsilon = 0$ and upper threshold h [Page, 1954a]. From now on, we consider a fixed value of θ, and we omit θ for simplicity. From Wald's identity (4.3.37) we have

$$L_0 = \mathbf{E}(T_{0,h} | S_{T_{0,h}} \leq 0, S_0 = 0) \cdot \mathbf{E}(c-1) + \mathbf{E}(T_{0,h} | S_{T_{0,h}} \geq h, S_0 = 0) \cdot 1 \tag{5.2.20}$$

where $\mathbf{E}(T_{-\epsilon,h} | S_{T_{-\epsilon,h}} \leq -\epsilon, S_0 = z)$ is the conditional ASN of one cycle of SPRT when the cumulative sum starting from z reaches the lower threshold $-\epsilon$. This quantity is now denoted by $\mathbf{E}(T_{-\epsilon,h} | S_T \leq -\epsilon, z)$.

In (5.2.20), $\mathbf{E}(c-1)$ is the mean number of cycles before the cycle of final decision, as depicted in figure 2.7. It is obvious that $c-1$ has a geometrical distribution $\mathbf{P}(k) = (1-p)p^k$ for $(k = 0, 1, 2, \ldots)$ where $p = \mathbf{P}(0|0)$ is the probability that the SPRT cumulative sum starting from $z = 0$ reaches the lower threshold $-\epsilon = 0$. Moreover, the probability p is nothing but the OC (see subsection 4.3.2). Thus,

$$\mathbf{E}(c-1) = \frac{1}{1-p} - 1 \tag{5.2.21}$$

and it results from (5.2.20) that

$$\begin{aligned} L_0 &= \frac{\mathbf{E}(T_{0,h}|S_T \leq 0, 0) \cdot p + \mathbf{E}(T_{0,h}|S_T \geq h, 0) \cdot (1-p)}{1-p} \\ &= \frac{\mathbf{E}(T_{0,h}|0)}{1-\mathbf{P}(0|0)} \end{aligned} \tag{5.2.22}$$

It is obvious from formula (5.2.22) and the definitions of $\mathbf{E}(T_{-\epsilon,h}|z)$ and $\mathbf{P}(-\epsilon|z)$ that the ARL is

$$L_0 = \frac{N(0)}{1-\mathbf{P}(0)} \tag{5.2.23}$$

where $N(z) = \mathbf{E}(T_{0,h}|z)$ and $\mathbf{P}(z) = \mathbf{P}(0|z)$.

It is also of interest to get a general formula for the ARL function L_z. Therefore, let $g_0 = z \geq 0$. In this case, the formula for the ARL function can be written as

$$L_z = \mathbf{E}(T_{0,h}|z) + \mathbf{P}(0|z)\ L_0 = N(z) + \mathbf{P}(z)L_0 \tag{5.2.24}$$

which we explain now. When the decision function g_k starts from z, there are two possible situations : either g_k goes down and reaches the lower boundary $-\epsilon = 0$ of the SPRT first, or g_k goes up and reaches the upper boundary h of the SPRT without reaching the lower one. The probabilities of these two cases are, respectively, $\mathbf{P}(0|z)$ and $1 - \mathbf{P}(0|z)$. For computing the ARL function L_z, it is necessary to weight the conditional means with these two probabilities, and thus

$$L_z = [1 - \mathbf{P}(0|z)]\ \mathbf{E}(T_{0,h}|S_T \geq h, z) + \mathbf{P}(0|z)\ \ [\mathbf{E}(T_{0,h}|S_T \geq h, z) + L_0] \tag{5.2.25}$$

From this results (5.2.24). Let us note that when $z = 0$, we recover (5.2.22).

For computing L_z from this equation, we need to compute $\mathbf{P}(z)$ (4.3.41) and $N(z)$ (4.3.42), which are solutions to the following Fredholm integral equations of the second kind :

$$\mathbf{P}(z) = \int_{-\infty}^{-z} f_\theta(x)dx + \int_0^h \mathbf{P}(x) f_\theta(x-z)dx \tag{5.2.26}$$

$$N(z) = 1 + \int_0^h N(x) f_\theta(x-z)dx \tag{5.2.27}$$

for $0 \leq z \leq h$, where f_θ is called the kernel of the integral equation. For solving these equations numerically, two approaches have been proposed :

- Solve the integral equations by the following iterative method [Page, 1954b] :

$$\mathbf{P}_n(z) = \int_{-\infty}^{-z} f_\theta(x)dx + \int_{-0}^h \mathbf{P}_{n-1}(x) f_\theta(x-z)dx \tag{5.2.28}$$

  where $\mathbf{P}_0(z)$ is the initial condition for the recursion and where the second integral on the right side of this equation is replaced by a finite sum. The problem of the convergence of this recursion is addressed in [Kemp, 1967a].

- Replace the integral equations by a system of linear algebraic equations and solve this system with respect to the unknown variables [Goel and Wu, 1971].

Let us describe in more detail the second approach, which consists of replacing the integral equation (5.2.26) by the following system of linear algebraic equations :

$$\left\{\begin{array}{ccccc} \tilde{P}(z_1) & = & F(-z_1) & + & \sum_{k=1}^{m} \varrho_k \tilde{P}(z_k) f_\theta(z_k - z_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \tilde{P}(z_m) & = & F(-z_m) & + & \sum_{k=1}^{m} \varrho_k \tilde{P}(z_k) f_\theta(z_k - z_m) \end{array}\right. \tag{5.2.29}$$

where $\tilde{P}(z_i) \approx \mathbf{P}(z_i)$. When the kernel $f_\theta(z)$ of this integral equation is a continuous function in $[0, h]$, there exists an effective solution to this problem, for example, the method of Gaussian quadrature [Kantorovich and Krilov, 1958].

Example 5.2.1 (ARL for CUSUM in the Gaussian case). *In this example, we follow [Goel and Wu, 1971]. Without loss of generality, we assume that the density of s_k is*

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} \tag{5.2.30}$$

where $\theta = \frac{\mu}{\sigma}$.

Algorithm 5.2.1 *The numerical solution of the Fredholm integral equation proceeds in three steps :*

- **Step 1 :** *Using the method of Gaussian quadrature [Kantorovich and Krilov, 1958], we can approximate the integral on the right side of equation (5.2.26) as*

$$\int_0^h \mathbf{P}(x) \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-z-\theta)^2}{2}} dx \approx \sum_{i=1}^{m} \varrho_i \tilde{P}(z_i) \frac{1}{\sqrt{2\pi}} e^{-\frac{(z_i-z-\theta)^2}{2}}$$

  where the $z_i \in [0, h]$ are the "Gaussian" points (namely the roots of the Legendre polynomial) and the ϱ_i are the weights of the Gaussian quadrature for the interval $[0, h]$. These values can be obtained from tables (or algorithms) giving the roots $\tilde{z}_i$ and the weights $\tilde{\varrho}_i$ in the interval $[0, 1]$. Then simply take z_i equal to $h\tilde{z}_i$ and ϱ_i equal to $h\tilde{\varrho}_i$.

- **Step 2 :** *Replace equation (5.2.26) by the following system of linear equations :*

$$A \cdot \tilde{P} = \tilde{B} \tag{5.2.31}$$

  The matrix $A(m \times m)$ and column vectors $\tilde{P}(m \times 1)$ and $\tilde{B}(m \times 1)$ are defined by

$$\begin{array}{rcl} A & = & (a_{ij}), \quad i, j = 1, \ldots, m; \\ \tilde{P}^T & = & [\tilde{P}(z_1), \ldots, \tilde{P}(z_m)] \\ \tilde{B}^T & = & [\phi(-z_1 - \theta), \ldots, \phi(-z_m - \theta)] \end{array}$$

  where

$$\begin{array}{rcl} a_{ij} & = & -\varrho_j \psi(z_k, z_i) \quad \textit{for} \quad i \neq j \\ a_{ii} & = & 1 - \varrho_i \psi(z_i, z_i) \\ \psi(z_k, z_i) & = & \frac{1}{\sqrt{2\pi}} e^{-\frac{(z_k - z_i - \theta)^2}{2}} \end{array}$$

- **Step 3 :** *Solve the system of m linear equations (5.2.31) with respect to the unknown vector $\tilde{P}$. Finally, for obtaining $\mathbf{P}$, z is taken equal to $z_1 = 0$ and the value $\tilde{P}(z_1)$ substituted for $\mathbf{P}(0)$:*

$$\mathbf{P}(0) \approx \tilde{P}(z_1) \tag{5.2.32}$$

The ASN $N(z)$ is computed in a similar manner, and the ARL is computed with the aid of (5.2.24).

The accuracy of this approximation is numerically evaluated and compared with other methods for computing the ARL in [Goel and Wu, 1971].

Example 5.2.2 (ARL for CUSUM in the case of a $\chi^2(1)$ distribution). *In this example we follow [Kireichikov et al., 1990, Mikhailova et al., 1990]. This problem arises in the case of a change in the variance σ^2, from σ_0^2 to σ_1^2, of an independent Gaussian sequence $(y_k)_{k\geq 1}$ with mean $\mathbf{E}(y_k) = 0$. After obvious simplifications and omission of a positive multiplicative factor, the increment s_k can be written as*

$$s_k = \left(\frac{y_k}{\sigma^*}\right)^2 - 1, \;\; \sigma^* = \frac{\ln \sigma_0^2 - \ln \sigma_1^2}{\sigma_1^{-2} - \sigma_0^{-2}} \tag{5.2.33}$$

Let us define $n_k = \frac{y_k}{\sigma}$, where $\sigma^2 = \mathbf{E}(y_k^2)$. We have $\mathcal{L}(n_k) = \mathcal{N}(0,1)$. In this case, the CUSUM stopping time t_a (5.2.16) and decision function g_k(5.2.17) can be rewritten without loss of generality as

$$\begin{array}{rcl} t_a & = & \min\{k \; : g_k \geq h\theta\} \\ g_k & = & (g_{k-1} + s_k)^+ \\ s_k & = & n_k^2 - \theta; \;\; \theta = \left(\frac{\sigma^*}{\sigma}\right)^2 \end{array} \tag{5.2.34}$$

Therefore, the density of the increment s_k is

$$f_\theta(x) = \begin{cases} \frac{e^{-\frac{x+\theta}{2}}}{\sqrt{2}\Gamma(\frac{1}{2})\sqrt{x+\theta}} & \text{if } \; x+\theta > 0 \\ 0 & \text{if } \; x+\theta \leq 0 \end{cases} \tag{5.2.35}$$

which is the shifted pdf of a $\chi^2(1)$ distribution with one degree of freedom.

The following algorithm is suggested in [Kireichikov et al., 1990] for obtaining the solution to the Fredholm equation.

Algorithm 5.2.2 *The numerical solution of the Fredholm integral equation is more complex in this case, because of the discontinuity of the kernel. The algorithm proceeds in six steps :*

- **Step 1 :** *Suppose that the discontinuity point of the kernel is $z_0 : z_0 \in [z_j, z_{j+1}]$. Let us replace the unknown function $\mathbf{P}(x)$ in this subinterval $[z_j, z_{j+1}]$ by the linear approximation :*

$$\mathbf{P}(z) \approx \tilde{P}(z_j) + \frac{[\tilde{P}(z_{j+1}) - \tilde{P}(z_j)](z - z_j)}{\tilde{n}} \tag{5.2.36}$$

where $\tilde{n} = \frac{\tilde{h}}{m-1}$.

- **Step 2 :** *For smoothing the numerical procedure, use the same type of approximation in the next two subintervals* $[z_{j+1}, z_{j+2}]$ *and* $[z_{j+2}, z_{j+3}]$. *Hence, the integral on the right side of equation (5.2.26) for* $\mathbf{P}(z_k)$ *can be approximated by*

$$\begin{aligned}\int_0^{\tilde{h}} \mathbf{P}(x) f_\theta(x - z_k) dx &\approx \sum_{i=1}^{j-1} \varrho_i f_\theta(z_i - z_k) \tilde{P}(z_i) \\ &+ \int_{z_j}^{z_{j+1}} f_\theta(x - z_k) \left[\tilde{P}(z_j) + \frac{[\tilde{P}(z_{j+1}) - \tilde{P}(z_j)](x - z_j)}{\tilde{n}} \right] dx \\ &+ \int_{z_{j+1}}^{z_{j+2}} \dots dx + \int_{z_{j+2}}^{z_{j+3}} \dots dx + \sum_{i=j+3}^{m} \varrho_i f_\theta(z_i - z_k) \tilde{P}(z_i)\end{aligned}$$

where $\varrho_{j+3} = \varrho_m = \frac{\tilde{h}}{2m-2}$; $\varrho_{j+4} = \dots = \varrho_{m-1} = \frac{\tilde{h}}{m-1}$. *Note here that*

$$\sum_{i=1}^{j-1} \varrho_i f_\theta(z_i - z_k) \tilde{P}(z_i) = 0 \tag{5.2.37}$$

because of the shape of the kernel and of the location of the discontinuity point. Suppose now that the number of the row (in the system of linear algebraic equations) where the discontinuity arises for the first time, is equal to k'.

- **Step 3 :** *Execute Step 1 and Step 2 for rows* $j = k', \dots, m$.
- **Step 4 :** *Apply Step 1 and Step 2 to rows* $k' - 1$ *and* $k' - 2$ *if they exist. In equation* $k' - 2$, *the linear approximation replaces the unknown function* $\mathbf{P}(z)$ *in the subinterval* $[z_1, z_2]$, *and in equation* $k' - 1$, *the linear approximation replaces* $\mathbf{P}(z)$ *in the two subintervals* $[z_1, z_2]$ *and* $[z_2, z_3]$.
- **Step 5 :** *Compute the elements of the vector* $\tilde{B} = (\tilde{b}_i)$ *on the right side of (5.2.31) :*

$$\tilde{b}_i = \int_{-\infty}^{-z_i} \frac{e^{-\frac{x+\theta}{2}}}{\sqrt{2}\,\Gamma(\frac{1}{2})\,\sqrt{x+\theta}} dx \tag{5.2.38}$$

- **Step 6 :** *Solve the system of* m *linear equations (5.2.29) with respect to the unknown vector* $\tilde{P}$. *Finally, for obtaining* $\mathbf{P}$, z *is taken to* $z_1 = 0$ *and the value* $\tilde{P}(z_1)$ *is substituted for* $\mathbf{P}(0)$ *:*

$$\mathbf{P}(0) \approx \tilde{P}(z_1) \tag{5.2.39}$$

The ASN $N(z)$ *is computed in a similar manner, and the ARL is computed with the aid of (5.2.24).*

5.2.2.2 Wald's Approximations

We now derive approximations for the ARL function, using (5.2.22) and Wald's approximation for ASN and OC. To use the theory of sequential analysis, we assume that the increment of the cumulative sum $S_k = \sum_{i=1}^{k} s_i$ satisfies conditions (4.3.55) and (4.3.56), which were formulated in subsection 4.3.2. From (4.3.66) and (4.3.61), we deduce that the Wald's approximations for the ASN $\mathbf{E}_\theta(T_{-\epsilon,h}|0)$ and the OC $\mathbf{P}_\theta(-\epsilon|0)$ are as follows :

$$\tilde{\mathbf{E}}_\theta(T_{-\epsilon,h}|0) = \frac{-\epsilon \tilde{\mathbf{P}}_\theta(-\epsilon|0) + h[1 - \tilde{\mathbf{P}}_\theta(-\epsilon|0)]}{\mathbf{E}_\theta(s_k)} \tag{5.2.40}$$

$$\tilde{\mathbf{P}}_\theta(-\epsilon|0) = \frac{e^{-\omega_0 h} - 1}{e^{-\omega_0 h} - e^{\omega_0 \epsilon}} \tag{5.2.41}$$

where $\omega_0 = \omega_0(\theta)$ is the single nonzero root of the equation :

$$\mathbf{E}_\theta(e^{-\omega_0 s_k}) = 1 \tag{5.2.42}$$

The difficulties when using these approximations are that, in the case of $\epsilon = 0$, which is the relevant case for the CUSUM algorithm (see subsection 4.3.2), the substitution of $\tilde{\mathbf{E}}_\theta(T_{-\epsilon,h}|0)$ and $\tilde{\mathbf{P}}_\theta(-\epsilon|0)$ in (5.2.22) leads to a degenerate ratio. We follow [Reynolds, 1975] for deriving an approximation of the ARL function $L_0(\theta)$ defined by the following limit :

$$\hat{L}_0(\theta) = \lim_{\epsilon \to 0} \frac{\tilde{\mathbf{E}}_\theta(T_{-\epsilon,h}|0)}{1 - \tilde{\mathbf{P}}_\theta(-\epsilon|0)} \tag{5.2.43}$$

After substitution of $\tilde{\mathbf{E}}_\theta(T_{-\epsilon,h})$ and $\tilde{\mathbf{P}}_\theta(-\epsilon|0)$, and simple computations, we get

$$L_0(\theta) \approx \hat{L}_0(\theta) = \frac{1}{\mathbf{E}_\theta(s_k)} \left(h + \frac{e^{-\omega_0 h}}{\omega_0} - \frac{1}{\omega_0} \right) \tag{5.2.44}$$

In this formula, we assume that $\mathbf{E}_\theta(s_k) \neq 0$. Now we choose

$$\theta = \theta^* \text{ such that } \mathbf{E}_{\theta^*}(s_k) = 0 \tag{5.2.45}$$

and we compute the ARL function $L_0(\theta^*)$. In practice, point θ^* is usually a boundary between the hypotheses about θ. In this case, the single root of equation (5.2.42) is equal to zero, and Wald's approximations (4.3.61) and (4.3.66) for the quantities $\mathbf{E}_{\theta^*}(T_{-\epsilon,h}|0)$ and $\mathbf{P}_{\theta^*}(-\epsilon|0)$ can be written as

$$\tilde{\mathbf{E}}_{\theta^*}(T_{-\epsilon,h}|0) = \frac{\tilde{\mathbf{P}}_{\theta^*}(-\epsilon|0)\epsilon^2 + [1 - \tilde{\mathbf{P}}_{\theta^*}(-\epsilon|0)]h^2}{\mathbf{E}_{\theta^*}(s_k^2)} \tag{5.2.46}$$

$$\tilde{\mathbf{P}}_{\theta^*}(-\epsilon|0) = \frac{h}{h + \epsilon} \tag{5.2.47}$$

After substitution of these approximations in the formula (5.2.22), we get

$$L_0(\theta^*) \approx \hat{L}_0(\theta^*) = \frac{h^2}{\mathbf{E}_{\theta^*}(s_k^2)} \tag{5.2.48}$$

Up to now, we have assumed that $z = 0$, but it is interesting to get an approximation of the ARL function $L_z(\theta)$ when $z > 0$. In this case, Wald's approximations (5.2.40) and (5.2.41) of $\mathbf{E}(T_{0,h}|z)$ and $\mathbf{P}(0|z)$ can be rewritten as

$$\tilde{\mathbf{E}}_\theta(T_{0,h}|z) = \frac{-z\tilde{\mathbf{P}}_\theta(0|z) + (h - z)[1 - \tilde{\mathbf{P}}_\theta(0|z)]}{\mathbf{E}_\theta(s_k)} \tag{5.2.49}$$

$$\tilde{\mathbf{P}}_\theta(0|z) = \frac{e^{-\omega_0(h-z)} - 1}{e^{-\omega_0(h-z)} - e^{\omega_0 z}} \tag{5.2.50}$$

After substitution of $\tilde{\mathbf{E}}_\theta(T_{0,h}|z)$ from this equation and of $L_0(\theta)$ from (5.2.44) into (5.2.24), we obtain

$$\hat{L}_z(\theta) = \frac{-z\tilde{\mathbf{P}}_\theta(0|z) + (h - z)[1 - \tilde{\mathbf{P}}_\theta(0|z)]}{\mathbf{E}_\theta(s_k)} + \frac{\tilde{\mathbf{P}}_\theta(0|z)}{\mathbf{E}_\theta(s_k)} \left(h + \frac{e^{-\omega_0 h}}{\omega_0} - \frac{1}{\omega_0} \right) \tag{5.2.51}$$

Finally, the substitution of the previous estimate of $\mathbf{P}_\theta(0|z)$ results in

$$L_z(\theta) \approx \hat{L}_z(\theta) = \frac{1}{\mathbf{E}_\theta(s_k)} \left(h - z + \frac{e^{-\omega_0 h}}{\omega_0} - \frac{e^{-\omega_0 z}}{\omega_0} \right) \tag{5.2.52}$$

Let us now derive the approximation for the ARL function $L_z(\theta)$ at $\theta = \theta^*$. Wald's approximations for $\mathbf{E}_{\theta^*}(T_{0,h}|z)$ and $\mathbf{P}_{\theta^*}(0|z)$ when $g_0 = z$ are as follows :

$$\tilde{\mathbf{E}}_{\theta^*}(T_{0,h}|z) = \frac{\tilde{\mathbf{P}}_{\theta^*}(0|z)z^2 + [1 - \tilde{\mathbf{P}}_{\theta^*}(0|z)](h-z)^2}{\mathbf{E}_{\theta^*}(s_k^2)} \tag{5.2.53}$$

$$\tilde{\mathbf{P}}_{\theta^*}(0|z) = \frac{h-z}{h} \tag{5.2.54}$$

After substitution of these approximations in (5.2.24), we get

$$L_z(\theta^*) \approx \hat{L}_z(\theta^*) = \frac{h^2 - z^2}{\mathbf{E}_{\theta^*}(s_k^2)} \tag{5.2.55}$$

Let us discuss equations (5.2.44) and (5.2.52). We want to show that these approximations are compatible with the worst-case inequality $L_z(\theta) \leq L_0(\theta)$ proved by Lorden and depicted in figure 4.5. In other words, we wish to show that the following inequality $\hat{L}_z(\theta) \leq \hat{L}_0(\theta)$ holds for all θ such that $\mathbf{E}_\theta(s_k) \neq 0$. It results from (5.2.44) and (5.2.52) that

$$\hat{L}_z(\theta) - \hat{L}_0(\theta) = \frac{\omega_0}{\mathbf{E}_\theta(s_k)} \left(-\omega_0 z - e^{-\omega_0 z} + 1\right) \tag{5.2.56}$$

The first term in this product is always positive because $\mathbf{E}_\theta(s_k)$ and $\omega_0 = \omega_0(\theta)$ have the same sign. The second term of the product is always negative because $e^x \geq 1 + x$. Therefore, we proved that, for all $\theta \neq \theta^*, \hat{L}_z(\theta) \leq \hat{L}_0(\theta)$.

Let us discuss now equations (5.2.48) and (5.2.55). They imply that, when $\theta = \theta^*$, we have

$$\hat{L}_z(\theta^*) - \hat{L}_0(\theta^*) = \frac{-z^2}{\mathbf{E}_{\theta^*}(s_k^2)} < 0 \tag{5.2.57}$$

Therefore, for all $\theta, \hat{L}_z(\theta) \leq \hat{L}_0(\theta)$.

5.2.2.3 Siegmund's Approximations

In the previous paragraph, we derived approximations of the ARL function based upon Wald's formulas for ASN and OC. The idea of Wald's approximations is to ignore the excess over the boundary : for example, to replace $\mathbf{E}_\theta(S_T|S_T \leq -\epsilon)$ by $-\epsilon$. As an illustration, the excess over the boundary is especially visible at time 24 in figure 2.7. Unfortunately, these approximations are not very accurate, especially when $\mathbf{E}(s_k) < 0$ and $\sqrt{\text{var}(s_k)}$ have the same order of magnitude as h. In this case, the error in $\mathbf{P}_\theta(-\epsilon|0)$ and $\mathbf{E}_\theta(T|0)$ resulting from the omission of the excesses is significant. Therefore, the issue of numerical accuracy and comparison between different approximations for ARL is of interest and is discussed next.

In this paragraph we explain another approximation which includes the calculation of the excess. These deep and useful results are described in [Siegmund, 1985a, Siegmund, 1985b], where they are called *corrected diffusion approximations*. The general theory of the corrected diffusion approximation is complex, and we derive only heuristic proof of the ARL formula, and refer to these references for exact mathematical results and details. Note that the original proof is based upon more sophisticated and subtle ideas than in the rough description which we give here. The main idea of the approximation is the following. The formula for ARL is the same as before (5.2.44), except that we replace the threshold h by $h + \kappa$, where κ is some positive constant. Let us consider the SPRT with boundaries $-\epsilon$ and h. The ASN (4.3.40) and OC (4.3.59)

can be written as

$$\begin{aligned}
\mathbf{E}_\theta(T_{-\epsilon,h}|0) &= \frac{\mathbf{E}_\theta(S_T|S_T \le -\epsilon)\ \mathbf{P}_\theta(-\epsilon|0) + \mathbf{E}_\theta(S_T|S_T \ge h)\ [1 - \mathbf{P}_\theta(-\epsilon|0)]}{\mathbf{E}_\theta(s_k)} \\
\mathbf{P}_\theta(-\epsilon|0) &= \mathbf{P}_\theta(S_T \le -\epsilon) \\
&= \frac{\mathbf{E}_\theta(e^{-\omega_0 S_T}|S_T \ge h) - 1}{\mathbf{E}_\theta(e^{-\omega_0 S_T}|S_T \ge h) - \mathbf{E}_\theta(e^{-\omega_0 S_T}|S_T \le -\epsilon)}
\end{aligned} \tag{5.2.58}$$

It results from formula (5.2.22) that

$$\begin{aligned}
L_0(\theta) &= \frac{\mathbf{E}_\theta(T_{0,h}|0)}{1 - \mathbf{P}_\theta(S_T \le 0)} \\
&= \frac{\mathbf{E}_\theta(S_T|S_T \le 0)\ \mathbf{P}_\theta(0|0) + \mathbf{E}_\theta(S_T|S_T \ge h)\ [1 - \mathbf{P}_\theta(0|0)]}{\mathbf{E}_\theta(s_k)[1 - \mathbf{P}_\theta(0|0)]} \\
&= \frac{1}{\mathbf{E}_\theta(s_k)} \left[\mathbf{E}_\theta(S_T|S_T \ge h) - \mathbf{E}_\theta(S_T|S_T \le 0) + \frac{\mathbf{E}_\theta(S_T|S_T \le 0)}{1 - \mathbf{P}_\theta(0|0)}\right] \\
&= \frac{1}{\mathbf{E}_\theta(s_k)} \left[h + \mathbf{E}_\theta(S_T - h|S_T - h \ge 0) - \mathbf{E}_\theta(S_T|S_T \le 0) + \frac{\mathbf{E}_\theta(S_T|S_T \le 0)}{1 - \mathbf{P}_\theta(0|0)}\right]
\end{aligned} \tag{5.2.59}$$

Let us compute the fourth term on the right side of the last equality. It is obvious from (5.2.58) that this term can be rewritten as

$$\begin{aligned}
\frac{\mathbf{E}_\theta(S_T|S_T \le 0)}{1 - \mathbf{P}_\theta(0|0)} &= \mathbf{E}_\theta(S_T|S_T \le 0) \frac{\mathbf{E}_\theta(e^{-\omega_0 S_T}|S_T \ge h) - \mathbf{E}_\theta(e^{-\omega_0 S_T}|S_T \le 0)}{1 - \mathbf{E}_\theta(e^{-\omega_0 S_T}|S_T \le 0)} \\
&= \mathbf{E}_\theta(S_T|S_T \le 0) \frac{e^{-\omega_0 h}\mathbf{E}_\theta(e^{-\omega_0 (S_T - h)}|S_T - h \ge 0) - \mathbf{E}_\theta(e^{-\omega_0 S_T}|S_T \le 0)}{1 - \mathbf{E}_\theta(e^{-\omega_0 S_T}|S_T \le 0)}
\end{aligned}$$

Assume now that $\omega_0 \to 0$. In this case, the expansion of the expectation $e^{-\omega_0 S_T}$ can be written as

$$\mathbf{E}_\theta(e^{-\omega_0 S_T}) \approx 1 - \omega_0 \mathbf{E}_\theta(S_T) + \frac{1}{2}\omega_0^2 \mathbf{E}_\theta(S_T^2) + \cdots \tag{5.2.60}$$

From (5.2.60) it results immediately that

$$\begin{aligned}
\frac{\mathbf{E}_\theta(S_T|S_T \le 0)}{1 - \mathbf{P}_\theta(0|0)} \approx \frac{1}{\omega_0} \Big\{ & e^{-\omega_0 h}[1 - \omega_0 \mathbf{E}_\theta(S_T - h|S_T - h \ge 0) + \cdots] \\
& - [1 - \omega_0 \mathbf{E}_\theta(S_T|S_T \le 0) + \cdots]\}
\end{aligned} \tag{5.2.61}$$

Using the approximation $1 - x = e^{-x} + o(x)$ as $x \to 0$, we obtain

$$\begin{aligned}
\frac{\mathbf{E}_\theta(S_T|S_T \le 0)}{1 - \mathbf{P}_\theta(0|0)} &\approx \frac{e^{-\omega_0(h + \varrho_+ - \varrho_-)} - 1}{\omega_0 e^{\omega_0 \varrho_-}} \\
&\approx \frac{e^{-\omega_0(h + \varrho_+ - \varrho_-)} - 1}{\omega_0}[1 + o(\omega_0)]
\end{aligned} \tag{5.2.62}$$

where

$$\begin{aligned}
\varrho_+ &= \mathbf{E}_\theta(S_T - h|S_T - h \ge 0) \\
\varrho_- &= \mathbf{E}_\theta(S_T|S_T \le 0)
\end{aligned} \tag{5.2.63}$$

are the expectations of the excesses over the boundaries h and 0. Inserting (5.2.63) and (5.2.62) in (5.2.59), we get the following ARL approximation :

$$L_0(\theta) \approx \frac{1}{\mathbf{E}_\theta(s_k)} \left(h + \varrho_+ - \varrho_- + \frac{e^{-\omega_0(h+\varrho_+-\varrho_-)}}{\omega_0} - \frac{1}{\omega_0} \right) \tag{5.2.64}$$

The direct comparison between (5.2.22) and (5.2.64) shows that *this new approximation has the same form as Wald's approximation with h replaced by $h + \varrho_+ - \varrho_-$.*

The main result of corrected diffusion approximation theory is the formula for calculation of the excesses over the boundaries ϱ_+ and ϱ_-. For the Gaussian case $p(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$, Siegmund has shown that $\varrho_+ - \varrho_- \approx 2\rho$, where

$$\rho = -\pi^{-1} \int_0^\infty x^{-2} \ln \left[\frac{2}{x^2} (1 - e^{-\frac{1}{2}x^2}) \right] \, dx \;\approx\; 0.583 \tag{5.2.65}$$

5.2.2.4 Bounds

In the previous paragraphs, we derived several approximations of the ARL function. Bounds for the ARL function are also desirable for two reasons. First, in practice, it is highly desirable to have reliable estimates of the statistical properties of change detection algorithms, and it turns out that the precision of the previous approximations may be not sufficient in some cases. Second, sometimes it is of interest to prove that one algorithm is better than another one for solving a particular change detection problem. In many cases, the use of bounds for the properties is sufficient for this purpose. Of course, relevant bounds are an upper bound for the mean delay for detection $\bar{\tau}^*$ and a lower bound for the mean time between false alarms $\bar{T}$.

Upper bound for the delay Let us derive an upper bound for the delay $\bar{\tau}^* = L_0(\theta_1)$ when $\mathbf{E}_{\theta_1}(s_k) > 0$. For simplicity, *we omit the index 1 for θ*. Note that we have $\omega_0 = \omega_0(\theta) > 0$. From (5.2.22) and (5.2.58) it results that

$$L_0(\theta) = \frac{\mathbf{E}_\theta(S_T | S_T \le -\epsilon)}{\mathbf{E}_\theta(s_k)} \, \frac{\mathbf{P}_\theta(-\epsilon|0)}{1 - \mathbf{P}_\theta(-\epsilon|0)} + \frac{\mathbf{E}_\theta(S_T | S_T \ge h)}{\mathbf{E}_\theta(s_k)} \tag{5.2.66}$$

Let us consider the first term on the right side of this equation. It is obvious that the value $\mathbf{E}_\theta(S_T | S_T \le -\epsilon)$ is not positive for all $\epsilon \ge 0$ at least. Thus, this first term is bounded from above by zero. Let us now derive an upper bound for the second term in (5.2.66). It is obvious from Wald's formula (4.3.74) for the bound of the ASN that

$$\frac{\mathbf{E}_\theta(S_T | S_T \ge h)}{\mathbf{E}_\theta(s_k)} \le \frac{h + \gamma(\theta)}{\mathbf{E}_\theta(s_k)} \tag{5.2.67}$$

where

$$\gamma(\theta) = \sup_{\lambda > 0} \, \mathbf{E}_\theta(s_k - \lambda | s_k \ge \lambda > 0) \tag{5.2.68}$$

is an upper bound for the expectation of the excess over the boundary.

Finally, we get the upper bound for $\bar{\tau}^*$:

$$\bar{\tau}^* \le \bar{L}_0(\theta) = \frac{1}{\mathbf{E}_\theta(s_k)} \left[h + \gamma(\theta) \right] \tag{5.2.69}$$

Let us add some comments about the computation of $\gamma(\theta)$. In general, the computation of this value is a difficult problem. In [Lorden, 1970] a very simple upper bound for the case $\mathbf{E}_\theta(s_k) > 0$ is suggested. If $\ln[p_\theta(s)]$, where $p_\theta(s)$ is the pdf of s_k, and is a continuous and convex function of s for some $\theta \in \Theta$, then $\gamma(\theta) = \mathbf{E}_\theta(s_k | s_k > 0)$. For example, this situation holds for the case of a change in the mean of a Gaussian random sequence and for some other distributions.

Lower bound for the mean time between false alarms As will become obvious from the following discussion, the estimation of the mean time between false alarms is the major problem as far as the ARL function is concerned. The key reason for this fact comes from the issue of excess over the boundary. When the drift of the CUSUM decision function is negative, the excess over the lower threshold 0 is *not* negligible, and thus the estimation of the mean time between false alarms is difficult. On the contrary, when the drift is positive, the excess over the lower boundary is negligible, and the estimation of the mean delay is less difficult.

We investigate two possible bounds for the mean time between false alarms. These bounds are compared numerically in section 5.5.

Bound 1 Let us now describe how to obtain a lower bound for the mean time between false alarms $\bar{T} = L_0(\theta_0)$ when $\mathbf{E}_{\theta_0}(s_k < 0)$. For simplicity, *we omit the index* 0 *for* θ. In this case, $\omega_0 = \omega_0(\theta) < 0$. Let us again start with equation (5.2.66). We can derive a lower bound of the first term on the right side of (5.2.66), which we call a. Using $\mathbf{P}_\theta(-\epsilon|0)$ given in (5.2.58), we get

$$a = \frac{\mathbf{E}_\theta(S_T|S_T \leq -\epsilon)}{\mathbf{E}_\theta(s_k)} \left[\frac{\mathbf{E}_\theta(e^{-\omega_0 S_T}|S_T \geq h) - 1}{1 - \mathbf{E}_\theta(e^{-\omega_0 S_T}|S_T \leq -\epsilon)} \right] \tag{5.2.70}$$

From the Jensen inequality,

$$1 - \mathbf{E}_\theta(e^{-\omega_0 S_T}|S_T \leq -\epsilon) \leq 1 - e^{\mathbf{E}_\theta(-\omega_0 S_T|S_T \leq -\epsilon)} \tag{5.2.71}$$

Therefore, a lower bound for a is given by

$$a \geq \frac{\mathbf{E}_\theta(S_T|S_T \leq -\epsilon)}{\mathbf{E}_\theta(s_k)} \left[\frac{\mathbf{E}_\theta(e^{-\omega_0 S_T}|S_T \geq h) - 1}{1 - e^{\mathbf{E}_\theta(-\omega_0 S_T|S_T \leq -\epsilon)}} \right] \tag{5.2.72}$$

Now, using the inequality $-x \geq 1 - e^x$, and the fact that the ratio in the square brackets of the previous formula is positive, we get

$$a \geq \frac{\mathbf{E}_\theta(S_T|S_T \leq -\epsilon)}{\mathbf{E}_\theta(s_k)} \left[\frac{\mathbf{E}_\theta(e^{-\omega_0 S_T}|S_T \geq h) - 1}{\omega_0 \mathbf{E}_\theta(S_T|S_T \leq -\epsilon)} \right] \tag{5.2.73}$$

And remembering that $\omega_0 < 0$ implies the inequality

$$\mathbf{E}_\theta(e^{-\omega_0 S_T}|S_T \geq h) \geq e^{-\omega_0 h} \tag{5.2.74}$$

we obtain

$$a \geq \frac{e^{-\omega_0 h} - 1}{\mathbf{E}_\theta(s_k)\omega_0} \tag{5.2.75}$$

On the other hand, a lower bound for the (negative) second term of (5.2.66) is

$$\frac{\mathbf{E}_\theta(S_T|S_T \geq h)}{\mathbf{E}_\theta(s_k)} \geq \frac{h + \gamma(\theta)}{\mathbf{E}_\theta(s_k)} \tag{5.2.76}$$

Therefore,

$$\bar{T} \geq \underline{L}_0(\theta) = \frac{1}{\mathbf{E}_\theta(s_k)} \left[\frac{e^{-\omega_0 h} - 1}{\omega_0} + h + \gamma(\theta) \right] \tag{5.2.77}$$

Bound 2 Another possible lower bound for mean time between false alarms $\bar{T} = L_0(\theta_0)$ can be obtained using the following idea. Let us rewrite formula (5.2.22) in the following manner :

$$L_0(\theta) = \frac{\mathbf{E}_\theta(T_{0,h}|0)}{\mathbf{P}_\theta(S_T \geq h)} \tag{5.2.78}$$

Note that this formula holds true because we use a *closed* SPRT. It is obvious from (4.3.70) that

$$L_0(\theta) \geq \frac{1}{\mathbf{P}_\theta(S_T \geq h)} \geq \frac{e^{-\omega_0 h} - \eta(\theta)}{1 - \eta(\theta)} \geq e^{-\omega_0 h} \tag{5.2.79}$$

The additional motivation for the derivation of this second bound lies in the following inequality :

$$L_0(\theta_0) \geq \frac{1}{\mathbf{P}_{\theta_0}(S_T \geq h)} \geq e^h \tag{5.2.80}$$

which holds under the assumption that s_n is the logarithm of the likelihood ratio (and *not* any i.i.d. sequence as we assume in (5.2.17)). This inequality for $L_0(\theta_0)$ is widely used for the investigation of the properties of CUSUM-type and GLR algorithms in [Lorden, 1971, Lorden, 1973, Pollak and Siegmund, 1975]. For this reason, in section 5.5, we compare the two bounds (5.2.77) and (5.2.80) to the "exact" value of the mean time between false alarms computed with the aid of the algorithm described in the example 5.2.1.

Bound for $L(\theta^*)$ Let us now investigate the ARL function $L_0(\theta)$ when θ goes to the value θ^*, which is such that $\mathbf{E}_{\theta^*}(s_k) = 0$. From formulas (5.2.69) and (5.2.77), it is obvious that the precision of our bounds when $\theta \to \theta^*$ is infinite. For this reason, let us compute directly the lower bound for $L_0(\theta^*)$. Using the lower bound (4.3.75) for $\mathbf{E}_{\theta^*}(T_{0,h})$ and formula (5.2.22), we get

$$\begin{aligned} L_0(\theta^*) &= \frac{\mathbf{E}_{\theta^*}(T_{0,h})}{1 - \mathbf{P}_{\theta^*}(-\epsilon|0)} \\ &= \frac{\mathbf{E}_{\theta^*}(S_T^2|S_T \leq -\epsilon)\mathbf{P}_{\theta^*}(-\epsilon|0) + \mathbf{E}_{\theta^*}(S_T^2|S_T \geq h)\,[1 - \mathbf{P}_{\theta^*}(-\epsilon|0)]}{\mathbf{E}_{\theta^*}(s_k^2)(1 - \mathbf{P}_{\theta^*}(-\epsilon|0))} \\ &= \frac{\mathbf{E}_{\theta^*}(S_k^2|S_k \leq -\epsilon)\mathbf{P}_{\theta^*}(-\epsilon|0)}{\mathbf{E}_{\theta^*}(s_k^2)\,[1 - \mathbf{P}_{\theta^*}(-\epsilon|0)]} + \frac{\mathbf{E}_{\theta^*}(S_k^2|S_k \geq h)}{\mathbf{E}_{\theta^*}(s_k^2)} \\ &\geq \underline{\mathrm{L}}_0(\theta^*) = \frac{h^2}{\mathbf{E}_{\theta^*}(s_k^2)} \end{aligned} \tag{5.2.81}$$

This bound $\underline{\mathrm{L}}_0(\theta^*)$ is very rough. In some cases, for particular pdf of s_k, it is possible to obtain a more accurate bound. Note here that this bound is the same as the Wald's approximation (5.2.48), but continues to be the lower bound for Siegmund's approximation. Therefore, we conclude that the Wald's approximation underestimates the ARL $L_0(\theta^*)$.

5.2.2.5 Two-sided CUSUM Algorithm

We now compute the ARL function of the two-sided CUSUM algorithm (2.2.24) and we follow [Van Dobben De Bruyn, 1968, Khan, 1981, Yashchin, 1985a, Siegmund, 1985b, Wetherill and Brown, 1991]. Under general conditions, the ARL function of the two-sided CUSUM algorithm can be computed from the ARL functions of the two one-sided CUSUM algorithms in the following manner :

$$\frac{1}{L^T(\theta)} \geq \frac{1}{L^l(\theta)} + \frac{1}{L^u(\theta)} \tag{5.2.82}$$

where

L^T is the ARL function of the two-sided CUSUM;

L^l is the ARL function of the one-sided CUSUM corresponding to (θ_0, θ_1^-);

L^u is the ARL function of the one-sided CUSUM corresponding to (θ_0, θ_1^+)

and $\theta_1^- < \theta_0 < \theta_1^+$. In the case (2.2.24) of a change in the mean μ of a Gaussian sequence, the previous inequality becomes an *equality* :

$$\frac{1}{L^T(\mu)} = \frac{1}{L^l(\mu)} + \frac{1}{L^u(\mu)} \tag{5.2.83}$$

Let us give a sketch of the proof of this equality. The interested reader is referred to [Siegmund, 1985b, Wetherill and Brown, 1991] for the details. The stopping time in (2.2.24) can be written as

$$t_a = \min(t_l, t_u) \tag{5.2.84}$$

where

$$\begin{aligned} t_l &= \min\{k \ : g_k^- \geq h\} \\ t_u &= \min\{k \ : g_k^+ \geq h\} \end{aligned} \tag{5.2.85}$$

We now fix one μ. The ARL function L^l can be computed as

$$\begin{aligned} L^l &= \mathbf{E}(t_l) \\ &= \mathbf{E}(t_a) + \mathbf{E}(t_l - t_a) \\ &= \mathbf{E}(t_a) + 0 \cdot \mathbf{P}(t_l = t_a) + \mathbf{E}(t_l - t_a | t_l - t_a > 0) \cdot \mathbf{P}(t_l > t_a) \end{aligned} \tag{5.2.86}$$

It is intuitively obvious, and can be formally proven [Siegmund, 1985b], that, if for one k, we have $g_k^+ \geq h$, this implies that $g_k^- = 0$ for the same k. Thus,

$$\begin{aligned} L^l &= L^T + L^l \cdot \mathbf{P}(t_l > t_a) \\ \text{or } L^T &= L^l \cdot \mathbf{P}(t_l = t_a) \end{aligned} \tag{5.2.87}$$

A similar result holds for L^u :

$$L^T = L^u \cdot \mathbf{P}(t_u = t_a) \tag{5.2.88}$$

Moreover, using $\mathbf{P}(t_u = t_a) + \mathbf{P}(t_l = t_a) = 1$ and the previous relations, we deduce (5.2.83).

From (5.2.83) and Wald's or Siegmund's approximations of the ARL function of the one-sided CUSUM algorithm, two approximations of the ARL function of the two-sided CUSUM algorithm can be computed, but it turns out that they are complex in general. Thus, we give here only the formula corresponding to the special case of a zero minimum magnitude of change in the mean of a Gaussian sequence, for which we use the decision function (2.2.26). It is interesting to note that, in [Nadler and Robbins, 1971], a Brownian motion approximation of this decision function is used and leads to the same formula for the ARL function, namely

$$\begin{aligned} L^T(\mu_s) &= \left(\frac{h}{\mu_s}\right) \coth\left(\frac{\mu_s h}{\sigma_s^2}\right) - \frac{\sigma_s^2}{2\mu_s^2} - \frac{h^2}{2\sigma_s^2 \sinh^2\left(\frac{\mu_s h}{\sigma_s^2}\right)} \quad \text{when} \quad \mu_s \neq 0 \\ L^T(\mu_s) &= \frac{h^2}{2\sigma_s^2} \qquad \text{when} \quad \mu_s = 0 \end{aligned} \tag{5.2.89}$$

where μ_s and σ_s are the mean and standard deviation of the increment of the cumulative sum.

5.2.3 Properties of CUSUM-type Algorithms

We now describe an extension of Lorden's results to the case of weighted CUSUM algorithms. As explained in chapter 2, these algorithms are Bayesian with respect to the *a priori* information about the parameter after change. Let us start from the definition of the stopping time T_k corresponding to the one-sided weighted SPRT in the case of an exponential family of distributions :

$$p_\theta(y) = h(y)e^{\theta y - d(\theta)} \tag{5.2.90}$$

The SPRT stopping time is

$$T = \begin{cases} \min\{n \geq 1 : \Lambda_1^n \geq e^h\} \\ \infty \text{ if no such } n \text{ exists} \end{cases} \tag{5.2.91}$$

where

$$\Lambda_k^n = \int_{-\infty}^{\infty} e^{(\theta_1-\theta_0)S_k^n - (n-k+1)[d(\theta_1)-d(\theta_0)]} dF(\theta_1) \tag{5.2.92}$$

and where

$$S_k^n = \sum_{i=k}^{n} y_i \tag{5.2.93}$$

As before, we define the stopping time T_k as the stopping time T when applied to the observations $y_k, y_{k+1}, \ldots$. The extended stopping time t_a corresponding to the weighted CUSUM is again $t_a = \min_k(T_k + k - 1)$ and can be written as

$$t_a = \min\{n \geq 1 : \max_{1 \leq k \leq n} \Lambda_k^n \geq e^h\} \tag{5.2.94}$$

In this case, the Kullback information is given by (4.1.55) :

$$\mathbf{K}(\theta_1, \theta_0) = d(\theta_0) - d(\theta_1) + (\theta_1 - \theta_0)\dot{d}(\theta_1) \tag{5.2.95}$$

Because the weighted likelihood ratio is a likelihood ratio also, it results from Wald's inequality again that

$$\mathbf{P}_{\theta_0}(T_k < \infty) \leq e^{-h} \tag{5.2.96}$$

From this and theorem 5.2.1 we get the mean time $\bar{T}$ between false alarms for the weighted CUSUM algorithm :

$$\bar{T} = \mathbf{E}_{\theta_0}(t_a) \geq e^h \tag{5.2.97}$$

Consider now the computation of the worst mean delay. For an open-ended weighted SPRT, the following theorem is proven in [Pollak and Siegmund, 1975] :

Theorem 5.2.3 *Assume that the weighting function F has a positive and continuous derivative in the neighborhood of θ_1. Then, when h goes to infinity, the ASN (or mean number of observations before end) has the following approximation :*

$$\mathbf{E}_{\theta_1}(T) \approx \frac{h + \frac{1}{2}\ln\left[\frac{h}{\mathbf{K}(\theta_1,\theta_0)}\right]}{\mathbf{K}(\theta_1, \theta_0)} - G[\theta_1, \theta_0, F(\theta_1)] + o(1) \tag{5.2.98}$$

where

$$G[\theta_1, \theta_0, F(\theta_1)] = \frac{1}{2\ \mathbf{K}(\theta_1, \theta_0)} \left\{ \ln\left[2\pi\ \frac{\dot{F}^2(\theta_1)}{\ddot{d}(\theta_1)}\right] - 1 \right\} \tag{5.2.99}$$

It results from theorem 5.2.1 that this mean number of observations can be used as an upper bound for the worst mean delay for detection. Therefore, we get the following approximation for the worst mean delay of the weighted CUSUM algorithm :

$$\bar{\tau}^*(\theta_1) \approx \frac{\ln \bar{T} + \frac{1}{2} \ln \left[\frac{\ln \bar{T}}{\mathbf{K}(\theta_1,\theta_0)} \right]}{\mathbf{K}(\theta_1,\theta_0)} - G[\theta_1, \theta_0, F(\theta_1)] + o(1) \tag{5.2.100}$$

In [Pollak and Siegmund, 1975], a more precise approximation is given, with an additional term that can be shown to have the same order of magnitude as $-1/\mathbf{K}(\theta_1, \theta_0)$.

Let us compare the mean delays of the weighted CUSUM and CUSUM algorithms. It is obvious that the mean delay for the weighted CUSUM does not reach the infimum of mean delay for the class of open-ended tests given in theorem 5.2.2, because of the main additional term when $\bar{T}$ goes to infinity :

$$\frac{1}{2\ \mathbf{K}(\theta_1,\theta_0)} \ln \left[\frac{\ln \bar{T}}{\mathbf{K}(\theta_1,\theta_0)} \right] \tag{5.2.101}$$

In some sense, this additional term is the price to be paid for the unknown *a priori* information. It can be shown that the next term G, which is constant with respect to h, can compensate this main additional term when the shape of the *a priori* distribution F for θ_1 is closed to a Dirac function.

5.2.4 χ^2-CUSUM Algorithms

We now briefly investigate the properties of another weighted CUSUM for which the regularity condition of the previous theorem does not apply. This algorithm has been introduced in the subsection 2.4.2 under the name of χ^2-CUSUM and corresponds to the case of a degenerated distribution for θ_1, which is concentrated on two values, $\theta_0 \pm \nu$. This algorithm is devoted to the case of a change in the mean of a Gaussian sequence where we can assume a variance equal to 1 without loss of generality.

In this case, the weighted likelihood ratio is

$$\tilde{\Lambda}_k^n = e^{-\frac{1}{2}(n-k+1)\nu^2} \cosh[\nu(n-k+1)\chi_k^n] \tag{5.2.102}$$

where

$$\chi_k^n = \frac{1}{n-k+1} |\tilde{S}_k^n|, \quad \tilde{S}_k^n = \sum_{i=k}^{n} (y_i - \mu_0) \tag{5.2.103}$$

A more extensive investigation of the properties of multidimensional χ^2-CUSUM algorithms (which include the scalar case) is described in section 7.3. For this algorithm, the result of theorem 5.2.1 can be applied for computing the mean time between false alarms. For computing the mean delay for detection, two approaches are possible. The first consists of using the theorem of [Berk, 1973] for invariant SPRT, and remembering that χ^2-SPRT is a special case of invariant SPRT [Ghosh, 1970]. This theorem is used in section 7.3 for a multidimensional parameter. The second approach consists of considering the stopping rule of an open-ended test as a stopping time associated with a random walk that crosses an almost linear boundary, namely the V-mask of the two-sided CUSUM algorithm, which we introduced in subsection 2.4.2. Let us explain this now. The stopping time (5.2.91) can be rewritten as

$$T = \left\{ \begin{array}{c} \min\{n \geq 1 : |\tilde{S}_1^n| \geq c_n\} \\ \infty \text{ if no such } n \text{ exists} \end{array} \right. \tag{5.2.104}$$

where $|c_n|$ is the positive solution of the equation $\ln \tilde{\Lambda}_1^n(|x|) = h$ and where $\tilde{\Lambda}_1^n = \tilde{\Lambda}_1^n(\tilde{S}_1^n)$. It results from the discussion in subsection 2.4.2 that the asymptotic stopping boundary has the equation

$$|c_n| \approx \frac{h + \ln 2}{\nu} + \frac{n\nu}{2} \tag{5.2.105}$$

which is nothing but the stopping boundary of the V-mask. As we discussed in subsection 2.4.2, the key difference between the asymptotes and the actual boundaries is negligible. Then the Wald's identity for the ASN of this one-sided SPRT with threshold $h + \ln 2$ gives an approximation to the mean delay for the χ^2-CUSUM :

$$\bar{\tau}^* \leq \mathbf{E}_{\theta_1}(T) \sim \frac{h + \ln 2}{\mathbf{K}(\mu_1, \mu_0)} \sim \frac{\ln \bar{T} + \ln 2}{\mathbf{K}(\mu_1, \mu_0)} \tag{5.2.106}$$

Now let us compare the mean delays of the χ^2-CUSUM and the CUSUM algorithms. From the last formula, it is obvious that for χ^2-CUSUM there arises an additional term $\ln 2/\mathbf{K}(\mu_1, \mu_0)$. This is again a consequence of the lack of *a priori* information, but it is obvious that this term is asymptotically negligible. Formula (5.2.106) provides us with an asymptotic relation between the delay for detection and the mean time between false alarms. But, from a practical point of view, it is more important to know that the stopping boundaries of the χ^2-CUSUM and two-sided CUSUM algorithms are approximately the same. Therefore, the formulas for the ARL function of the *two-sided CUSUM* algorithm can be applied to the case of the χ^2-CUSUM algorithm.

5.3 The GLR Algorithm

In this section, we continue to discuss the application of Lorden's results when the available *a priori* information about θ_1 is minimum. We follow the ideas of [Lorden, 1971, Lorden, 1973]. We first describe the properties of the GLR algorithm and then discuss the relation between these statistical properties and different levels of *a priori* information. We also compare the GLR and CUSUM-type algorithms.

5.3.1 Properties of the GLR Algorithm

Let us continue the previous discussion concerning Lorden's results. We still consider a change in the scalar parameter of an independent sequence modeled with the aid of an exponential family of distributions (5.2.90) as before, but now the parameter after change θ_1 is unknown. Again we start from an open-ended test. Theorem 5.2.1 states the relation between the properties of a class of open-ended tests and the properties of the GLR algorithm for change detection. Thus, we first investigate the properties of the relevant open-ended test in this case, and then describe the consequence for the properties of GLR using this result.

In the present case, we consider hypotheses $\mathbf{H}_0 : \{\theta = \theta_0\}$ and $\mathbf{H}_1 : \{\theta \geq \underline{\theta}\}$, where $\theta_0 < \underline{\theta}$. The open-ended test is then defined as

$$\hat{T} = \begin{cases} \min\{n \geq 1 : \Lambda_1^n \geq e^h\} \\ \infty \text{ if no such } n \text{ exists} \end{cases} \tag{5.3.1}$$

where

$$\Lambda_k^n = \sup_{\underline{\theta} \leq \theta} e^{(\theta - \theta_0) S_k^n - (n-k+1)[d(\theta) - d(\theta_0)]} \tag{5.3.2}$$

and where

$$S_k^n = \sum_{i=k}^{n} y_i \tag{5.3.3}$$

This stopping rule can be rewritten as

$$S_1^n \geq \inf_{\underline{\theta} \leq \theta} \left[\frac{h}{\theta - \theta_0} + n \frac{d(\theta) - d(\theta_0)}{\theta - \theta_0} \right] \tag{5.3.4}$$

It results from theorem 5.2.1 that bounds for the mean time between false alarms and mean delay for detection can be estimated when the error probability and the ASN of the open-ended test are known. But in the present case, Wald's inequality cannot be used because GLR is a *generalized* likelihood ratio test and not simply a likelihood ratio test. Thus, it is necessary to estimate the error probability α and the ASN of the GLR test. For this purpose, we make use of the following theorem [Lorden, 1973] :

Theorem 5.3.1 *When the threshold h and the error probability α are connected through*

$$e^{-h} = \frac{\alpha}{3 \ln \alpha^{-1} \left[1 + \frac{1}{\mathbf{K}(\underline{\theta}, \theta_0)} \right]^2} \tag{5.3.5}$$

then the ASN of the open-ended test satisfies

$$\mathbf{E}_\theta(\hat{T}) \leq \frac{\ln \alpha^{-1} + \ln \ln \alpha^{-1}}{\mathbf{K}(\theta, \theta_0)} + 2 \, \frac{\ln \left\{ \sqrt{3} \left[1 + \frac{1}{\mathbf{K}(\underline{\theta}, \theta_0)} \right] \right\}}{\mathbf{K}(\theta, \theta_0)} + \frac{\theta^2 \, \ddot{d}(\theta)}{\mathbf{K}^2(\theta, \theta_0)} + 1 \tag{5.3.6}$$

for all θ such that $\underline{\theta} \leq \theta$.

The main lines of the proof of this theorem are as follows. We start with the computation of the ASN. Let us fix one θ such that $\theta \geq \underline{\theta}$. It results from (5.3.4) that approximately, when h goes to infinity, we can assume that the ASN of the GLR test can be computed as the ASN of the one-sided SPRT with this θ, with the aid of Wald's identity, namely

$$\mathbf{E}_\theta(\hat{T}) \approx \frac{h}{\mathbf{K}(\theta, \theta_0)} \tag{5.3.7}$$

As already discussed in section 5.2, this relation does not include the excess over the boundary. If we add the upper bound for the expectation of this excess on the right side of this relation, we obtain an upper bound for the ASN. This idea was discussed in subsection 4.3.2 when computing the bounds (4.3.74) for the ASN function. Because we are dealing with an open-ended test, it results from Wald's identity that

$$\mathbf{E}_\theta(\hat{T}) \leq \frac{h + \sup_{r>0} \mathbf{E}_\theta(y - r | y \geq r > 0)}{\mathbf{K}(\theta, \theta_0)} \tag{5.3.8}$$

For an exponential family of distributions, the supremum of the expectation of the excess can be computed analytically [Lorden, 1970]. From this, we get

$$\mathbf{E}_\theta(\hat{T}) \leq \frac{h}{\mathbf{K}(\theta, \theta_0)} + \frac{\theta^2 \, \ddot{d}(\theta)}{\mathbf{K}^2(\theta, \theta_0)} + 1 \tag{5.3.9}$$

This concludes the computation of the ASN.

Let us now explain the relation between the threshold and the error probability as given in theorem 5.3.1. We follow here the same lines of reasoning as in the geometrical interpretation of GLR in chapter 2. For $n \geq n^* = \frac{h}{\mathbf{K}(\underline{\theta}, \theta_0)}$, the infimum in the equation (5.3.4) is reached at $\underline{\theta}$. Let us prove that

$$\mathbf{P}_{\theta_0}(\hat{T} \leq n^*) \leq e^{-h} n^* \tag{5.3.10}$$

For this, first notice that

$$\mathbf{P}_{\theta_0}(\hat{T} \leq n^*) = \sum_{n=1}^{n^*} \mathbf{P}_{\theta_0}(\hat{T} = n) \tag{5.3.11}$$

Then, fix one $n \leq n^*$. Consider the event $\{\hat{T} = n\}$. This event occurs exactly when the CUSUM S_1^n reaches the threshold

$$\inf_{\theta \geq \underline{\theta}} \left[\frac{h}{\theta - \theta_0} + n \, \frac{d(\theta) - d(\theta_0)}{\theta - \theta_0} \right] \tag{5.3.12}$$

Let us assume that this infimum is reached for θ^n. The probability of this event is equal to the error probability of a one-sided SPRT for the parameter θ^n. Therefore, from Wald's inequality we deduce

$$\mathbf{P}_{\theta_0}(\hat{T} = n | n \leq n^*) \leq e^{-h} \tag{5.3.13}$$

which concludes the computation of the bound (5.3.10).

Finally, noting again that for $n \geq n^*$ the infimum is reached for $\theta = \underline{\theta}$, Wald's inequality can be applied, which leads to

$$\mathbf{P}_{\theta_0}(n^* \leq \hat{T} < \infty) \leq e^{-h} \tag{5.3.14}$$

Therefore, from this discussion, it results that

$$\mathbf{P}_{\theta_0}(\hat{T} < \infty) = \sum_{n=1}^{n^*} \mathbf{P}_{\theta_0}(\hat{T} = n) + \mathbf{P}_{\theta_0}(n^* \leq \hat{T} < \infty) \leq \frac{h e^{-h}}{\mathbf{K}(\underline{\theta}, \theta_0)} + e^{-h} \tag{5.3.15}$$

It is possible to prove that the right side of this inequality is less than α for the choice of threshold (5.3.5) given in theorem 5.3.1.

Now, we discuss the connection between theorem 5.3.1 and the properties of the GLR algorithm for change detection. Recall that the corresponding stopping time is

$$t_a = \min\{k \ : \max_{1 \leq j \leq k} \sup_{\theta \geq \underline{\theta}} [(\theta - \theta_0) S_j^k - (k - j + 1)(d(\theta) - d(\theta_0))]\} \tag{5.3.16}$$

It results from theorem 5.2.1 and theorem 5.3.1 that the mean time $\bar{T}$ between false alarms is such that

$$\bar{T} = \mathbf{E}_{\theta_0}(t_a) \geq \alpha^{-1} \tag{5.3.17}$$

and the worst mean delay for detection satisfies

$$\bar{\tau}^*(\theta) \leq \frac{\ln \bar{T} + \ln \ln \bar{T}}{\mathbf{K}(\theta, \theta_0)} + 2 \, \frac{\ln \left\{ \sqrt{3} \left[1 + \frac{1}{\mathbf{K}(\underline{\theta}, \theta_0)} \right] \right\}}{\mathbf{K}(\theta, \theta_0)} + \frac{\theta^2 \, \ddot{d}(\theta)}{\mathbf{K}^2(\theta, \theta_0)} + 1 \tag{5.3.18}$$

when $\theta \geq \underline{\theta}$.

5.3.2 Discussion : Role of A Priori Information

We now have a result similar to the case of the weighted CUSUM, namely that the price to be paid for the unknown *a priori* information is an additional term in the delay for detection. The main quantity of this additional term, when the mean time between false alarms goes to infinity, is

$$\frac{\ln \ln \bar{T}}{\mathbf{K}(\theta, \theta_0)} \tag{5.3.19}$$

Figure 5.1 *A priori* information for the CUSUM and GLR algorithms. The CUSUM algorithm uses the knowledge of θ_1, whereas the GLR algorithm uses only an interval $[\underline{\theta}, \bar{\theta}]$ of possible values for θ_1.

In the second term, the information about the minimum magnitude of change can be negligible in some asymptotic sense. Assume that $\underline{\theta} \to \theta_0$ as $\bar{T} \to \infty$ in such a way that

$$\mathbf{K}^{-1}(\underline{\theta}, \underline{\theta_0}) \approx (\ln \alpha^{-1})^m \to \infty \tag{5.3.20}$$

where m is a positive real number. In this case, if we choose the threshold h as in theorem 5.3.1, namely

$$e^{-h} = \frac{\alpha}{3 \ln \alpha^{-1} [1 + (\ln \alpha^{-1})^m]^2} \tag{5.3.21}$$

then we get the following relation for the worst mean delay :

$$\bar{\tau}^* \leq \frac{\ln \bar{T} + (2m+1) \ln \ln \bar{T}}{\mathbf{K}(\theta, \theta_0)} + M(\theta) \tag{5.3.22}$$

where $M(\theta)$ is a function of θ and of the excess over the boundary alone, and not of the mean time between false alarms.

Let us now assume, moreover, that the value of the parameter θ after change has an upper bound $\bar{\theta}$. In this case, the error probability can be estimated by using an "approximate inequality" [Lorden, 1973] :

$$\mathbf{P}_{\theta_0}(\hat{T} < \infty) \overset{<}{\approx} e^{-h} \left[1 + \left(\frac{h}{\pi} \right)^{\frac{1}{2}} \ln \left(\frac{\bar{\theta} - \theta_0}{\underline{\theta} - \theta_0} \right) \right] \tag{5.3.23}$$

The main idea underlying the proof of this result is close to the proof of theorem 5.3.1. In this case, the stopping boundary for the cumulative sum S_n is made of two straight segments connected to a curve line. The additional idea is to use the local limit Laplace theorem for Bernoulli schemes [Shiryaev, 1984], for approximating the terms in the sum (5.3.11) as the probability of the number of trials being equal to its expectation. In this case the variance of the sample number of a one-sided SPRT is also used.

Another possible solution for this problem of performance evaluation of the GLR algorithm consists of approximating S_n by a Wiener process and using a formula concerning the associated exit time [Lorden, 1973] given in chapter 3 in (3.1.110). The resulting approximation of the GLR properties are then the same as before.

Finally, let us comment upon the *robustness* issue. In several places in this book, we discuss the issue of robustness of the CUSUM and GLR algorithms with respect to the amount of *a priori* information about the parameters before and after change, as depicted in figure 5.1. Two limit cases are of interest. If we have no *a priori* information about θ_1, the relevant algorithm is GLR. If θ_1 is perfectly known, the relevant algorithm is CUSUM. Intermediate amounts of *a priori* information are discussed later, but it is important to recall that the CUSUM algorithm looses its optimality in these situations. In the second part of the book, we generalize the CUSUM and GLR algorithms for the multidimensional case, and it turns out that the problem of robustness with respect to *a priori* information is even more complex in this case. Nevertheless,

we then use an analogous of the CUSUM algorithm, which we call *linear CUSUM*. This algorithm assumes that the available *a priori* information is concentrated in a separating hyperplane between Θ_0 and Θ_1, and in the shape and size of these sets. The additional motivation for using this algorithm is its computational simplicity and the availability of analytical and numerical results for its performances.

5.3.3 Comparison Between the GLR and CUSUM Algorithms

Let us now compare the properties of the GLR and CUSUM algorithms in the case of a change in the mean of a Gaussian process. Without loss of generality, we assume that $\theta_0 = 0$ and $\sigma^2 = 1$. First, it is necessary to explain carefully the conditions under which such a comparison can take place. Actually, as stated several times before, these two algorithms do not work with the same level of *a priori* information, as depicted in figure 5.1. The CUSUM algorithm knows the value of the parameter after change θ_1, while the GLR algorithm knows only an interval of possible values of θ_1 with lower bound $\underline{\theta}$ and upper bound $\bar{\theta}$. The amount of *a priori* information for the GLR algorithm is measured by the ratio $\frac{\bar{\theta}}{\underline{\theta}}$, and from this point of view the CUSUM algorithm is considered as a degenerate case of the GLR algorithm, for $\bar{\theta} = \underline{\theta} = \theta_1$.

From the results given in this and the previous sections, it results that a comparison between the GLR and CUSUM performances can be obtained through the comparison between the upper bounds for the ASN of the two corresponding open-ended tests, for an equal error probability α.

Let us insist upon the fact that *equal error probabilities result in different thresholds h_C and h_G for CUSUM and GLR*. The comparison between (5.3.23) and (5.2.9) shows that, for the same mean time between false alarms, the threshold has to be greater for the GLR than for the CUSUM algorithm. For this reason, the mean delay for GLR is greater than the delay for CUSUM. Let us now define the *efficiency* of the GLR algorithm with respect to the CUSUM one by the ratio of the ASN of the open-ended stopping times T and $\hat{T}$ associated with the CUSUM and GLR algorithms :

$$e = \frac{\mathbf{E}_{\theta_1}(T)}{\mathbf{E}_{\theta_1}(\hat{T})} \tag{5.3.24}$$

where θ_1 is the true value of the parameter after change. It results from (5.2.7), (5.3.23), and theorem 5.3.1 that

$$\mathbf{E}_{\theta_1}(T) \approx 2\,\frac{h_G - \ln A(h_G)}{\theta_1^2} \tag{5.3.25}$$

$$\mathbf{E}_{\theta_1}(\hat{T}) \approx 2\,\frac{h_G}{\theta_1^2} \tag{5.3.26}$$

where

$$A(h_G) = 1 + \left(\frac{h_G}{\pi}\right)^{\frac{1}{2}} \ln\left(\frac{\bar{\theta}}{\underline{\theta}}\right) \tag{5.3.27}$$

and h_G and h_C are the thresholds for the GLR and CUSUM algorithms, respectively. Now, from (5.3.23) and (5.2.9), we conclude that

$$e = 1 - \frac{\ln A(h_G)}{h_G} \tag{5.3.28}$$

This formula shows that the efficiency does go to 1 when the ratio $\frac{\bar{\theta}}{\underline{\theta}}$ goes to 1. Moreover, numerical computations show that the efficiency of GLR increases with the mean time between false alarms, whatever the ratio of *a priori* information is.

5.4 Bayes-type Algorithms

We now investigate the properties of Bayes-type algorithms, which we described in section 2.3, and which have been proven to be *optimal* in [Shiryaev, 1961, Shiryaev, 1963, Shiryaev, 1978]. Even though these algorithms are introduced for discrete time processes as usual in this book, we consider *continuous* time processes for the investigation of their properties. Therefore, we consider the following model of change :

$$dy_t = \nu \, \mathbf{1}_{\{t \geq t_0\}} \, dt + \sigma \, dw_t \tag{5.4.1}$$

where $(w_t)_t$ is a normalized Brownian motion. In other words, we consider the continuous time counterpart of the problem of detecting a change in the mean value of a Gaussian random sequence. Without loss of generality, we assume that, before the change time t_0, the mean of the observed process $(y_t)_t$ is zero, and after the change time, this mean is $\nu \, (t - t_0)$. Moreover, in the present Bayesian framework, we assume that the *a priori* distribution of the change time is

$$\mathbf{P}_\pi(t_0 < t) = 1 - e^{\lambda \, t} \tag{5.4.2}$$

and we define

$$\alpha = \mathbf{P}_\pi(t_a < t_0) \tag{5.4.3}$$

to be the probability of false alarm.

Let us consider the asymptotic situation where λ goes to zero, and α goes to one, in such a way that

$$\bar{T} = \frac{1 - \alpha}{\lambda} \tag{5.4.4}$$

In other words, we assume that the mean time between changes goes to infinity and consequently that the probability of false alarm goes to one. When the mean time between false alarms $\bar{T}$ goes to infinity, the delay of the optimal algorithm is [Shiryaev, 1965, Shiryaev, 1978]

$$\bar{\tau}(\bar{T}) = \frac{2\sigma^2}{\nu^2} \left[\ln \left(\frac{\nu^2}{2\sigma^2} \bar{T} \right) - 1 - C + O\left(\frac{2\sigma^2}{\nu^2 \bar{T}} \right) \right] \tag{5.4.5}$$

where C is the Euler constant $C = 0.577\ldots$ More recent investigations of Bayes-type algorithms can be found in [Pollak and Siegmund, 1985].

5.5 Analytical and Numerical Comparisons

The goal of this section is twofold. First, because of the central role played by the CUSUM algorithm in this book, we compare the various available expressions for the ARL function of the CUSUM algorithm in the basic example of change in the mean of an independent Gaussian sequence. Second, we compare analytically and numerically different algorithms, namely CUSUM, Bayes, FMA, GMA, and Shewhart's charts.

5.5.1 Comparing Different ARL Expressions for the CUSUM Algorithm

We discuss the case of a change in the mean of a Gaussian sequence, and compare the "exact" ARL function to its various approximations and bounds derived in section 5.2. We consider a change in the mean μ_y of

an independent Gaussian sequence $(y_n)_{n\geq 1}$ with known variance σ_y^2. We use the CUSUM algorithm with preassigned values μ_0 and μ_1. In this case, the increment of the cumulative sum (5.2.17) is

$$s_k = \frac{\mu_1 - \mu_0}{\sigma_y^2}\left(y_k - \frac{\mu_1 + \mu_0}{2}\right) \tag{5.5.1}$$

and is a Gaussian random value, $\mathcal{L}(s_k) = \mathcal{N}(\mu, \sigma^2)$. Its mean μ is

$$\mu = \frac{\mu_1 - \mu_0}{\sigma_y^2}\left(\mu_y - \frac{\mu_1 + \mu_0}{2}\right) \tag{5.5.2}$$

and its variance is $\sigma^2 = \frac{(\mu_1-\mu_0)^2}{\sigma_y^2}$. When $\mu_y = \mu_0$, the mean of the increment is $-\mathbf{K}(\mu_0, \mu_1)$ and when $\mu_y = \mu_1$, it is $+\mathbf{K}(\mu_0, \mu_1)$, where $\mathbf{K}(\mu_0, \mu_1) = \frac{(\mu_1-\mu_0)^2}{2\sigma_y^2}$ is the Kullback information.

Now, we are interested in the ARL function as a function of the parameter μ_y or equivalently of μ.

5.5.1.1 Different Approximations

Let us first compute $\gamma(\theta)$, which we denote now as $\gamma(\mu)$, as follows :

$$\gamma(\mu) = \mathbf{E}_\mu(s_k | s_k > 0) = \frac{\int_0^\infty x p_\mu(x)\, dx}{\int_0^\infty p_\mu(x)\, dx} \tag{5.5.3}$$

where $p_\mu(x)$ is the density of s_k. Obvious computations give rise to

$$\gamma(\mu) = \frac{\sigma\varphi(\frac{\mu}{\sigma})}{\phi(\frac{\mu}{\sigma})} + \mu \tag{5.5.4}$$

where $\varphi(x)$ and $\phi(x)$ are the Gaussian density and cdf defined in (3.1.14).

In the Gaussian case, the nonzero solution of the equation $\mathbf{E}_\mu(e^{-\omega_0 s_k}) = 1$ is given by $\omega_0 = \frac{2\mu}{\sigma^2}$. From (5.2.69) and (5.2.77), it results that

$$\bar{L}_0(\mu > 0) = \frac{h}{\mu} + \frac{\sigma\varphi(\frac{\mu}{\sigma})}{\mu\phi(\frac{\mu}{\sigma})} + 1 \tag{5.5.5}$$

$$\underline{L}_0(\mu < 0) = \frac{e^{-\frac{2\mu}{\sigma^2}h} - 1 + \frac{2\mu}{\sigma^2}h}{\frac{2\mu^2}{\sigma^2}} + \frac{\sigma\varphi(\frac{\mu}{\sigma})}{\mu\phi(\frac{\mu}{\sigma})} + 1 \tag{5.5.6}$$

From (5.2.44) and (5.2.48), the Wald's approximation of the ARL function can be written as

$$\hat{L}_0(\mu \neq 0) = \frac{e^{-\frac{2\mu}{\sigma^2}h} - 1 + \frac{2\mu}{\sigma^2}h}{\frac{2\mu^2}{\sigma^2}} \tag{5.5.7}$$

$$\hat{L}_0(0) = \frac{h^2}{\sigma^2} \tag{5.5.8}$$

On the other hand, from (5.2.64) and (5.2.65), the Siegmund's approximation of the ARL function can be written as

$$\tilde{L}_0(\mu \neq 0) = \frac{\exp\left[-2\left(\frac{\mu h}{\sigma^2} + \frac{\mu}{\sigma}\cdot 1.166\right)\right] - 1 + 2\left(\frac{\mu h}{\sigma^2} + \frac{\mu}{\sigma}\cdot 1.166\right)}{\frac{2\mu^2}{\sigma^2}} \tag{5.5.9}$$

$$\tilde{L}_0(0) = \left(\frac{h}{\sigma} + 1.166\right)^2 \tag{5.5.10}$$

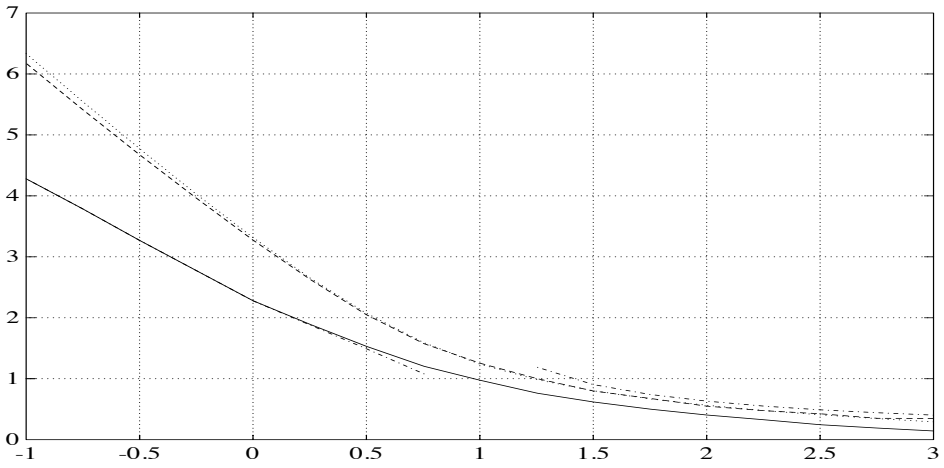

Figure 5.2 ARL function for the Gaussian case ($\log L(\mu_y)$) : "exact" function (dashed line); Wald's approximation (solid line); Siegmund's approximation (dotted line); and bounds (dash-dot lines).

Table 5.1 Comparison between the "exact" ARL, Wald's, and Siegmund's approximations, and bounds.

$\mu = \mathbf{E}(s)$	"Exact" ARL	Wald's approx.	Siegmund approx.	Bounds
-2.0	1.5×10^6	2.03×10^4	2.16×10^6	2.03×10^4
-1.5	4.8×10^4	1.8×10^3	5.95×10^4	1.8×10^3
-1.0	2.0×10^3	198	2.07×10^3	197
-0.5	112.2	32.2	118.6	30.9
0.0	17.3	9.0	17.36	-
0.5	6.53	4.1	6.36	8.02
1.0	3.75	2.5	3.67	4.29
1.5	2.69	1.78	2.56	3.09
2.0	2.12	1.38	1.96	2.53

The typical behavior of the ARL computed by the numerical solution of the Fredholm integral equations ("exact" ARL), of the approximations of the ARL and of these bounds, is depicted in figure 5.2, which we discuss later in this subsection. The numerical values of these different quantities are listed in the table 5.1. These results are obtained in the case where $\mu_0 = 0, \mu_1 = 2, \sigma_y = 1$ and for the threshold $h = 3$. The figure and the table both show that the Siegmund's approximation is very close to the "exact" value of the ARL function. Wald's approximation is worse, especially for a negative drift of the increment of the decision function, namely for the mean time between false alarms, as we show now.

Let us consider the limit of the difference $\hat{L}_0(\mu) - \underline{L}_0(\mu)$ when $\mu \to -\infty$. For this purpose, we use the asymptotic formula

$$\phi(-x) \sim \frac{1}{x\sqrt{2\pi}} e^{-\frac{x^2}{2}} (1 - \frac{1}{x^2} + \frac{3}{x^4} + \cdots) \text{ when } x \to +\infty \tag{5.5.11}$$

Then

$$\lim_{\mu\to-\infty}\frac{\sigma\varphi(\frac{\mu}{\sigma})}{\mu\phi(\frac{\mu}{\sigma})}=\sigma\lim_{\mu\to-\infty}\frac{e^{-\frac{\mu^2}{2}}}{\mu\left[-\frac{1}{\mu}e^{-\frac{\mu^2}{2}}(1-\frac{1}{\mu^2}+\frac{3}{\mu^4}+\cdots)\right]}=-1 \tag{5.5.12}$$

and

$$\lim_{\mu\to-\infty}[\hat{L}_0(\mu)-\underline{L}_0(\mu)]=0 \tag{5.5.13}$$

In other words, when $\mu\to-\infty$, the Wald's approximation acts as a bound.

5.5.1.2 Two Bounds for the Mean Time Between False Alarms

We now compare the two bounds (5.2.77) - bound 1 - and (5.2.80) - bound 2 - again in the Gaussian case. It results from inequality (5.2.80) that bound 2 should be more accurate when the absolute value $|\mu|$ of the expectation of the increment s_k of the decision function (5.2.17) is large, and conversely should be poor when this absolute value is small. The accuracy of bound 1 should vary in reverse order. To prove these intuitive statements, we compute these bounds in the case where $\sigma^2=1$, for increasing values of $|\mu|$ and different values of the threshold h, and compare them to the "exact" value of $\bar{T}$ computed through the solution of the Fredholm integral equations, as discussed in example 5.2.1. The results are summarized in tables 5.2 to 5.4, where missing values indicate either a loss of numerical accuracy for the "exact" value or a negative value for the bounds. These tables prove the above-mentioned intuitive statement about the accuracy of the two bounds : For small values of $|\mu|$ (or equivalently of the change magnitude), namely in tables 5.2 and 5.3, bound 1 is more accurate; for intermediate values of the change magnitude, the two bounds are comparable, though nonaccurate; and for large values of $|\mu|$ (table 5.4), bound 2 is more accurate, though nonaccurate also.

5.5.2 Comparison Between Different Algorithms

Many comparative results can be found in the literature concerning the algorithms described in chapter 2. In [Shiryaev, 1961], three algorithms are compared with the aid of numerical computation of complex and nonasymptotic formulas for the delay and mean time between false alarms. These algorithms are Shewhart's control charts, CUSUM, and Bayes algorithms. The problem statement that is assumed is that of *random* change time, which is described in sections 4.4 and 5.4. Moreover, asymptotically when $\bar{T}$ goes to infinity, and assuming that $\frac{\nu^2}{2\sigma^2}=1$, the following approximations are given in [Shiryaev, 1961] :

$$\begin{array}{lcll}\bar{\tau} & = & \ln\bar{T}-1-C+o(1) & \text{for the Bayes algorithm}\\ \bar{\tau} & = & \ln\bar{T}-\frac{3}{2}+o(1) & \text{for the CUSUM algorithm}\\ \bar{\tau} & \sim & \frac{3}{2}\ln\bar{T} & \text{for Shewhart's chart}\end{array} \tag{5.5.14}$$

Note that we here recover (5.4.5) in the first of these formulas. For the CUSUM and Shewhart algorithms, it can be proven that, in the two other previous formulas, we recover (5.5.7) and (5.1.12). Let us explain these last two formulas. The characteristic feature of Shiryaev's comparison is the assumption of the existence of a steady state reached by each of the algorithms that are compared. In other words, before t_0, the decision functions are assumed to reach stationary distributions, and the mean delays are computed with respect to *these* distributions. Therefore, the direct application of (5.5.7) and (5.1.12) would lead to results slightly different from the previous approximations of Shiryaev, but basically the order of magnitude of the main term is the same. It results from these relations that, for large $\bar{T}$, the CUSUM algorithm has practically the same delay as the optimal Bayes algorithm for *this* problem statement. More recently, another comparison between Bayes-type and CUSUM algorithms has been done through analytical formulas and simulations

Table 5.2 Comparison between the "exact" value of $\bar{T}$ and the two bounds - $\mu = -10^{-2}$.

Threshold (h)	"exact" $\bar{T}$	bound 1	bound 2
5	39.6	-	1.1
10	134	27.6	1.22
15	295	170	1.35
16	343	206	1.38
18	-	287	1.43

Table 5.3 Comparison between the "exact" value of $\bar{T}$ and the two bounds - $\mu = -10^{-1}$.

Threshold (h)	"exact" $\bar{T}$	bound 1	bound 2
1	5.51	-	1.22
2	12.5	-	1.49
3	23.4	3.48	1.82
4	38.8	13.7	2.22
5	59.9	28.3	2.72
6	87.9	48.4	3.32
10	305	212	7.39
15	1090	797	20

Table 5.4 Comparison between the "exact" value of $\bar{T}$ and the two bounds - $\mu = -5$.

Threshold (h)	"exact" $\bar{T}$	bound 1	bound 2
0.1	5.87×10^6	0.01	2.71
0.2	9.99×10^6	0.08	7.39
0.4	2.96×10^7	1	54.6
0.6	8.95×10^7	7.93	403.4
0.8	2.65×10^8	59.4	2981
1.0	6.94×10^8	440	2.2×10^4
1.1	1.02×10^9	1.2×10^3	5.99×10^4
1.2	1.36×10^9	3.25×10^3	1.63×10^5

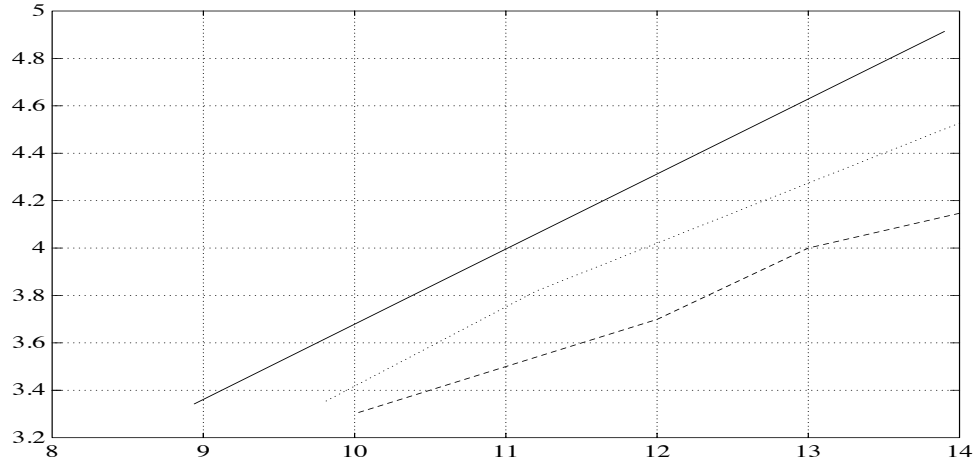

Figure 5.3 Comparison between the Shewhart, CUSUM, and GMA algorithms in the Gaussian case : $\log(\bar{T})$ as a function of $\bar{\tau}$ - Shewhart (dashed line); CUSUM (solid line); GMA (dotted line).

in [Pollak and Siegmund, 1985]. The result is that both algorithms are almost indistinguishable if the preassigned and actual change magnitudes are the same.

In [S.Roberts, 1966], five algorithms are compared, namely Shewhart's chart and the GMA, FMA, CUSUM, and Bayes algorithms, with the aid of simulations, again in our basic case of change in the mean of a Gaussian distribution. Two types of results are obtained. First, for a given $\bar{T}$ and a given change magnitude ν, the optimal free parameters are chosen and the resulting mean delays for detection are compared. Second, the robustness of these algorithms is investigated with respect to the change magnitude ν. The main conclusion of this comparison is that, when the change magnitude ν is *small*, the CUSUM algorithm is better than the GMA, FMA, and Shewhart algorithms, and is approximately as efficient as the Bayes algorithm. A similar conclusion was reached in [Basseville, 1981] where the CUSUM algorithm is shown, through simulations, to be both more efficient and more robust than the filtered derivative algorithms.

Other comparisons are discussed in [Goldsmith and Whitfield, 1961, Gibra, 1975] and in the survey papers [Taylor, 1968, Phillips, 1969, Montgomery, 1980].

The numerical results of a comparison between the Shewhart, CUSUM, and GMA algorithms are depicted in figure 5.3, again in the Gaussian case with $\nu = 1.2$ and $\sigma = 1$.

5.6 Notes and References

Section 5.1

Shewhart control chart

The optimal tuning of the parameters of a simple Shewhart chart and the computation of its ARL function are addressed in [Page, 1954c]. More complex Shewhart algorithms are discussed in [Page, 1955, Page, 1962, Shiryaev, 1961] and the influence of data correlation is discussed in [Vasilopoulos and Stamboulis, 1978].

GMA algorithm

The first numerical investigation of the properties of the GMA algorithm was proposed in [Robinson and Ho, 1978]. Further investigations are in [Crowder, 1987, Novikov and Ergashev, 1988, Fishman, 1988] and more sophisticated GMA algorithms are investigated in [Hines, 1976a, Hines, 1976b].

FMA algorithm

To our knowledge, the only papers concerning the investigation of the properties of the FMA algorithms are [Laï, 1974, Böhm and Hackl, 1990], and no analytical results exist for the filtered derivative algorithms.

Section 5.2

The history of the investigations of the properties of the CUSUM algorithm is quite long, as can be seen from the books [Van Dobben De Bruyn, 1968, Nikiforov, 1983, Montgomery, 1985, Duncan, 1986, Wetherill and Brown, 1991] and the survey papers [Phillips, 1969, Montgomery, 1980].

Optimal properties

The main references concerning the proof of optimality of the CUSUM algorithm are [Lorden, 1971, Moustakides, 1986, Ritov, 1990].

ARL function

The numerical computation of the ARL function of the CUSUM algorithm is discussed in [Page, 1954b, Van Dobben De Bruyn, 1968, Kemp, 1967a, Goel and Wu, 1971, Kireichikov *et al.*, 1990]. The use of a Brownian motion for the approximation of the ARL function was introduced in [R.Johnson and Bagshaw, 1974, Bagshaw and R.Johnson, 1975a, Taylor, 1975]. Wald's approximations and their modifications are discussed in [Page, 1954b, Ewan and Kemp, 1960, Kemp, 1958, Kemp, 1961, Reynolds, 1975, Nikiforov, 1978, Khan, 1978, Nikiforov, 1980, Khan, 1981]. Different types of corrected diffusion approximations are discussed in [Siegmund, 1975, Siegmund, 1979, Siegmund, 1985a, Siegmund, 1985b]. Bounds for the ARL function are discussed in [Kemp, 1967b, Nikiforov, 1980].

A Markov chain approach to the investigation of the properties of CUSUM algorithms was introduced in [Lucas and Crosier, 1982] and was further investigated in [Woodall, 1984, Yashchin, 1985a, Yashchin, 1985b]. Martingale techniques were introduced in [Kennedy, 1976] and used in [Basseville, 1981].

The ARL function of the two-sided CUSUM algorithm has been investigated in several papers and books [Van Dobben De Bruyn, 1968, Khan, 1981, Yashchin, 1985a, Siegmund, 1985b, Wetherill and Brown, 1991].

CUSUM-type algorithms

The properties of the weighted CUSUM algorithm were investigated in [Pollak and Siegmund, 1975]. The properties of the χ^2-CUSUM algorithm given in this chapter are new.

Section 5.3

The properties of the GLR algorithm were investigated first in [Lorden, 1971, Lorden, 1973] and more recently in [Dragalin, 1988].

Section 5.4

The first investigation of the Bayes-type algorithms can be found in [Shiryaev, 1961, Shiryaev, 1963, Shiryaev, 1965]. Since then, many references have appeared in the literature [Taylor, 1967, Shiryaev, 1978, Pollak and Siegmund, 1985, Pollak, 1985, Pollak, 1987].

Section 5.5

Comparisons between the algorithms described in chapter 2 are reported in [Shiryaev, 1961, S.Roberts, 1966, Phillips, 1969, Montgomery, 1980, Basseville, 1981, Pollak and Siegmund, 1985].

5.7 Summary

ARL of the Elementary Algorithms

Shewhart's chart

$$L(\theta) = \frac{N}{\beta(\theta)}$$

GMA chart

$$L_z = 1 + \frac{1}{\alpha}\int_{-\lambda}^{\lambda} L_y f_\theta \left[\frac{y-(1-\alpha)z}{\alpha}\right] dy$$

where z is the initial value of the decision function.

FMA chart

$$1 + \frac{q_N(h)}{p_N(h)} \leq L(\mu) \leq N + \frac{q_N(h)}{p_N(h)}$$

CUSUM-type Algorithms

Optimal properties of the CUSUM algorithm

$$\bar{\tau}^* \sim \frac{\ln \bar{T}}{\mathbf{K}(\theta_1, \theta_0)}$$

ARL of the CUSUM algorithm

"Exact" ARL

$$L_0 = \frac{N(0)}{1 - \mathbf{P}(0)}$$

$$\begin{aligned}\mathbf{P}(z) &= \int_{-\infty}^{-z} f_\theta(x)dx + \int_0^h \mathbf{P}(x) f_\theta(x-z)dx, \;\; 0 \le z \le h \\ N(z) &= 1 + \int_0^h N(x) f_\theta(x-z)dx, \;\; 0 \le z \le h\end{aligned}$$

Wald's approximation

$$L_0(\theta) \approx \frac{1}{\mathbf{E}_\theta(s_k)} \left(h + \frac{e^{-\omega_0 h}}{\omega_0} - \frac{1}{\omega_0} \right)$$

Siegmund's approximation

$$L_0(\theta) \approx \frac{1}{\mathbf{E}_\theta(s_k)} \left(h + \varrho_+ - \varrho_- + \frac{e^{-\omega_0 (h+\varrho_+ - \varrho_-)}}{\omega_0} - \frac{1}{\omega_0} \right)$$

Bounds

$$\bar{\tau}^* \le \frac{1}{\mathbf{E}_\theta(s_k)} \left[h + \gamma(\theta) \right]$$

$$\bar{T} \ge \frac{1}{\mathbf{E}_\theta(s_k)} \left[\frac{e^{-\omega_0 h} - 1}{\omega_0} + h + \gamma(\theta) \right]$$

Two-sided CUSUM algorithm

$$\frac{1}{L^T(\mu)} = \frac{1}{L^l(\mu)} + \frac{1}{L^u(\mu)}$$

Properties of the CUSUM-type algorithms

$$\bar{\tau}^*(\theta_1) \approx \frac{\ln \bar{T} + \frac{1}{2} \ln \left[\frac{\ln \bar{T}}{\mathbf{K}(\theta_1, \theta_0)} \right]}{\mathbf{K}(\theta_1, \theta_0)} - \frac{1}{2\, \mathbf{K}(\theta_1, \theta_0)} \left\{ \ln \left[2\pi \frac{\dot{F}^2(\theta_1)}{\ddot{d}(\theta_1)} \right] - 1 \right\} + o(1)$$

Properties of χ^2-CUSUM algorithms

$$\bar{\tau}^* \sim \frac{\ln \bar{T} + \ln 2}{\mathbf{K}(\theta_1, \theta_0)}$$

GLR Algorithm

$$\bar{\tau}^*(\theta) \le \frac{\ln \bar{T} + \ln \ln \bar{T}}{\mathbf{K}(\theta, \theta_0)} + 2\, \frac{\ln \left\{ \sqrt{3} \left[1 + \frac{1}{\mathbf{K}(\underline{\theta}, \theta_0)} \right] \right\}}{\mathbf{K}(\theta, \theta_0)} + \frac{\theta^2\, \ddot{d}(\theta)}{\mathbf{K}^2(\theta, \theta_0)} + 1$$

Bayes-type Algorithms

$$\bar{\tau}(\bar{T}) = \frac{2\sigma^2}{\nu^2} \left[\ln \left(\frac{\nu^2}{2\sigma^2} \bar{T} \right) - 1 - C + O\left(\frac{2\sigma^2}{\nu^2 \bar{T}} \right) \right]$$

Part II

Extension to More Complex Changes

6

Introduction to Part II

We now enter the second part of the book, which is devoted to the investigation of change detection problems in situations that are more complex than those of part I. In part I, we discussed the case of *independent* random sequences with a distribution parameterized by a *scalar parameter*. In part II, we extend the algorithms and results presented in part I to **dependent** processes characterized by a **multidimensional parameter**. As discussed in section 1.3, we distinguish between two main problems :

1. detecting *additive changes*;
2. detecting *nonadditive or spectral changes*.

In part I, we did not distinguish between these two types of changes. The reasons that we make this distinction now will become clear later. Part II is organized as follows. In chapter 7, we investigate additive changes. Chapters 8 and 9 are devoted to nonadditive changes for scalar and multidimensional signals, respectively. Here in the introduction, we describe the two above-mentioned problems, discuss the corresponding modeling issues, and introduce the key ideas and tools to be used throughout part II.

6.1 Additive and Nonadditive Changes

Let us now describe the two problems : additive and nonadditive changes. Most change detection problems can be classified into one of these two categories. Of course, these two types of changes can occur simultaneously, but it is of interest to investigate their main characteristics separately. We describe these two types of changes using both intuitive and theoretical points of view.

Intuitively, anyone who has ever looked at a real signal knows the qualitative difference between a jump, namely a change in the mean value of the signal, as depicted in figure 1.1, and a change in the behavior around a mean level, as depicted in figure 6.1. It is thus easy to guess that the two corresponding change detection problems do not present the same characteristics or the same degree of difficulty. The fact that most of the practically relevant change detection problems fall into one of the two above-mentioned classes is illuminated by the theoretical point of view of the underlying models. In the next section, we shall show this, using the mathematical statistics of some parameterized families of conditional probability densities. We first consider the case of linear models, namely regression and ARMA models, and then the control point of view of state-space models. Finally, we consider a class of nonlinear models belonging to the family of conditional probability distributions.

Before investigating this modeling issue in more detail, let us summarize the main features of the detection of the two types of changes. In this discussion, we consider the ideal case where the parameters before and after change are known.

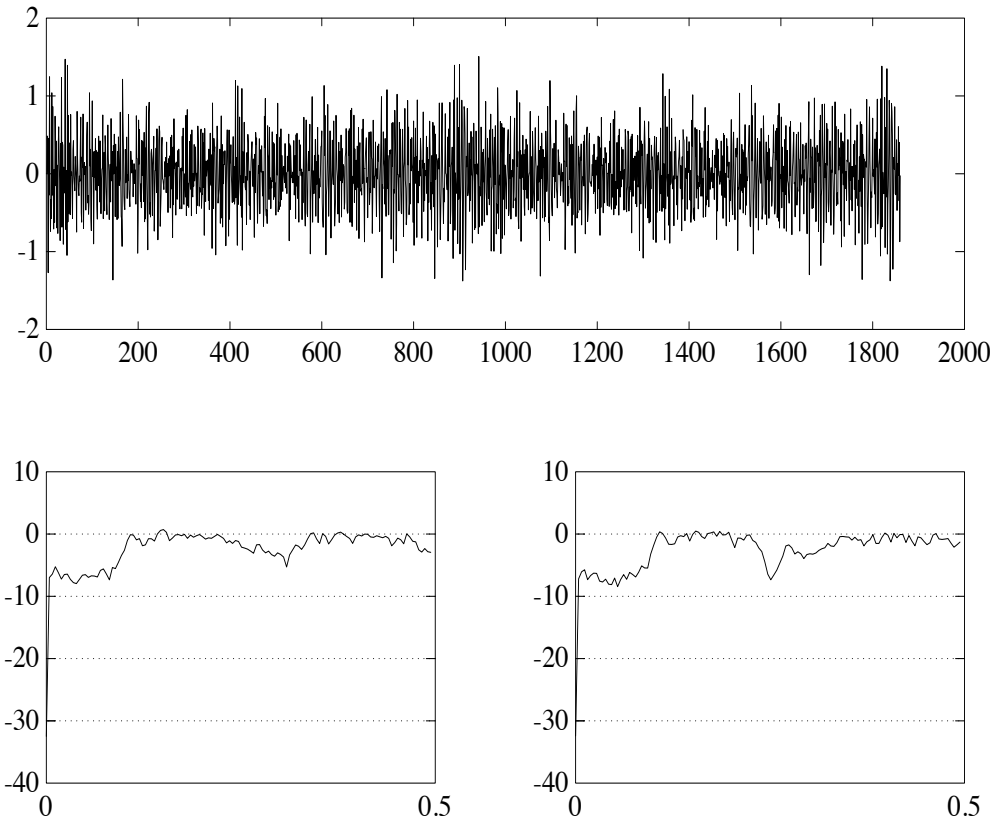

Figure 6.1 A spectral change (first row; change time 1000) and spectral densities before and after change (log-scale - second row).

6.1.1 Additive Changes

Here we refer to any changes in a signal or a linear system that result in *changes only in the mean value* of the sequence of observations. In chapter 7, we investigate the important case of linear systems excited by Gaussian white noises and perturbed by additive changes. This type of dynamic system can be represented with the aid of the transfer function $\mathcal{T}$ of the filter, which generates the observations from a white noise sequence and describes the entire dynamics of the system. We assume that additive changes occur in the mean value θ of the input V to this system, and thus these changes do not affect the dynamics in $\mathcal{T}$. For such systems, the extension of the methods and theoretical results of part I is possible in many cases and relatively easy. The main reason for this is that the problem of detecting additive changes remains unchanged under the transformation from observations to innovations, which corresponds to the whitening filter with transfer function $\mathcal{T}^{-1}$. We continue this discussion when describing figure 6.4.

We show later that the solution to this particular problem is of crucial interest for solving more complex change detection problems.

6.1.2 Nonadditive or Spectral Changes

Here we refer to more general and difficult cases where changes occur in the variance, correlations, spectral characteristics, dynamics of the signal or system. In other words, these changes act as multiplicative changes in the transfer function $\mathcal{T}$. Two cases must be distinguished. In the first case, the change is assumed to occur in the energy of the input excitation V. Thus it does not affect the dynamics of the system and the change detection problem remains unchanged under the transformation from observations to innovations, as before.

In the second case, namely for all the changes that affect the dynamics of the system itself, the problem is more complex. For simplicity, we assume that the dynamics are modeled by the parameterized transfer function $\mathcal{T}_\theta$, and that the change in the dynamics is summarized as a change in the parameter θ from θ_0 to θ_1. For reasons that are explained in section 6.3 and in chapter 8, for detecting such a change, a useful guideline is the following : Process the observations Y through both the inverse filters $\mathcal{T}_{\theta_0}^{-1}$ and $\mathcal{T}_{\theta_1}^{-1}$, and build the decision function of the change detection algorithm upon *both* innovation (or more precisely residual) processes. A more precise statement is presented later. We will show that the extension of the methods of the first part is less simple for nonadditive changes than for additive changes, and that there are fewer available theoretical results.

In chapters 8 and 9, we distinguish scalar and multidimensional observed signals, respectively.

6.2 Modeling Issues

We now investigate the distinction between additive and nonadditive changes using several different modeling points of view. The key model used in part II is a parameterized conditional probability distribution. We discuss additive and nonadditive changes in some special cases of interest. We first concentrate basically on linear models, considering regression, ARMA models, and state-space models. Then, we describe changes in a larger class of models belonging to the family of parameterized conditional probability densities. These nonlinear models are encountered in stochastic approximation theory and in many applications.

From the point of view of conditional distributions, there exist several ways of generating changes. Because the importance of this issue basically arises for nonadditive changes, we address it in chapter 8.

6.2.1 Changes in a Regression Model

We consider regression models such as

$$Y_k = HX_k + V_k \tag{6.2.1}$$

where X_k is the unknown state, $(V_k)_k$ is a white noise sequence with covariance matrix R, and H is a full rank matrix of size $r \times n$ with $r > n$. The characteristic feature of this model is the existence of *redundancy* in the information contained in the observations, which is of crucial importance because of the unknown state. A typical example in which such a model arises is a measurement system where the number of sensors is greater than the number of physical quantities that have to be estimated.

In these models, we discuss only additive changes that occur in the input noise sequence and are modeled by

$$Y_k = HX_k + V_k + \Upsilon(k, t_0) \tag{6.2.2}$$

where $\Upsilon(k, t_0)$ is the dynamic profile of the change occurring at time t_0, namely $\Upsilon(k, t_0) = 0$ for $k < t_0$. We consider different levels of available *a priori* information about Υ, and sometimes we find it useful to consider separately the change magnitude ν and profile Υ.

6.2.2 Changes in an ARMA Model

Now we describe the difference between additive and nonadditive changes in an ARMA model such as

$$Y_k \;=\; \textstyle\sum_{i=1}^{p} A_i Y_{k-i} \;+\; \sum_{j=0}^{q} B_j V_{k-j} \tag{6.2.3}$$

where $(V_k)_k$ is a white noise sequence with covariance matrix R, $B_0 = I$, and the stability assumption is enforced.

1. **Additive changes** : As explained in detail in subsection 7.2.3, additive changes in this model are additive changes on the white noise sequence $(V_k)_k$ as modeled by

$$Y_k \;=\; \textstyle\sum_{i=1}^{p} A_i Y_{k-i} \;+\; \sum_{j=0}^{q} B_j [V_{k-j} \;+\; \Upsilon(k-j, t_0)] \tag{6.2.4}$$

   where $\Upsilon(k, t_0)$ is again the dynamic profile of the change occurring at time t_0. In other words, without loss of generality, we assume that the mean value θ of V_k changes from $\theta_0 = 0$ to $\theta_1 = \Upsilon(k, t_0)$. The reasons this assumption leads to the above model are made clearer in subsection 7.2.3.

2. **Changes in variance** : The first type of nonadditive changes is a change in the variance R of the sequence $(V_k)_k$. Let $(\xi_k)_k$ be a normalized white noise sequence. Then the change is modeled as

$$V_k = \begin{cases} R_0^{\frac{1}{2}} \xi_k & \text{if } k < t_0 \\ R_1^{\frac{1}{2}} \xi_k & \text{if } k \geq t_0 \end{cases} \tag{6.2.5}$$

3. **Spectral changes** : The second type of nonadditive changes contains changes in the shape of the spectrum of the observations. In other words, these changes are changes in the matrix coefficients A_i and B_j :

$$Y_k = \begin{cases} \sum_{i=1}^{p} A_i^0 Y_{k-i} \;+\; \sum_{j=0}^{q} B_j^0 V_{k-j} & \text{if } k < t_0 \\ \sum_{i=1}^{p} A_i^1 Y_{k-i} \;+\; \sum_{j=0}^{q} B_j^1 V_{k-j} & \text{if } k \geq t_0 \end{cases} \tag{6.2.6}$$

These three types of changes are depicted in the three rows of figure 6.4, which we discuss when we introduce the key ideas of part II.

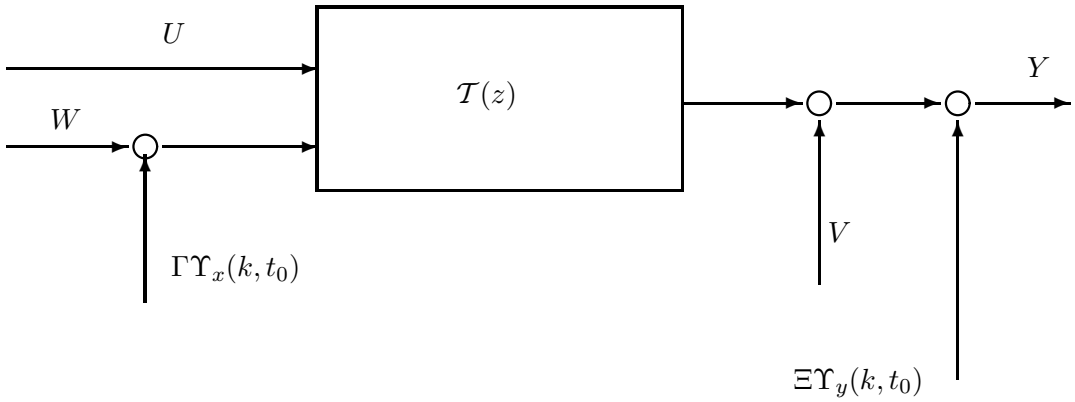

Figure 6.2 Additive changes in a state-space model.

6.2.3 Changes in a State-Space Model

We consider the linear dynamic system described by the state-space representation of the observed signals $(Y_k)_k$:

$$\left\{ \begin{array}{rcl} X_{k+1} & = & FX_k + GU_k + W_k \\ Y_k & = & HX_k + JU_k + V_k \end{array} \right. \tag{6.2.7}$$

where the state X, the measurement Y, and the control U have dimensions n, r, and m, respectively, and where $(W_k)_k$ and $(V_k)_k$ are two independent Gaussian white noises, with covariance matrices Q and R, respectively. We shall comment upon parameterization problems in such multivariable systems later.

The additive and nonadditive types of changes, occurring at an unknown time instant t_0, can be formally defined in the following manner.

1. **Additive changes** : These changes are additive either in the state transition equation or in the observation equation, and thus result in *changes in the mean of the output observations* Y_k. They are modeled by

$$\left\{ \begin{array}{rcl} X_{k+1} & = & FX_k + GU_k + W_k + \Gamma\,\Upsilon_x(k,t_0) \\ Y_k & = & HX_k + JU_k + V_k + \Xi\,\Upsilon_y(k,t_0) \end{array} \right. \tag{6.2.8}$$

   where Γ and Ξ are matrices of dimensions $n \times \tilde{n}$ and $r \times \tilde{r}$, respectively, and where $\Upsilon_x(k,t_0)$ and $\Upsilon_y(k,t_0)$ are the *dynamic profiles* of the assumed changes, of dimensions $\tilde{n} \leq n$ and $\tilde{r} \leq r$, respectively. These matrices and profiles are not necessarily completely known *a priori*. The instant t_0 is again the unknown change time, so $\Upsilon_x(k,t_0) = \Upsilon_y(k,t_0) = 0$ for $k < t_0$. These additive changes can be represented with the aid of figure 6.2. They are discussed in detail in chapter 7, where we exhibit a basic difference between additive changes in ARMA models and state-space models : The profile of the resulting change in the innovation is left unchanged in the ARMA case, but is modified in the case of a state-space model. This is the main motivation for considering any dynamic profile in (6.2.8) and not only steps.

2. **Changes in variances** : Nonadditive changes are changes in the *dynamics* of the system (6.2.7). Obviously, these changes are also changes in the spectrum. They can be represented with the aid of figure 6.3. As in the previous case of ARMA models, the first type of nonadditive changes consists of changes in the covariance matrices Q and R of the noise sequences $(W_k)_k$ and $(V_k)_k$ and are modeled as in (6.2.5).

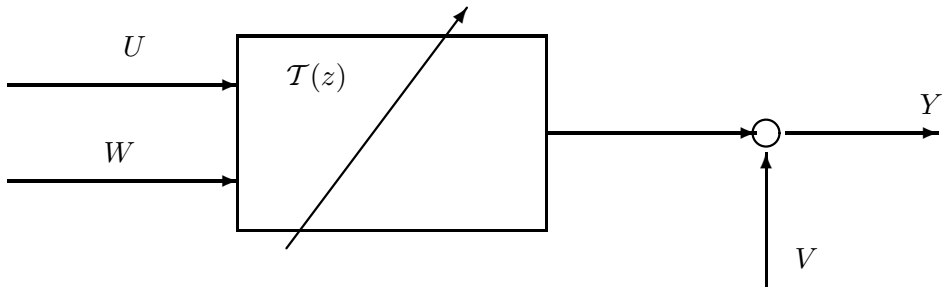

Figure 6.3 Spectral or nonadditive changes.

We also include in this first type of nonadditive changes, changes in the control matrices G and J, modeled by

$$G = \begin{cases} G_0 & \text{if } k < t_0 \\ G_1 & \text{if } k \geq t_0 \end{cases} \tag{6.2.9}$$

and similarly for J. Finally, we also include in this type similar changes in the observation matrix H. The reason for this classification is that these changes affect only the numerator of the transfer function.

Let us emphasize that all the above-mentioned change types, namely the additive changes and the "changes in variance," correspond to *sensor or actuator faults*. The second type of nonadditive changes (described next) corresponds to what are often called *components faults*.

3. **Spectral changes :** The second type of nonadditive changes can be modeled by changes in the state transition matrix F of the state-space model (6.2.7), in the following manner :

$$F = \begin{cases} F_0 & \text{if } k < t_0 \\ F_1 & \text{if } k \geq t_0 \end{cases} \tag{6.2.10}$$

These changes affect the denominator of the transfer function.

Because both types of nonadditive changes (changes in variance and spectral changes) act in a *nonlinear* way with respect to the observations Y_k, the corresponding change detection problem is more difficult to solve than in the case of additive changes, even though a common methodological basis exists for designing change detection and estimation algorithms in both cases, as will be seen later.

We conclude with modeling issues concerning multivariable state-space models. Let us emphasize that this classification of changes is valid only for the particular parameterization that we consider here, and that completely different classifications can arise from alternative parameterizations of the system. More precisely, in some applications the control matrices G and J and/or the observation matrix H contain parts of the dynamics of the system itself. Similarly, changes in the dynamics of the system are sometimes more conveniently modeled with the aid of changes in the pair (H, F) than changes in the matrix F alone. We shall not discuss further this complex multivariable parametric modeling issue and we refer the reader to [Hannan and Deistler, 1988] for a thorough investigation.

6.2.4 Changes in Other Models

In many applications, such as telecommunications or monitoring of complex industrial plants, it is of interest to deal with conditional probability densities that cannot be modeled using the linear problem settings

(transfer functions, ARMA, or state-space models). An example of this situation can be found in a phase-locked loop system [Benveniste *et al.*, 1990] or in a gas turbine [Zhang *et al.*, 1994]. Such types of processes $(Y_k)_k$ can be modeled with the aid of the following Markov representation :

$$\left\{ \begin{array}{rcl} \mathbf{P}(X_k \in B | X_{k-1}, X_{k-2}, \ldots) & = & \int_B \pi_\theta(X_{k-1}, dx) \\ Y_k & = & f(X_k) \end{array} \right. \tag{6.2.11}$$

where $\pi_\theta(X, dx)$ is the transition probability of the Markov process $(X_k)_k$ and where f is a nonlinear function. These processes are called *controlled semi-Markov processes*. Their properties are briefly described in chapter 8. Especially, we shall show that, if $(Y_k)_k$ is an AR(p) process, it can be written in the form (6.2.11) with a linear f, using the Markov process $X_k = \check{\mathcal{Y}}_{k-p}^{k-1}$, the transition probability of which is parameterized with the set of autoregressive coefficients. This type of model can thus be thought of as the *nonlinear counterpart of ARMA models*. It can be shown to be the most general class of conditional probability densities that can be generated with the aid of a *finite dimensional state-space* - compare (6.2.7) and (6.2.11). These models are investigated in detail in [Benveniste *et al.*, 1990].

The changes of interest in such a model are changes in the parameter θ of the transition probability π_θ :

$$\theta = \left\{ \begin{array}{ll} \theta_0 & \text{if } k < t_0 \\ \theta_1 & \text{if } k \geq t_0 \end{array} \right. \tag{6.2.12}$$

The point we want to make here is that, apart from the general likelihood ratio framework that we develop for changes in the general case of conditional distributions, there exists, for the nonlinear models (6.2.11), another way of designing change detection algorithms, which is based not upon the likelihood function, but upon another statistic which comes from stochastic approximation theory and which is used for solving the problem of identification in such models. We discuss this point in the next section, while introducing the key ideas of part II.

6.3 Introducing the Key Ideas of Part II

In part II, we mainly investigate the extension of the algorithms developed in the simplest case of part I to the more complex case of *dependent* processes parameterized by a *multidimensional* parameter. We also address the diagnosis or isolation issue, which is specific of the case of changes in a multidimensional parameter.

In this section, we first describe the main ideas for the design of change detection algorithms, and then we discuss the properties of these algorithms together with possible definitions of the detectability of a change.

6.3.1 Design of Algorithms

In explaining the key issues of this design, we make extensive use of figure 6.4. It is worth emphasizing that this figure is more a convenient guideline for changes in AR and ARMA models than a general picture. The key reason is that, in our general likelihood framework, the whitening operation that is used in this figure should be thought of as being nothing but a useful tool for computing the necessary likelihood functions.

For designing change detection algorithms, we mainly use the statistical framework of the **likelihood ratio** as before. But we also discuss some other tools, namely several geometrical system theory techniques, statistical non-likelihood-based algorithms, and some statistical and geometrical solutions to the diagnosis problem. Finally, we investigate the connections between statistical and geometrical techniques for detecting additive changes.

A traditional approach to failure detection consists of considering that the design of detection algorithms is basically made of two steps :

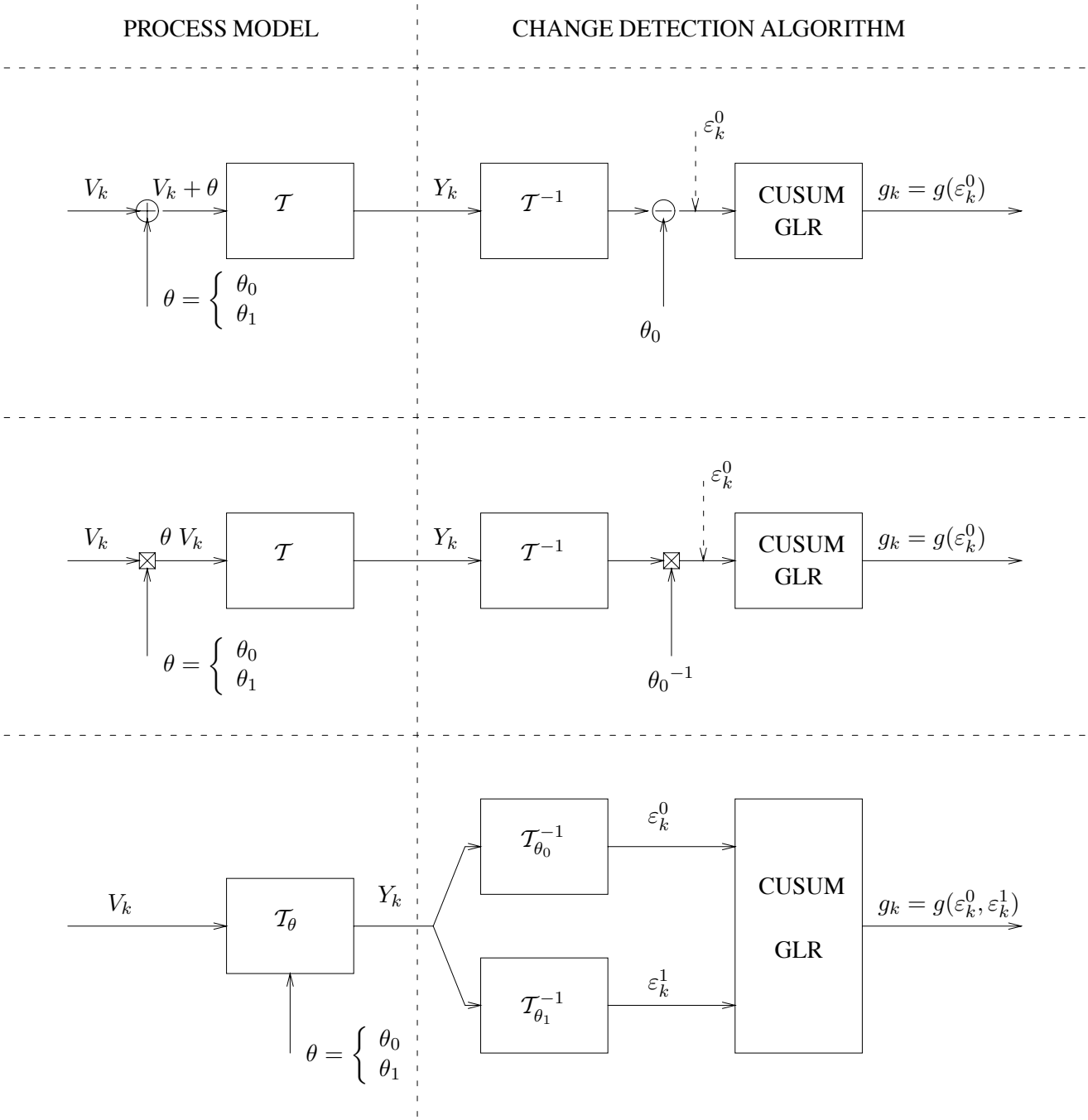

Figure 6.4 Additive and nonadditive change detection algorithms. The algorithms for detecting *additive* changes (first row) are based upon the innovation of the model before the change. The algorithms for detecting *nonadditive* changes (third row) should make use of the innovations of *both* models before and after the change. The only case of nonadditive change that can be detected with the aid of only one innovation is the case of a change in the input variance (second row).

1. generation of "residuals," namely of artificial measurements that reflect the changes of interest; for example, these signals are ideally close to zero before the change and nonzero after;
2. design of decision rules based upon these residuals.

It is interesting to note that this approach has a statistical interpretation because the likelihood ratio is a function of a particular type of residuals, namely those that come from the inverse filtering operation shown in figure 6.4. But geometric tools, relying upon the only deterministic part of the system, can also be used for step 1 and are reported in chapter 7 together with their connection to the statistical ones.

6.3.1.1 Additive Changes

The key tool for detecting additive changes consists of achieving step 1 in such a way that the change detection problem that results from the transformation of the observations into the residuals is nothing but the problem of *detecting changes in the mean value of a multidimensional Gaussian process*, which we call the **basic problem**. Note that for detecting nonadditive changes, the solution to the basic problem is also of crucial importance because of the theoretical results of section 4.2 about the asymptotic local approach. This approach allows us to simplify a complex change detection problem into this particular one, through the use of the efficient score which is the relevant sufficient statistic for small changes in the initial model. But in the case of chapter 7, namely additive changes in a Gaussian distribution, the local approach does not provide anything new, because the efficient score is nothing but the normalized observation itself.

Now, as stated in section 6.1, the main difference between the solutions to the additive and nonadditive change detection problems lies in the way by which the "residuals" are generated. In chapter 7, we investigate additive changes in the case of linear systems excited by Gaussian random noises. In this case, the generation of residuals can be achieved using either the mathematical statistics of **innovation** or the geometrical system theory of **redundancy** relationships or parity checks, as we explain in sections 7.2 and 7.4, respectively. The main advantage of the statistical log-likelihood ratio approach is that it automatically takes into account sensor noises and calibration problems. But it turns out that these two approaches do have common features in both cases of regression and state-space models, as explained in section 7.5. The basic common tools are in terms of the projection operation associated with least-squares estimation in a regression model, and in terms of signatures, of the changes on the residuals, evaluated through transfer function computations. Considering a general profile of additive change on a state-space model, we show that the signature of such a change on the innovation of a Kalman filter is the same as the signature of the same change on the generalized parity check derived through a factorization of the input-output transfer function.

6.3.1.2 Nonadditive Changes

In chapters 8 and 9, we investigate nonadditive changes for scalar and multidimensional signals, respectively. In these cases, the key tool for achieving step 1 is neither the transformation to innovations nor the transformation to parity checks, which are no longer sufficient statistics in this case. The sufficient statistic here is the **likelihood ratio**, which is a function of *both* the residuals ε_k^0 and ε_k^1, which are depicted in the third row of figure 6.4.

In some nonadditive changes, the likelihood ratio is computationally complex, both for the design of the detection algorithm and for the investigation of its properties. In these cases, this complexity can be reduced in two ways. The first key idea is the so-called local approach which leads to the use of the **efficient score** - or first derivative, with respect to the parameter, of the log-likelihood function - which allows us to transform complex change detection problems into the simpler basic problem mentioned before.

The second solution consists of using a statistic that is less complex than the likelihood function, but nevertheless efficient from the statistical inference point of view. This function is basically the function of both the observations and the parameter that is used in stochastic approximation theory for identification purposes. The key idea is then again to use the local approach and the solution to the basic problem. This leads to what we call **non-likelihood-based algorithms**, which are of interest in a number of applications and are investigated in detail in chapter 8.

6.3.1.3 Diagnosis

The diagnosis or isolation issue is concerned with the determination of the origin of the change, once the change in the multidimensional parameter has been detected. Our purpose in this book is not to make a complete investigation of this issue, but rather to describe two possible statistical solutions, based upon decoupling and sensitivity notions, and to investigate the deep connection between statistical and geometrical techniques for failure decoupling in state-space models. More precisely, in section 7.2, we use a general tool for statistical inference about a parameter of interest in the presence of nuisance parameters, called *min-max robust approach*. Then we show that, when applied to the diagnosis of one additive change in a state-space model in the presence of other ones, this approach results in a processing whose first step is nothing but a standard geometrical decoupling technique. In chapter 9, we describe a general sensitivity technique that provides us with a diagnosis of nonadditive changes using either likelihood-based or non-likelihood-based detection algorithms.

6.3.2 Properties of the Algorithms and Detectability

We now discuss the key issues concerning the properties of the algorithms and the definition of the detectability of a change.

6.3.2.1 Computation of the ARL Function

For investigating the properties of the change detection algorithms in part II, we again use the criteria that we define in section 4.4 and use in chapter 5 to investigate the properties of the change detection algorithms corresponding to the simplest case of part I. In part II, we follow the same ideas as in chapter 5, namely the computation of the mean time between false alarms and mean delay for detection through the computation of the ARL function. However, in the more complex situations of part II, the computation of these properties is much more difficult, because the parameter is *multidimensional* and the sequence of observations is *dependent*. Both these factors result in a more complex behavior of the decision function, and thus the computation of the ARL function is more involved.

The key results that are available and that we describe in part II are the following. The computation of the ARL function is possible in the case of additive changes and when we know at least either the change magnitude or the change direction. In the case of nonadditive changes, there are fewer available results, and the computation of the ARL function can be achieved only in more restricted cases. For example, in ARMA models, we can compute the ARL function when the change occurs in the input variance. For changes in the AR and/or MA coefficients, the direct computation of the ARL function is not possible, and we can only estimate the ARL function through an approximation of the decision function by a Brownian motion. But the key issue is that the estimation of the quality of this approximation is basically an *open* problem.

Because of the difficulty in estimating the ARL function, it is of interest to introduce a special kind of weak performance index which allows us to get preliminary insights into the properties of the algorithms. The detectability of a change, defined with the aid of the Kullback information between the two conditional distributions before and after change, can be seen as such a weak performance index, as we discuss now.

6.3.2.2 Statistical Detectability

This issue of detectability is basically discussed for the first time here, and is investigated in detail in all the chapters in part II.

Let us first recall that one of the key results of chapter 5 is theorem 5.3.1 by Lorden concerning the ARL function. This theorem basically states that for an *on-line* algorithm, the delay for detection is inversely proportional to the Kullback information. Therefore, a definition of statistical detectability that is relevant from the point of view of performance indexes is obviously in terms of the Kullback information between the two conditional distributions before and after change. Note that such a definition is stated from *an intrinsic information-based point of view* using the mutual information between the models before and after change. Another useful definition of the detectability could use *a detection performance-based point of view*, measuring the ability of a particular algorithm to detect a particular change. From the latter point of view, a change could be said to be detectable if the expectation of the decision function is not the same before and after change. A simple example is the log-likelihood ratio, the mean of which changes from a negative to a positive value. This is discussed further later.

Consider a change from a distribution p_{θ_0} to a distribution p_{θ_1}. Let s be the log-likelihood ratio :

$$s(y) = \ln \frac{p_{\theta_1}(y)}{p_{\theta_0}(y)} \tag{6.3.1}$$

and let $\mathbf{K}(\theta_1, \theta_0)$, defined by

$$\mathbf{K}(\theta_1, \theta_0) = \mathbf{E}_{\theta_1}[s(Y)] \geq 0 \tag{6.3.2}$$

be the Kullback information. The change is said to be *detectable* if the Kullback information exists and satisfies

$$\mathbf{K}(\theta_1, \theta_0) > 0 \tag{6.3.3}$$

Recall that, in the case of a random process, $\mathbf{K}$ is defined as a limit value when the sample size goes to infinity.

This condition implies that the reversed Kullback information $\mathbf{K}(\theta_0, \theta_1)$ is also positive. Consequently, the Kullback divergence

$$\mathbf{J}(\theta_0, \theta_1) = \mathbf{K}(\theta_0, \theta_1) + \mathbf{K}(\theta_1, \theta_0) \tag{6.3.4}$$

is also positive. The reason we prefer the Kullback information and not divergence becomes clear in chapter 8, which is devoted to spectral changes. Note, however, that the positivity of the Kullback divergence is equivalent to

$$\mathbf{E}_{\theta_1}[s(Y)] - \mathbf{E}_{\theta_0}[s(Y)] > 0 \tag{6.3.5}$$

In other words, a change is detectable if the mean value of the log-likelihood ratio is greater after change than before change. We thus recover the fact that, when using the log-likelihood ratio, the intrinsic information-based and the detection performance-based points of view for defining the statistical detectability are the same. Furthermore, because of Lorden's theorem, for all optimal on-line algorithms, these two definitions are equivalent. Note also that the detectability of a change increases with the Kullback information. This means, in the case of a change in mean, that the detectability of a change increases with the signal-to-noise ratio, which corresponds exactly to an intuitive definition.

Moreover, as explained in section 4.2, the power of optimal *off-line* hypotheses testing algorithms is an increasing function of the Kullback information. Therefore, the detectability defined in terms of Kullback information is also relevant for off-line hypotheses testing algorithms.

The statistical detectability is investigated in detail in chapter 7 for additive changes. In the case of state-space models, we discuss the relations between this criterion and the geometrical detectability used in control

theory and some intuitive transfer function-based detectability definitions . The statistical detectability is also discussed for nonadditive changes in chapters 8 and 9, in less detail.

The contents of part II are summarized in table 6.1.

Table 6.1 Contents of Part II.

chapter	MODELS	GOALS	TOOLS
7	regression ARMA state-space	extension of chapter 2 (vector θ, dependent seq.) detectability diagnosis statistical/geometrical links algorithms properties	innovation redundancy basic problem decoupling
8	conditional law AR ARMA nonlinear ARMA	extension CUSUM,GLR (non additive changes) simplifications of GLR local approach detectability	likelihood ratio efficient score local non-likelihood
9	AR ARMA state-space	extension of chapter 8 (vector signals) diagnosis algorithms properties	same as chapter 8 diagnosis use of state-space

7

Additive Changes in Linear Models

In this chapter, we investigate the problem of detecting additive changes in situations more complex than the case of independent random variables characterized by only one scalar parameter, which was investigated in chapter 2. Additive changes will be considered in four types of models :

1. basic models : multidimensional independent Gaussian sequences;
2. regression models;
3. ARMA models;
4. state-space models.

The main **goals** of this chapter are the following. First, in section 7.2, we *extend to these more complex situations* several of the basic algorithms introduced in chapter 2, namely *CUSUM-type and GLR algorithms*. This is done in subsections 7.2.1, 7.2.2, 7.2.3, and 7.2.4 for the basic, regression, ARMA, and state-space models, respectively. The properties of these algorithms are described in section 7.3. Second, we investigate the *connections between the statistical and geometrical points of view for change detection and diagnosis*. Therefore, we investigate the links between the statistical methods and some geometrical system theory techniques for change detection known in the control literature as analytical redundancy and parity spaces, which we introduce in section 7.4. We address the *diagnosis* issue, mainly from the statistical point of view in subsection 7.2.5. The *detectability* issue is also discussed from both statistical and geometrical points of view in subsections 7.2.6 and 7.4.4, respectively. In section 7.5, we consider the connections between both points of view. We investigate this issue for the design of detection rules, the diagnosis problem, and the detectability definition.

The **tools** for reaching these goals can be summarized as follows. The *transformation of the initial observations into a sequence of innovations* is a key tool for analyzing additive changes in the four types of models mentioned before. The GLR test for detecting a *change in the mean value of an independent vector Gaussian sequence*, which from now on we will refer to as the *basic problem*, will serve as the key approach for deriving all the statistical detection algorithms of this chapter. The concept of *redundancy* is widely used in the geometrical approach; it is often implemented with the aid of projections or observers or spectral factorizations of the input-output transfer function. The use of the concepts of innovation and redundancy for detecting additive changes is introduced in section 7.1.

7.1 Introducing the Tools

In this section, we introduce additive changes in the four types of models. We also introduce the key concepts that are to be used for solving the corresponding detection problem, namely innovations and redundancy.

These concepts will also be useful when discussing the detectability issue from the statistical and geometrical points of view in subsections 7.2.6 and 7.4.4, respectively. We finish with an introduction to the basic problem, which results from the application of these concepts.

7.1.1 Additive Changes in Linear Models

As we discussed when introducing part II, the additive changes considered in this chapter are changes that are additive on the equations of a linear time invariant model, and thus are basically **changes in the mean value** of the distribution of the observed signals. More precisely, let $(Y_k)_{1\leq k\leq N}$ be a sequence of observations with dimension r. We consider the four following types of models :

- **Basic model** : The multidimensional observations are assumed to be Gaussian and to form an independent sequence. The detection problem concerns changes in the mean value of these observations. This is what we call the basic problem. All the other problems are solved using a transformation of the observations resulting in this particular problem.
- **Regression models** :
$$Y_k \;=\; HX_k \;+\; V_k \tag{7.1.1}$$
where the state X has dimension $n < r$, H is of rank n, and where $(V_k)_k$ is a Gaussian white noise sequence with covariance matrix R.
- **ARMA models** :
$$Y_k \;=\; \sum_{i=1}^{p} A_i Y_{k-i} \;+\; \sum_{j=0}^{q} B_j V_{k-j} \tag{7.1.2}$$
where $B_0 = I$, $(V_k)_k$ is a white noise sequence with covariance matrix R, and where we assume that the stability condition holds.
- **State-space models** :
$$\left\{ \begin{array}{rcl} X_{k+1} &=& FX_k + GU_k + W_k \\ Y_k &=& HX_k + JU_k + V_k \end{array} \right. \tag{7.1.3}$$
where the state X and the control U have dimensions n and m, respectively, and where $(W_k)_k$ and $(V_k)_k$ are two independent white noise sequences, with covariance matrices Q and R, respectively.

These four types of models can be viewed in the single framework of state-space models (7.1.3), but it is of interest to consider them separately for the investigation of change detection problems, as will become clear later.

The additive changes in the models that are to be considered in this chapter were introduced in chapter 6 and will be discussed in detail in section 7.2. Here we recall these changes only to clarify the discussion about the tools.

- **Regression models :** We consider the following model of additive change :
$$Y_k \;=\; HX_k \;+\; V_k \;+\; \Upsilon(k, t_0) \tag{7.1.4}$$
The instant t_0 is the unknown change time, so $\Upsilon(k, t_0) = 0$ for $k < t_0$. Sometimes, we consider separately the change magnitude ν and direction Υ. As we explain in chapter 11, this type of model is of key interest for sensor failure detection in inertial navigation systems. For example, a quadruplicate $(r = 4)$ set of sensors measure the vector of accelerations $(n = 3)$. This application example is introduced in chapter 1.

- **ARMA models :** Here additive changes are modeled by

$$Y_k \;=\; \sum_{i=1}^{p} A_i Y_{k-i} \;+\; \sum_{j=0}^{q} B_j [V_{k-j} \;+\; \Upsilon(k-j, t_0)] \tag{7.1.5}$$

  The key point here is that this equation should be interpreted as additive changes in the mean of V_k, as we show later. Changes of this type are of particular interest in industrial quality control and for monitoring continuous-type technological processes.

- **State-space models :** We extensively use, both for the statistical and geometrical approaches, the following model of change :

$$\left\{ \begin{array}{rcl} X_{k+1} &=& FX_k + GU_k + W_k + \Gamma\, \Upsilon_x(k, t_0) \\ Y_k &=& HX_k + JU_k + V_k + \Xi\, \Upsilon_y(k, t_0) \end{array} \right. \tag{7.1.6}$$

  where Γ and Ξ are matrices of dimensions $n \times \tilde{n}$ and $r \times \tilde{r}$, respectively, and $\Upsilon_x(k, t_0)$ and $\Upsilon_y(k, t_0)$ are the *dynamic profiles* of the assumed changes, which are not necessarily completely known *a priori*. The assumption about Γ and Ξ depends upon the level of available *a priori* information, and is discussed in subsection 7.2. As we explain in chapter 11 again, this type of model is of key interest for sensor failure detection in inertial navigation systems.

The *a priori* information about the dynamic profiles of changes $\Upsilon(k, t_0)$ is discussed later. Typical examples of the change situations (7.1.4), (7.1.5), and (7.1.6) are discussed in the subsections 7.2.2, 7.2.3, and 7.2.4, respectively. Because all the inputs to these models have zero mean, these changes act as changes in the mean value of the observations Y_k.

7.1.2 Innovation and Redundancy

Let us now introduce the two key concepts that we use for solving these three additive change detection problems, namely *innovation* and *redundancy*. These concepts are also useful for discussing the detectability issue from both statistical and geometrical points of view.

7.1.2.1 Innovation

In sections 3.1 and 3.2, we introduced the concept of *innovation* in random processes as residual of the projection associated with the conditional expectation given the past observations. The key issue, as far as the detection of additive changes is concerned, is that the effect of these changes on the innovation, or more precisely on the residual, is also a change on its mean value. In other words, *additive changes remain additive under the transformation from observations to innovations*. The importance of the innovations as a tool for detecting additive changes comes from the computation of the likelihood ratio. Let us discuss this issue, distinguishing between two cases : constant changes and changes with dynamic profiles.

Constant changes We assume here that the change vector has a step profile :

$$\Upsilon(k, t_0) = \Upsilon\; \mathbf{1}_{\{k \geq t_0\}} \tag{7.1.7}$$

As we explained in section 3.1, the log-likelihood function of a sample of size N of observations can be written in terms of the innovations in the following manner :

$$l(\mathcal{Y}_1^N) = \ln p_Y(\mathcal{Y}_1^N) = \sum_{i=1}^{N} \ln p_\varepsilon(\varepsilon_i) \tag{7.1.8}$$

In the present Gaussian case, this results in

$$-2\, l(\mathcal{Y}_1^N) = \sum_{i=1}^{N} \ln(\det \Sigma_i) + \sum_{i=1}^{N} \varepsilon_i^T \Sigma_i^{-1} \varepsilon_i \tag{7.1.9}$$

up to an additive constant, where Σ_i is the covariance matrix of the innovation ε_i.

Example 7.1.1 (Innovation and likelihood ratio in ARMA models). *Let us now consider the* parametric *case :*

$$\ln p_\theta(\mathcal{Y}_1^N) = \sum_{i=1}^{N} \ln p(\varepsilon_i) \tag{7.1.10}$$

In the case of a stable ARMA model,

$$Y_k \;=\; \textstyle\sum_{i=1}^{p} A_i Y_{k-i} \;+\; \sum_{j=0}^{q} B_j (V_{k-j} \;+\; \theta) \tag{7.1.11}$$

where $(V_k)_k$ is an independent Gaussian sequence with zero mean and covariance matrix R, we can write

$$Y_k = \frac{B(z)}{A(z)}\,(V_k + \theta) \tag{7.1.12}$$

as in (3.2.34). Therefore, the innovation $\varepsilon_k(\theta) = Y_k - \mathbf{E}_\theta(Y_k|\mathcal{Y}_1^{k-1})$ satisfies

$$\varepsilon_k(\theta) = \frac{A(z)}{B(z)}\, Y_k - \theta \tag{7.1.13}$$

and has covariance matrix R. The first term on the right side of this equation is nothing but the output of the whitening filter with transfer function $\frac{A(z)}{B(z)}$, and is also the innovation for $\theta = 0$. Moreover, when $\theta = 0$, we use ε_k instead of $\varepsilon_k(0)$.

Now we consider the two hypotheses $\mathbf{H}_0 : \{\theta = 0\}$ and $\mathbf{H}_1 : \{\theta = \Upsilon\}$. The log-likelihood ratio in the present Gaussian case is then

$$\begin{aligned} S_1^N &= \frac{1}{2} \sum_{i=1}^{N} \left[\varepsilon_i^T R^{-1} \varepsilon_i - (\varepsilon_i - \Upsilon)^T R^{-1} (\varepsilon_i - \Upsilon) \right] \\ &= \Upsilon^T \sum_{i=1}^{N} R^{-1} \left(\varepsilon_i - \frac{\Upsilon}{2} \right) \end{aligned} \tag{7.1.14}$$

Therefore, the log-likelihood ratio is a function of the innovations ε_k. It should be obvious that these quantities are the output of the whitening filter corresponding to the ARMA model before change. Moreover, a comparison of this formula and formula (2.1.6) shows that the basic problem for additive changes in linear systems excited by Gaussian noises is the detection of changes in the mean of an independent *Gaussian sequence.*

Finally, let us add one comment about computational issues regarding the likelihood function. It results from (7.1.13) that the recursive computation of ε_k, and thus of the likelihood function, requires the knowledge of the initial conditions $Y_0, \ldots, Y_{1-p}, \varepsilon_0, \ldots, \varepsilon_{1-q}$. Several solutions exist for getting rid of these values. The simplest one consists of taking them all as zero, and waiting a sufficiently long time to cancel the effect of these zero initial data. When this is not admissible, an alternative solution consists of using special computational procedures aimed at the exact computation of the likelihood function [Gueguen and Scharf, 1980].

Changes with dynamic profiles The first motivation for considering the case of dynamic profiles of changes $\Upsilon_x(k, t_0)$ and $\Upsilon_y(k, t_0)$ is the following. Even if the additive changes, occurring on the state and observation equations of a state-space model as described in (7.1.6), are actually constant step changes, their effect on the innovations contains a dynamic profile, because of the dynamics of the system itself, as we show in subsection 7.2.4. From this fact arises the necessity to consider dynamic profiles of changes for the basic problem.

Let us thus consider the problem of detecting a change with a dynamic profile in the basic model. In this case, because the observations are independent, the above-mentioned formulas for computing the log-likelihood function and ratio are still valid, except that the ℓ-dimensional parameter $\theta_1 = \Upsilon(k, t_0)$ after change is now time-varying. Let us thus define the parameter θ_1 as a new parametric function :

$$\theta_1(k|k \geq t_0) = \Upsilon_\xi(k) = \begin{cases} 0 & \text{if } k < t_0 \\ \Upsilon_\xi(k) & \text{if } k \geq t_0 \end{cases} \tag{7.1.15}$$

where ξ is an unknown *fixed dimensional* parameter vector. It is important to note here that this parametric function contains all the available *a priori* information concerning the dynamic profile of the actual change, and thus that we automatically exclude the case of completely unknown dynamic profile of change, which obviously cannot be solved in our framework of a change between two parameter values of a parametric distribution. We refer to these dynamic profiles of changes as *parametrically unknown dynamic profiles*.

7.1.2.2 Redundancy

The concept of *redundancy* is widely used in control theory and applications and is basically related to the availability of several real or artificial measurements of the same physical quantities. *Direct redundancy* can be exploited when several sensors measure the same quantities, as in the example described in subsection 1.2.2. The natural and relevant model in this case is the *regression model* (7.1.1). *Analytical redundancy* refers to the exploitation of both physically available and computed measurements. Computed artificial measurements are built with the aid of available dynamical relationships, as summarized in a state-space model (7.1.3). As we explain in section 7.4, redundancy relationships - the so-called parity vectors - are linear combinations of either present values of outputs or present and past values of both inputs and outputs. They can be obtained through *projections* onto the orthogonal complement of the range of either the observation or the observability matrix. In these subsections, we also show that the key common concept underlying these two types of redundancy is the *residual of least-squares estimation in a regression model*. Because of the Gaussian assumption made on the noises in (7.1.1) and (7.1.3), this feature is of key interest to derive the connection with the statistical approach based upon the generalized likelihood ratio for composite hypotheses, which involves maximum likelihood - and thus least-squares - estimation of unknown states X under unfailed and failed hypotheses. Note that redundancy relationships can also be obtained through factorization of the input-output transfer function, and can be shown to be related to the Kalman filter innovation in some cases.

7.1.3 Toward the Basic Problem

As discussed before, additive changes, defined as changes in the mean value of observations, can be characterized by the fact that they result in changes in the mean values of both innovations and redundancy residuals. Because the stochastic inputs to models (7.1.1), (7.1.2), and (7.1.3) are Gaussian, the relevant problem to be investigated now is the detection of *changes in the mean value of a multidimensional Gaussian process*.

Because the main emphasis of this book is on statistical methods, and because innovation sequences are *independent*, as shown in subsection 3.1.2, from now on we use the term **basic problem** for the problem

of detecting changes in the mean value of an *independent* Gaussian sequence. The CUSUM-type and GLR solutions of this problem are recalled in subsection 7.2.1, considering various levels of *a priori* information.

It should be noted that applying the solution of the basic problem to the redundancy residuals results necessarily in a suboptimal algorithm except when the residuals are generated by the Kalman filter - viewed here as a full order state observer. The key reason for this is that when state and/or measurement noises are effectively present in the model (7.1.3), the Kalman filter is the only observer that results in an *independent* sequence of residuals, namely the innovations. Another way to look at this suboptimality issue is through the possibly nondiagonal form of the covariance matrix of the noise inputs. This point is further discussed in section 7.4.

7.2 Statistical Approach

In this section, we investigate the statistical approach to the detection of additive changes. The section is organized in the following manner. In subsection 7.2.1 we discuss several basic problems of detection of change in the mean vector of an independent identically distributed (i.i.d.) Gaussian sequence. As explained before, this problem is the central issue of this section. Then, we discuss problems of additive changes in more complex situations : regression models in subsection 7.2.2, ARMA models in subsection 7.2.3, and state-space models in subsection 7.2.4. In these subsections, we first show how these problems can be reduced to one of the basic problems analyzed in subsection 7.2.1; the additive feature of the change plays a key role in this reduction. Then we investigate the diagnosis problem in subsection 7.2.5 using a statistical point of view. The detectability issue is addressed in subsection 7.2.6. The statistical properties of the change detection algorithms presented in this section are described in section 7.3.

7.2.1 The Basic Problem

We consider here the following special but important case. Assume that we have an independent sequence $(Y_k)_{k\geq 1}$ of r-dimensional random vectors Y_k, with distribution

$$\mathcal{L}(Y_k) = \mathcal{N}(\theta, \Sigma) \tag{7.2.1}$$

Consider the on-line detection of a change in the mean vector θ. Until time $t_0 - 1$ included, the vector θ is equal to θ_0, and then from time t_0 the vector θ is equal to θ_1. Note that the parameter θ_1 is possibly a function of time, as it can be seen from the dynamic profiles often used for sensor failure detection in dynamical systems. As we discussed in section 1.4, change detection problems can be solved differently according to the various levels of the *a priori* information available about the vector parameters θ_0 and θ_1. From a practical point of view, it is useful to consider several special cases. For example, this is the case for the inertial navigation system application, as we explain in chapter 11.

7.2.1.1 Different Cases and their Motivations

The cases of interest are the following :

1. θ_0 and θ_1 are known; see figure 7.1;
2. Θ_0 and Θ_1 are separated by some known hyperplane, as depicted in figure 7.2;
3. θ_0 is known and the magnitude of the change is known, but not its direction; see figure 7.3;
4. θ_0 is known and the direction of the change is known, but not its magnitude; this change is depicted in figure 7.4;

5. θ_0 is known and the magnitude of the change has a known lower bound, but the direction of change is unknown; see figure 7.5;

6. θ_0 is unknown, but its magnitude has a known bound; the magnitude of the change has a known lower bound, but the direction of change is unknown; see figure 7.6;

7. θ_0 is known and the dynamic profile of the change is known, but not its magnitude, as in figure 7.7.

8. θ_0 is known and nothing is known about θ_1; the corresponding change is depicted in figure 7.8.

Some algorithms designed under the hypothesis of known model parameters before and after change can be used in the case where θ_0 and θ_1 are separated by an hyperplane, as depicted in figure 7.2, or by an ellipsoid, as in figures 7.5 and 7.6. We discuss this later. Situation 4 is obviously a special case of situation 7, but is nevertheless interesting in itself. In the case of *parametrically unknown* dynamic profiles, a simple transformation of the considered parametric space leads to case 8. On the other hand, let us also emphasize that, even though situation 3 is a special case of situation 6, we solve these two problems using two different ideas because basically we make quite different uses of the two levels of available *a priori* information.

There exist several possible points of view for distinguishing between the eight different cases. The first is mathematical statistics, which considers the different tools that exist for the design of decision rules and for the investigation of their properties. The second is the point of view of the applications where many cases of changes can be defined. From this point of view, the subdivision resulting from the statistical point of view may seem strange, and actually in each application, the investigator has to solve a kind of tradeoff between these two points of view for defining the final cases of interest in this particular application. This is discussed in chapter 10.

We first consider mathematical statistics, and give the motivations leading to the eight cases. From chapters 2 and 5, as depicted in figure 5.1, for a one-dimensional parameter, the following cases are of interest :

- Known θ_0 and θ_1 : This case is the simplest. It can be solved with the aid of the CUSUM algorithm for which an optimality result and the ARL function exist; the obvious extension of this case is the present case 1.
- Known middle point θ^* : This case corresponds to another traditional use of the CUSUM algorithm, which is not optimal but is practically very useful, for example, in quality control. Again the ARL function exists in this situation. The multidimensional counterpart of this case is the present case 2.

These two cases are the simplest. Let us now discuss the cases where θ_1 is partially or completely unknown. To extend the case of a one-dimensional parameter to the multidimensional one, we need to consider levels of *a priori* information for two characteristic features : the change magnitude and the change direction. Although it can be proven in a formal manner only in some particular cases, the following statement is likely to be true in general : These two features basically do not play the same role, as far as the design and the performance evaluation of the change detection algorithms are concerned. In other words, the key point is the *a priori* knowledge concerning the Kullback information between the two distributions before and after the change. This is the main motivation for the following cases :

- θ_0 is known, and $\theta_1 = \theta_0 \pm \nu$, where $\nu > 0$ is known. This leads to the two-sided CUSUM algorithm, for which the ARL function and the first-order optimality still exist. The multidimensional counterpart of this situation is the present case 3.
- θ_0 is known, and $\theta_1 = \theta_0 + \nu$, where ν is unknown. This leads to the GLR algorithm for which the ARL function exists. In the multidimensional case, this situation corresponds to the present case 4, but the corresponding change detection problem deals with a scalar parameter ν as well.

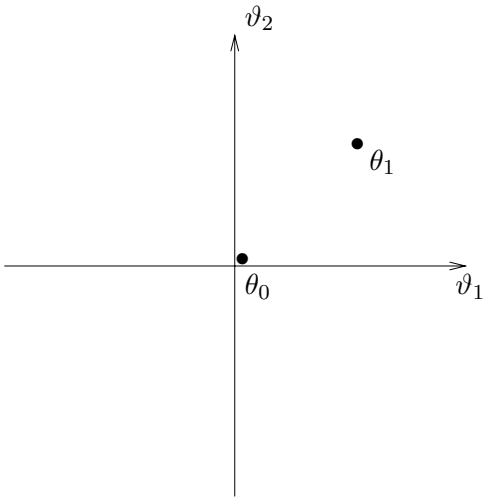

Figure 7.1 Known model parameters before and after change.

- θ_0 is known, and $|\theta_1 - \theta_0| \geq \nu_m$, where ν_m is a known minimum value of change magnitude. This leads again to the GLR algorithm, and the natural generalization is the present case 5. If we additionally assume that only an upper bound for the norm of θ_0 is known, the corresponding generalization is case 6.
- θ_0 is known and θ_1 is unknown. This leads to the GLR algorithm again, and corresponds to the present case 8.

Note that case 7 does not have any counterpart for a scalar parameter. It should be clear that this discussion is only a motivation for considering our eight cases, and in no way implies that the corresponding properties of the algorithms generalize to the multidimensional parameter case. We investigate the properties of the algorithms for each of these eight cases in section 7.3.

Now recall that the CUSUM and GLR algorithms are based upon the concept of likelihood ratio, and that the main difference between them lies in the *a priori* information about the parameter θ_1 after change. Therefore, we solve the first three problems with the aid of the CUSUM algorithm, and the remaining five with the GLR algorithm. Furthermore, let us also recall also that the CUSUM algorithm (2.2.9) is a recursive one, whereas the lack of information about θ_1 results in the fact that the GLR algorithm (2.4.32) is not recursive because of the double maximization with respect to t_0 and to θ_1. However, in the present Gaussian case, the maximization with respect to the unknown parameters (namely magnitude and direction of change) turns out to be explicit. Nevertheless, this explicit maximization cannot be obtained in a recursive manner. Finally, we also make use of an additional key idea, which we introduced in section 4.2, namely the concept of invariant SPRT which leads to another use of the *a priori* information.

Let us now investigate all the above cases. The estimation of the change time is investigated after these cases have been discussed.

7.2.1.2 Case 1 : Known θ_0 and θ_1

This situation is depicted in figure 7.1 and is addressed in [Nikiforov, 1980]. As we explained in chapter 2, in this case, the relevant algorithm is the standard CUSUM algorithm :

$$t_a \quad = \quad \min\{k \geq 1 : g_k > h\}$$

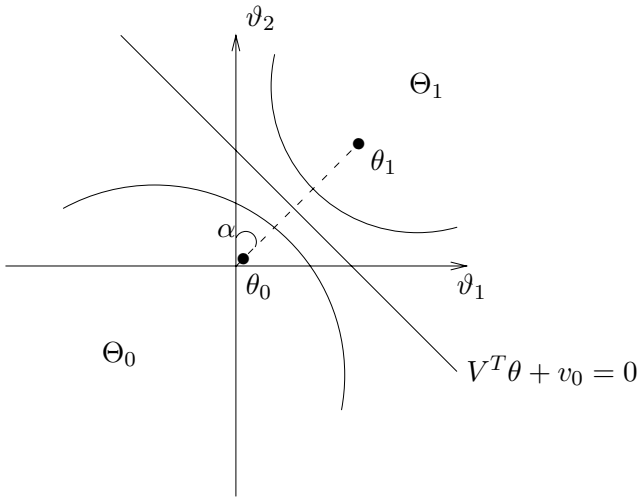

Figure 7.2 Linear discriminant function between the two parameter sets.

$$\begin{aligned} g_k &= (g_{k-1} + s_k)^+ \\ s_k &= \ln \frac{p_{\theta_1}(Y_k)}{p_{\theta_0}(Y_k)} \\ g_0 &= 0 \end{aligned} \tag{7.2.2}$$

Because of the p.d.f. of a Gaussian vector given in formula (3.1.37), we get

$$\begin{aligned} s_k &= (\theta_1 - \theta_0)^T \Sigma^{-1} \left[Y_k - \frac{1}{2}(\theta_0 + \theta_1) \right] \\ &= (\theta_1 - \theta_0)^T \Sigma^{-1} (Y_k - \theta_0) - \frac{1}{2}(\theta_1 - \theta_0)^T \Sigma^{-1} (\theta_1 - \theta_0) \end{aligned} \tag{7.2.3}$$

Note here that this case is not a new sequential change detection problem, and that the corresponding algorithm, when θ_0 and θ_1 are known, is the same in both situations of vector or scalar parameter θ, as is obvious from (7.2.3) and (2.2.20).

It should be noted that (7.2.3) can be seen as the well-known linear coherent detector, which uses the concept of *correlation* between a known constant signal and observations. Here we have correlation between the known magnitude of change and the shifted observations $Y_k - \theta_0$, which can be viewed as innovations. We will see that this concept plays a central role in all the additive change detection problems with known change direction or profile.

7.2.1.3 Case 2 : Θ_0 and Θ_1 Separated by a Hyperplane

These regions and the linear discriminant function are depicted in figure 7.2. This case is addressed in [Nikiforov, 1980]. First, let us discuss the above CUSUM algorithm (7.2.2)-(7.2.3). As we explained in chapter 5, this algorithm is optimal only when the actual parameter values are the values that are assumed in the decision function. However, we show here that, when the two sets of parameter values are separated by a hyperplane, we can design a decision function using the above simple case. Formula (7.2.3) for the increment of the CUSUM decision function can be rewritten as

$$s_k = \Upsilon^T \Sigma^{-1} (Y_k - \theta^*) \tag{7.2.4}$$

where $\theta^* = \frac{1}{2}(\theta_0 + \theta_1),\ (\theta_1 - \theta_0) = \nu\Upsilon,\ \|\Upsilon\| = 1$. As in chapter 4, let us use the following decomposition of the covariance matrix $\Sigma^{-1} = (R^{-1})^T R^{-1}$. We get

$$s_k = \check{\Upsilon}^T(\check{Y}_k - \check{\theta}) \tag{7.2.5}$$

where $\check{Y}_k = R^{-1}Y,\ \check{\theta} = R^{-1}\theta^*,\ \check{\Upsilon} = R^{-1}\Upsilon$. Let us rewrite this equation as

$$s_k = V^T Y_k + v_0 \tag{7.2.6}$$

where $V = \check{\Upsilon}$ and $v_0 = -\check{\theta}^T\check{\Upsilon}$. Taking the expectation leads to

$$\mathbf{E}_\theta(s_k) = V^T\theta + v_0 \tag{7.2.7}$$

As we discussed in chapter 2, the key property of the increment s_k of the CUSUM decision function is

$$\begin{array}{rcl} \mathbf{E}_\theta(s_k) & < & 0 \text{ for } \theta \in \Theta_0 \\ \mathbf{E}_\theta(s_k) & > & 0 \text{ for } \theta \in \Theta_1 \end{array} \tag{7.2.8}$$

Therefore, the equation of the hyperplane in the parameter space is

$$V^T\theta + v_0 = 0 \tag{7.2.9}$$

It is now obvious that, when the available *a priori* information about the parameters is not in terms of θ_0 and θ_1 but in terms of the discriminant hyperplane defined by V and v_0, the increment of the decision function is given by (7.2.6). From now on, this version of the CUSUM algorithm is called *linear CUSUM*. This term can be further explained by the following comment. In statistical pattern recognition theory [Fukunaga, 1990], the function $s(Y)$ defined by

$$s(Y) = V^T Y + v_0 \tag{7.2.10}$$

is called *linear discriminant function*. Thus, the increment s_k in (7.2.5) is nothing but a linear discriminant function.

7.2.1.4 Case 3 : Known Magnitude but Unknown Direction of Change

This situation is depicted in figure 7.3 and addressed in [Nikiforov, 1980, Nikiforov, 1983]. As in chapter 2, we solve this problem using two possible approaches. The first solution is based upon Wald's idea of weighting function, which we introduced for the one-dimensional case in subsection 2.4.2 under the name of weighted CUSUM. For the multidimensional case, this concept was extended in subsection 4.2.6 in the framework of the theory of fixed sample size invariant tests. In subsection 4.3.4 we described the weighted SPRT, which is the sequential counterpart of this test. The second solution is based upon the GLR algorithm. In this paragraph, we investigate the connection between these two solutions.

Invariant CUSUM We first describe the weighting function approach and we discuss the magnitude of a change in a vector parameter. For this purpose, we distinguish the simple case of a unit covariance matrix and the case of a general covariance matrix.

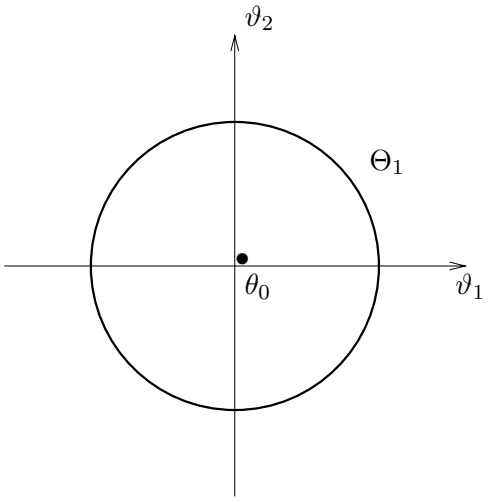

Figure 7.3 Known magnitude of change.

Unit covariance matrix In this case, the concept of magnitude is simply the Euclidian norm of the vector $\theta_1 - \theta_0$. In other words, the change detection problem is

$$\theta(k) = \begin{cases} \theta_0 & \text{when } k < t_0 \\ \theta : (\theta - \theta_0)^T(\theta - \theta_0) = b^2 & \text{when } k \geq t_0 \end{cases} \tag{7.2.11}$$

where $b > 0$.

As discussed in subsections 4.2.6 and 4.3.4, the relevant statistical decision function in this case is based upon the χ^2 distribution. This can be seen from the weighted log-likelihood ratio for the observations $Y_j, \ldots, Y_k$, which can be written as

$$\tilde{S}_j^k = -(k-j+1)\,\frac{b^2}{2} + \ln G\left[\frac{r}{2}, \frac{b^2(k-j+1)^2(\chi_j^k)^2}{4}\right] \tag{7.2.12}$$

$$(\chi_j^k)^2 = (\bar{Y}_j^k - \theta_0)^T(\bar{Y}_j^k - \theta_0) \tag{7.2.13}$$

where

$$\bar{Y}_j^k = \frac{1}{k-j+1}\sum_{i=j}^{k} Y_i \tag{7.2.14}$$

and where $G(m, x)$ is the generalized hypergeometric function. Note that this algorithm is essentially a *quadratic* detector and no longer a linear one as in the case of a known change direction. We often call this algorithm the χ^2-CUSUM algorithm, as in chapter 2.

As we explained in subsection 2.4.2, the stopping rule for the weighted CUSUM algorithm can be written as

$$t_a = \min\{k \geq 1 : \max_{1 \leq j \leq k} \tilde{S}_j^k \geq h\} \tag{7.2.15}$$

This algorithm cannot be written in a recursive manner, but another algorithm close to this one can be defined as a repeated version of the SPRT with zero lower threshold :

$$t_a = \min\{k \geq 1 : g_k \geq h\} \tag{7.2.16}$$

$$g_k = \left(\tilde{S}_{k-N_k+1}^k\right)^+$$

where

$$N_k = N_{k-1}\ \mathbf{1}_{\{g_{k-1}>0\}} + 1 \tag{7.2.17}$$

Note that the sufficient statistics $\bar{Y}$ can be written in a recursive manner as follows :

$$\begin{aligned} \bar{Y}_k &= N_k\ \bar{Y}^k_{k-N_k+1} \\ \bar{Y}_k &= \mathbf{1}_{\{g_{k-1}>0\}}\bar{Y}_{k-1} + Y_k \end{aligned} \tag{7.2.18}$$

General covariance matrix In the present case of known magnitude but unknown direction of change, the set Θ_1 is naturally defined with the aid of the concept of constant separability between the distributions before and after the change. As we explained in section 4.1, a relevant measure of distance between two probability distributions is the Kullback divergence, which, in the case of multidimensional Gaussian distributions, is simply a quadratic form (4.1.90) of the parameters with respect to the inverse covariance matrix Σ^{-1}. In other words, the set Θ_1 corresponds to an ellipsoid, and not to a sphere. Thus, the change detection problem can be stated as follows :

$$\theta(k) = \begin{cases} \theta_0 & \text{when } k < t_0 \\ \theta : (\theta-\theta_0)^T\Sigma^{-1}(\theta-\theta_0) = b^2 & \text{when } k \geq t_0 \end{cases} \tag{7.2.19}$$

where $b > 0$.

As we explained in subsection 4.3.4, the log-likelihood ratio $\tilde{S}^k_j$ can be written as in the previous case of a unit covariance matrix, except that χ should be now defined by

$$(\chi^k_j)^2 = (\bar{Y}^k_j - \theta_0)^T\Sigma^{-1}(\bar{Y}^k_j - \theta_0) \tag{7.2.20}$$

GLR algorithm We now describe the GLR solution. Recall that we defined the likelihood ratio test in example 4.2.6. Let us first investigate the simplest case of a unit covariance matrix.

Unit covariance matrix In this case, the GLR algorithm can be written as

$$\begin{aligned} t_a &= \min\{k \geq 1 : g_k > h\} \\ g_k &= \max_{1\leq j\leq k}\ \sup_{\|\theta-\theta_0\|=b} S^k_j \\ S^k_j &= \ln \frac{\prod_{i=j}^k p_\theta(Y_i)}{\prod_{i=j}^k p_{\theta_0}(Y_i)} \end{aligned} \tag{7.2.21}$$

We follow example 4.2.6 to obtain

$$\begin{aligned} \ln \frac{\sup_{\|\theta-\theta_0\|=b}\prod_{i=j}^k p_\theta(Y_i)}{\prod_{i=j}^k p_{\theta_0}(Y_i)} &= \ln \sup_{\|\theta-\theta_0\|=b}\prod_{i=j}^k p_\theta(Y_i) - \ln\prod_{i=j}^k p_{\theta_0}(Y_i) \\ &= -\frac{k-j+1}{2}\ \left(\|\bar{Y}^k_j - \theta_0\| - b\right)^2 + \frac{k-j+1}{2}\|\bar{Y}^k_j - \theta_0\|^2 \\ &= (k-j+1)\left(b\|\bar{Y}^k_j - \theta_0\| - \frac{b^2}{2}\right) \\ &= (k-j+1)\left(b\chi^k_j - \frac{b^2}{2}\right) \end{aligned} \tag{7.2.22}$$

where χ^k_j is defined in (7.2.13).

General covariance matrix The solution then consists of transforming this change detection problem into the previous one. Therefore, in this case, the GLR decision function is the same as above except that χ_j^k should be taken as defined in (7.2.20).

Connection between the two solutions Let us now investigate the connection between the invariant CUSUM and GLR solutions in the asymptotic situation where the threshold h goes to infinity. From the previous discussion, it is obvious that, without loss of generality, we can assume unit covariance matrix. The stopping rule of the invariant CUSUM can be interpreted as follows. At time t_a, there exists an integer j_0 such that the following inequality holds :

$$\tilde{S}_{j_0}^{t_a} = -(t_a - j_0 + 1)\,\frac{b^2}{2} + \ln G\left(\frac{r}{2}, \frac{b^2\|\check{S}_{j_0}^{t_a}\|^2}{4}\right) \geq h \tag{7.2.23}$$

$$\text{where } \check{S}_j^k = \sum_{i=j}^{k}(Y_i - \theta_0) \tag{7.2.24}$$

From this, we deduce the equation for the generatrix $\tilde{c}_n$ of the stopping surface of the invariant CUSUM :

$$h = -n\,\frac{b^2}{2} + \ln G\left(\frac{r}{2}, \frac{b^2\tilde{c}_n^2}{4}\right) \tag{7.2.25}$$

This is the direct extension of the U-mask and V-mask which we discussed in chapter 2. Now similar computations lead to the generatrix $\hat{c}_n$ of the stopping surface of the GLR algorithm :

$$h = -n\,\frac{b^2}{2} + b\hat{c}_n \tag{7.2.26}$$

which does not depend upon r. Let us compare these two generatrix. For this purpose, rewrite the function G with the aid of the so-called confluent hypergeometric function M [Ghosh, 1970] :

$$G(d, x) = e^{-2\sqrt{x}} M\left(d - \frac{1}{2}, 2d - 1; 4\sqrt{x}\right) \tag{7.2.27}$$

for $x > 0$ and $d > \frac{1}{2}$ (and thus is nonvalid in the one-dimensional case). The function M has the following asymptotic expansion for $x \to \infty$:

$$M(d, d'; x) = \frac{\Gamma(d')}{\Gamma(d)}\; e^x x^{d-d'}\left[1 + O\left(\frac{1}{x}\right)\right] \tag{7.2.28}$$

for fixed d and d'. Because $G(d, x)$ is an increasing function of x for $x > 0$, for each n and when h goes to infinity, the equation for the generatrix $\tilde{c}_n$ can be approximated by

$$h = -n\,\frac{b^2}{2} + b\tilde{c}_n - \frac{r-1}{2}\ln(2b\tilde{c}_n) + \ln\frac{\Gamma(r-1)}{\Gamma\left(\frac{r-1}{2}\right)} + \ln\left[1 + O\left(\frac{1}{2b\tilde{c}_n}\right)\right] \tag{7.2.29}$$

The comparison with the equation for the generatrix $\hat{c}_n$ leads, for each n and when h goes to infinity, to the following relations :

$$\hat{c}_n < \tilde{c}_n < \hat{c}_n\left[1 + O\left(h^{-1}\ln h\right)\right] \tag{7.2.30}$$

For finite h, the difference $\tilde{c}_n - \hat{c}_n$ depends upon the particular values of the change magnitude b and the dimension r. But for many cases of practical interest, this difference turns out to be negligible. Therefore, the invariant CUSUM and the GLR algorithms basically have the same behavior.

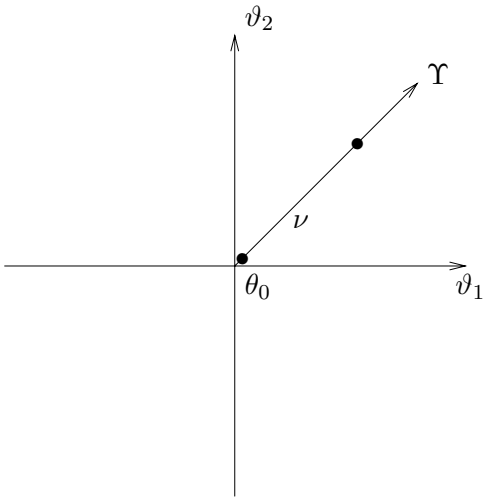

Figure 7.4 Known direction of change.

7.2.1.5 Case 4 : Known Direction but Unknown Magnitude of Change

In this case, which is depicted in figure 7.4, we assume that the change detection problem is

$$\theta(k) = \begin{cases} \theta_0 & \text{when } k < t_0 \\ \theta_0 + \nu\Upsilon & \text{when } k \geq t_0 \end{cases} \tag{7.2.31}$$

where Υ is the *unit* vector of the change direction. The corresponding GLR decision function is thus

$$g_k = \max_{1\leq j\leq k} \ln \frac{\sup_\nu \prod_{i=j}^k p_{\theta_0+\nu\Upsilon}(Y_i)}{\prod_{i=j}^k p_{\theta_0}(Y_i)} \tag{7.2.32}$$

It results that

$$g_k = \max_{1\leq j\leq k} (k-j+1) \left[\hat{\nu}_k(j) \Upsilon^T \Sigma^{-1} (\bar{Y}_j^k - \theta_0) - \frac{\hat{\nu}_k^2(j)}{2} \Upsilon^T \Sigma^{-1} \Upsilon \right] \tag{7.2.33}$$

where

$$\hat{\nu}_k(j) = \frac{\Upsilon^T \Sigma^{-1} (\bar{Y}_j^k - \theta_0)}{\Upsilon^T \Sigma^{-1} \Upsilon} \tag{7.2.34}$$

is the estimate, at time k, of the magnitude of the change occurring at time j. Note that (7.2.33) is the relevant extension of (7.2.3). It can be seen as a matched filtering operation between the known change direction Υ and the mean value of the shifted observations.

Note that the link between the present algorithm and the CUSUM solution to case 2 is the same as the link between GLR and CUSUM algorithms in the case of a scalar parameter, which we discussed in chapter 5.

7.2.1.6 Case 5 : Known θ_0 and Lower Bound for the Magnitude but Unknown Change Direction

This case is depicted in figure 7.5, and is a special case of situation 6, namely it corresponds to the situation where $a = 0$.

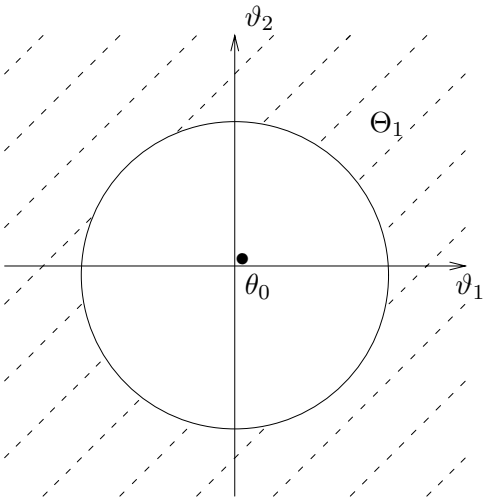

Figure 7.5 Known lower bound for the magnitude of θ_1.

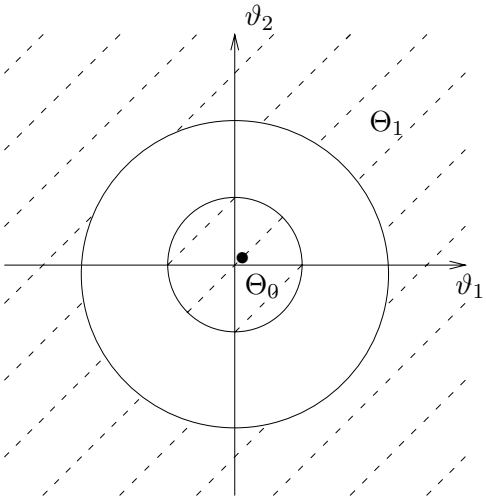

Figure 7.6 Known upper bound for the magnitude of θ_0 and known lower bound for the magnitude of θ_1.

7.2.1.7 Case 6 : Known Upper Bound for θ_0 and Lower Bound for the Change Magnitude

In this case, we assume that

$$\theta(k) = \begin{cases} \theta \in \Theta_0 & \text{when } k < t_0 \\ \theta \in \Theta_1 & \text{when } k \geq t_0 \end{cases} \tag{7.2.35}$$

This situation is depicted in figure 7.6. We derive the change detection algorithm by using a generalization of the idea of the one-dimensional GLR algorithm introduced in chapter 2 and the theory of the likelihood ratio test discussed in section 4.1.

Unit covariance matrix First, let us discuss the special case of a unit covariance matrix $\Sigma = I$. In this case, the above change detection problem is equivalent to the following :

$$\theta(k) = \begin{cases} \theta : (\theta - \theta_0)^T(\theta - \theta_0) \leq a^2 & \text{when } k < t_0 \\ \theta : (\theta - \theta_0)^T(\theta - \theta_0) \geq b^2 & \text{when } k \geq t_0 \end{cases} \tag{7.2.36}$$

where $a < b$. The corresponding GLR algorithm is then

$$\begin{aligned} t_a &= \min\{k \geq 1 : g_k \geq h\} \\ g_k &= \max_{1 \leq j \leq k} S_j^k \\ S_j^k &= \ln \frac{\sup_{\|\theta - \theta_0\| \geq b} \prod_{i=j}^k p_\theta(Y_i)}{\sup_{\|\theta - \theta_0\| \leq a} \prod_{i=j}^k p_\theta(Y_i)} \end{aligned} \tag{7.2.37}$$

From the formula (4.2.67) of section 4.1,

$$\frac{2}{k-j+1} S_j^k = \begin{cases} -(\|\bar{Y}_j^k - \theta_0\| - b)^2 & \text{when} \quad \|\bar{Y}_j^k - \theta_0\| < a \\ -(\|\bar{Y}_j^k - \theta_0\| - b)^2 + (\|\bar{Y}_j^k - \theta_0\| - a)^2 & \text{when} \quad a \leq \|\bar{Y}_j^k - \theta_0\| \leq b \\ +(\|\bar{Y}_j^k - \theta_0\| - a)^2 & \text{when} \quad \|\bar{Y}_j^k - \theta_0\| > b \end{cases}$$

General covariance matrix Let us discuss now the general case : $\mathcal{L}(Y) = \mathcal{N}(\theta, \Sigma)$. In this case, the change detection problem is

$$\theta(k) = \begin{cases} \theta : (\theta - \theta_0)^T\Sigma^{-1}(\theta - \theta_0) \leq a^2 & \text{when } k < t_0 \\ \theta : (\theta - \theta_0)^T\Sigma^{-1}(\theta - \theta_0) \geq b^2 & \text{when } k \geq t_0 \end{cases} \tag{7.2.38}$$

where $a < b$ again. As we explained for case 3 above, the log-likelihood ratio of the GLR algorithm (7.2.37) for the change detection problem (7.2.36) can be rewritten as

$$\frac{2}{k-j+1} S_j^k = \begin{cases} -(\chi_j^k - b)^2 & \text{when} \quad \chi_j^k < a \\ -(\chi_j^k - b)^2 + (\chi_j^k - a)^2 & \text{when} \quad a \leq \chi_j^k \leq b \\ +(\chi_j^k - a)^2 & \text{when} \quad \chi_j^k > b \end{cases} \tag{7.2.39}$$

For the change detection problem (7.2.38), we use the formula

$$\chi_j^k = [(\bar{Y}_j^k - \theta_0)^T\Sigma^{-1}(\bar{Y}_j^k - \theta_0)]^{\frac{1}{2}} \tag{7.2.40}$$

This algorithm can also be derived directly from the solution of the following optimization problem :

$$\sup_{\theta^T\Sigma^{-1}\theta \geq b^2} \left[-\frac{N}{2}(\theta - \bar{Y}_j^k)^T\Sigma^{-1}(\theta - \bar{Y}_j^k) \right] \tag{7.2.41}$$

or

$$\sup_{\theta^T\Sigma^{-1}\theta \leq a^2} \left[-\frac{N}{2}(\theta - \bar{Y}_j^k)^T\Sigma^{-1}(\theta - \bar{Y}_j^k) \right] \tag{7.2.42}$$

Using Lagrange's method, we get the solution of (7.2.41) :

$$\sup_{\theta^T\Sigma^{-1}\theta \geq b^2} \left[-\frac{N}{2}(\theta - \bar{Y}_j^k)^T\Sigma^{-1}(\theta - \bar{Y}_j^k) \right] = \begin{cases} 0 & \text{when} \quad \chi_j^k > b \\ -\frac{N}{2}(\chi_j^k - b)^2 & \text{when} \quad \chi_j^k \leq b \end{cases} \tag{7.2.43}$$

The maximization (7.2.42) can be solved in the same way. From these, (7.2.39) results.

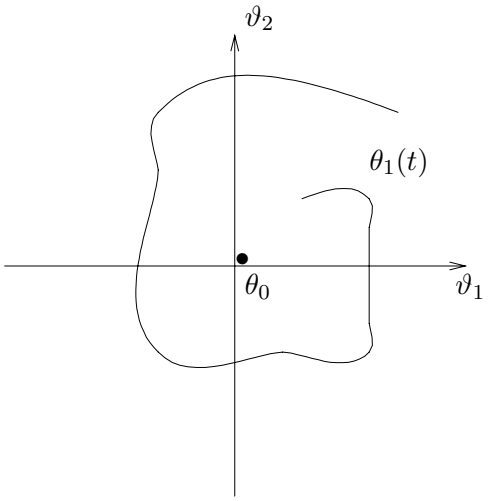

Figure 7.7 Known profile of change.

7.2.1.8 Case 7 : Known θ_0 and Dynamic Profile of the Change, but Unknown Magnitude

In this case, which is depicted in figure 7.7, we consider the following change detection problem :

$$\theta(k) = \begin{cases} \theta_0 & \text{when } k < t_0 \\ \theta_0 + \nu\Upsilon(k, t_0) & \text{when } k \geq t_0 \end{cases} \tag{7.2.44}$$

where $\Upsilon(k, t_0)$ is the known *dynamic profile* of the change and ν is the unknown change magnitude. A trivial example of profile is $\Upsilon(k, t_0) = \mathbf{1}_{\{k \geq t_0\}}\ \Upsilon$, where $\|\Upsilon\| = 1$, which is used in all the remaining paragraphs of this subsection. The corresponding GLR algorithm is

$$g_k = \max_{1 \leq j \leq k} \ln \frac{\sup_\nu \prod_{i=j}^k p_{\theta_0 + \nu\Upsilon(k,t_0)}(Y_i)}{\prod_{i=j}^k p_{\theta_0}(Y_i)} \tag{7.2.45}$$

As in the case where the change direction is known, but not the magnitude, and where the GLR algorithm is given by (7.2.32), it is straightforward to obtain

$$g_k = \max_{1 \leq j \leq k} \left[\hat{\nu}_k(j) \sum_{i=j}^k \Upsilon(i,j)^T \Sigma^{-1}(Y_i - \theta_0) - \frac{\hat{\nu}_k^2(j)}{2} \sum_{i=j}^k \Upsilon(i,j)^T \Sigma^{-1} \Upsilon(i,j) \right] \tag{7.2.46}$$

where

$$\hat{\nu}_k(j) = \frac{\sum_{i=j}^k \Upsilon(i,j)^T \Sigma^{-1}(Y_i - \theta_0)}{\sum_{i=j}^k \Upsilon(i,j)^T \Sigma^{-1} \Upsilon(i,j)} \tag{7.2.47}$$

is the estimate at time k of the change magnitude, assuming a change at time j. Note that we again have, as in (7.2.32), a *correlation* operation between the profile of the change and the shifted observations.

7.2.1.9 Case 8 : Known θ_0 but Unknown θ_1

In this case, depicted in figure 7.8, the change detection problem statement is as follows :

$$\theta(k) \begin{cases} = \theta_0 & \text{when } k < t_0 \\ \neq \theta_0 & \text{when } k \geq t_0 \end{cases} \tag{7.2.48}$$

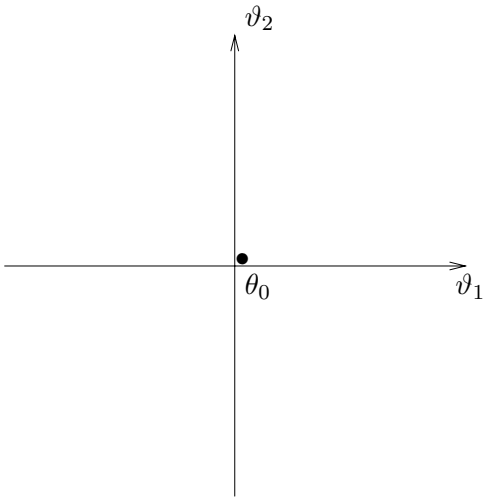

Figure 7.8 Unknown parameter after change.

As is obvious from this figure, the case where nothing is known about θ_1 can be considered the limit of the cases depicted in figures 7.5 and 7.6, in which we assume a lower bound for the change magnitude. From a more formal point of view, we consider that the present case (7.2.48) of unknown θ_1 is the limit, when $a = b = 0$, of the change detection problem (7.2.38).

The GLR solution to this problem is based upon the following decision function :

$$g_k = \max_{1\leq j\leq k} \ln \frac{\sup_\theta \prod_{i=j}^k p_\theta(Y_i)}{\prod_{i=j}^k p_{\theta_0}(Y_i)} \tag{7.2.49}$$

and considering the limit case of (7.2.39) when $a = b = 0$, we get

$$g_k = \max_{1\leq j\leq k} \frac{k-j+1}{2}(\bar{Y}_j^k - \theta_0)^T \Sigma^{-1}(\bar{Y}_j^k - \theta_0) = \max_{1\leq j\leq k} \frac{k-j+1}{2}(\chi_j^k)^2 \tag{7.2.50}$$

7.2.1.10 Geometrical Interpretation of the CUSUM and GLR Algorithms

The geometrical interpretation of the CUSUM and GLR algorithms is now given in the case of a unit covariance matrix, using figures 7.9 and 7.10. We discuss cases 3 and 8, and the other cases can obviously be deduced as a superposition of these two cases. Here we continue to investigate the connections between the invariant CUSUM and the GLR algorithms in case 3.

Case 3 We have shown before that, in this case, the CUSUM and GLR algorithms have asymptotically strong connection. Furthermore, these algorithms can be interpreted as extended stopping times associated with parallel open-ended tests, as we explained in section 2.2. Let us begin our discussion with the GLR algorithm. The generatrix $\hat{c}_n$ of the stopping surface is

$$\hat{c}_n = \frac{h}{b} + \frac{nb}{2} \tag{7.2.51}$$

and is depicted in figure 7.9. We call this cone a multidimensional V-mask because it is a natural generalization of the V-mask used for the geometrical interpretation of the CUSUM algorithm in the one-dimensional case in chapter 2. It results from equation 7.2.25 that for the invariant CUSUM the the surface of revolution is more complex than a simple cone.

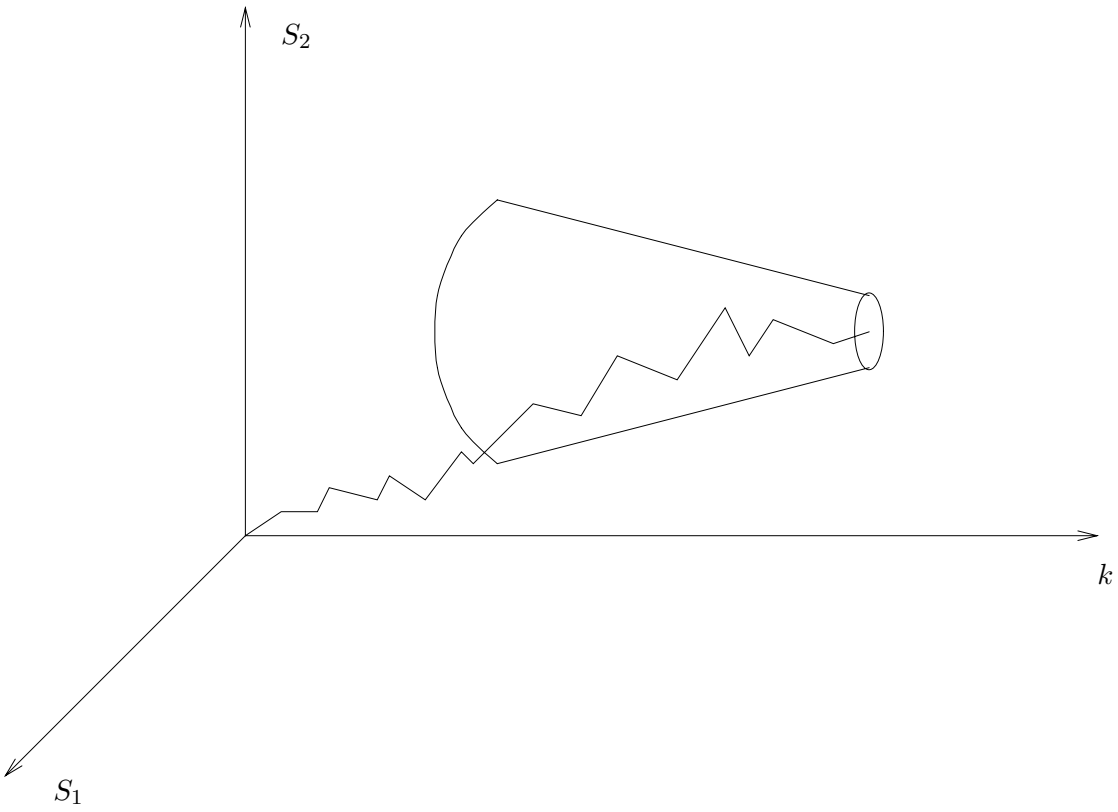

Figure 7.9 Multidimensional V-mask for the CUSUM algorithm, case 3 : (S_1, S_2) are the coordinates of $\check{S}_1^k = \sum_{i=1}^{k}(Y_i - \theta_0)$.

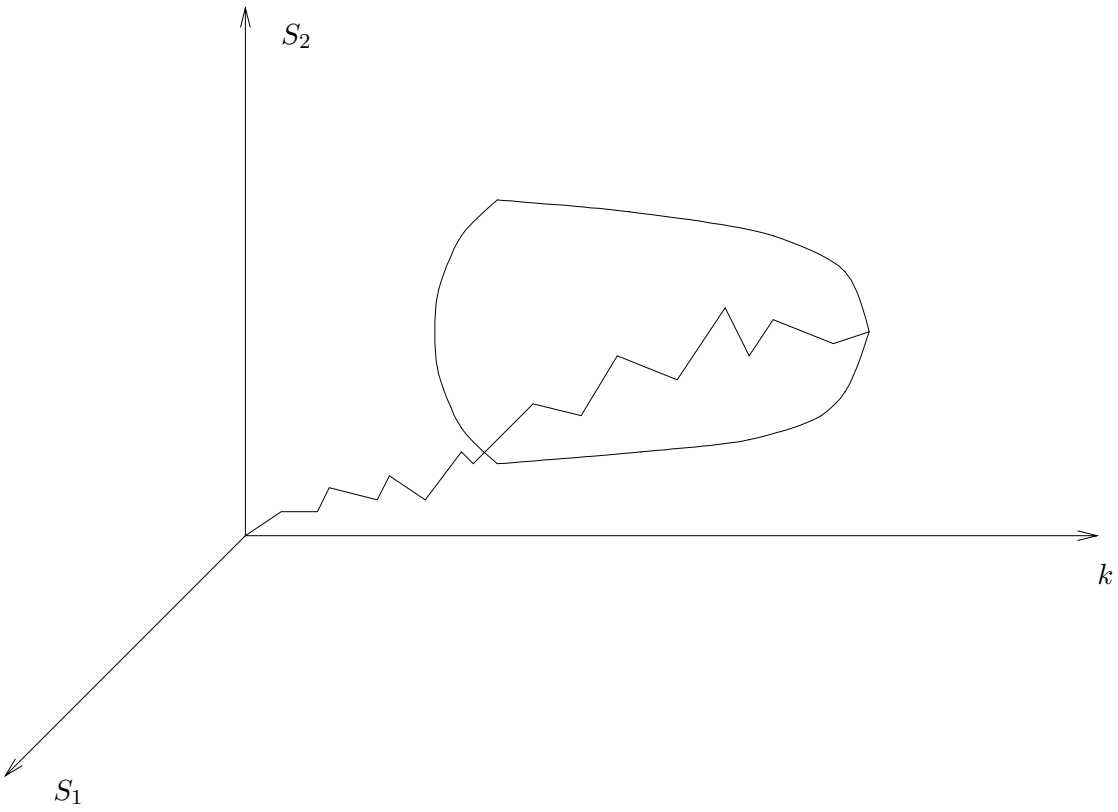

Figure 7.10 Multidimensional U-mask for the GLR algorithm, case 8 : (S_1, S_2) are the coordinates of $\check{S}_1^k = \sum_{i=1}^{k}(Y_i - \theta_0)$.

Case 8 We have shown before that, in this case, the GLR algorithm can be interpreted as an extended stopping time associated with parallel open-ended tests, as we explained in subsection 2.4.3. In the corresponding decision rule, the alarm is set the first time k at which there exists a time j_0 such that

$$\hat{S}_{j_0}^k = \frac{k - j_0 + 1}{2} \ \|\bar{Y}_{j_0}^k - \theta_0\|^2 \geq h \tag{7.2.52}$$

We can write

$$\hat{S}_1^n = \frac{1}{2n} \ \|\check{S}_1^n\|^2 \tag{7.2.53}$$

The generatrix c_n of the stopping surface thus in this case has the following equation :

$$\|c_n\| = \sqrt{2nh} \tag{7.2.54}$$

and is depicted in figure 7.10. We call this paraboloid a multidimensional U-mask, because it is a natural generalization of the U-mask used for the geometrical interpretation of the GLR algorithm in the one-dimensional case in chapter 2.

7.2.1.11 Estimation of the Change Time

As in subsections 2.2.3 and 2.4.3 and section 2.6, the unknown change time is estimated with the aid of the following maximization of the likelihood function :

$$(\tilde{j}, \tilde{\theta}_1) = \arg \max_{1 \leq j \leq t_a} \ \sup_{\theta_1} \ \hat{S}_j^{t_a}(\theta_1) \tag{7.2.55}$$

and the estimated change time is $\hat{t}_0 = \tilde{j}$.

From now on, we consider cases 1, 3, 6, and 7 of the basic problem. In the case of state-space models, we also consider case 8. The solutions to the other cases can be obtained in similar ways. The basic problems that we investigate in this subsection serve as tools for solving more complex additive change detection problems in the three next subsections. As we explained in the introduction to this chapter, we basically solve these problems by first using the transformation from observations to innovations, and then solving the corresponding additive change detection problem for the innovations using the solution to the basic problems in the Gaussian independent case.

7.2.2 Regression Models

We now investigate additive changes in an r-dimensional independent sequence $(Y_k)_k$, which can be described as the output of a regression model :

$$Y_k = HX_k + V_k \tag{7.2.56}$$

where X is an unknown vector with dimension $n < r$, and $(V_k)_k$ is a white noise sequence with positive definite covariance matrix R. The observation matrix H is assumed to be a matrix with full rank n. In the sequel, we use the following factorization of the covariance matrix R :

$$\begin{aligned} R &= AA^T \\ R^{-1} &= A^{-T}A^{-1} \end{aligned} \tag{7.2.57}$$

As far as the additive changes in this model are concerned, we assume the following :

$$Y_k = HX_k + V_k + \Upsilon(k, t_0) \tag{7.2.58}$$

where $\Upsilon(k, t_0)$ is the vector of change. For many practical situations, it is convenient to assume that, in this case, $\theta_0 = 0$ and $\theta_1 = \Upsilon(k, t_0)$ for $k \geq t_0$. Recall that, as we explained in section 7.1, in case of dynamic profiles of changes, we consider only parametrically unknown dynamic profiles.

From now on, we follow subsection 7.2.1, considering different levels of knowledge about the parameter after change. These different levels of *a priori* information result in different types of constraints on $\Upsilon(k, t_0)$.

7.2.2.1 Case 1 : Known θ_1

In this case, we consider the following model :

$$Y_k = HX_k + V_k + \Upsilon \, \mathbf{1}_{\{k \geq t_0\}} \tag{7.2.59}$$

where Υ is the known vector of change, namely $\Upsilon = \theta_1$. The specificity of this change detection problem with respect to the above first basic problem lies in the fact that the vector X is unknown. Because the dimension r of the observations is greater than the dimension n of the unknown state X, we can make use of this *redundancy* to solve the detection problem. As we explain in subsection 4.2.8, the standard statistical approach in this case, namely the minmax approach, is equivalent to the GLR approach, and thus consists of replacing the unknown (nuisance) values by their maximum likelihood estimates. Therefore, and because the sequence of noise $(V_k)_k$ here is independent, we take, as a solution to the present change detection problem, the CUSUM algorithm (7.2.2) where the log-likelihood ratio can be written as

$$s_k = \ln \frac{\sup_{X_k} p_\Upsilon(Y_k|X_k)}{\sup_{X_k} p_0(Y_k|X_k)} \tag{7.2.60}$$

and thus as

$$s_k = \ln \frac{p_\Upsilon(Y_k|\hat{X}_{k,\Upsilon})}{p_0(Y_k|\hat{X}_{k,0})} \tag{7.2.61}$$

where $\hat{X}_{k,\Upsilon}$ and $\hat{X}_{k,0}$ are the maximum likelihood estimates of X under both hypotheses. It is known [Seber, 1977] that under the linear and Gaussian assumptions, the maximum likelihood estimate coincides with the least-squares estimate, which can be written as

$$\hat{X}_{k,\Upsilon} = (H^T R^{-1} H)^{-1} H^T R^{-1} (Y_k - \Upsilon) \tag{7.2.62}$$

The residual corresponding to this estimation is

$$\begin{aligned}
e_k &= Y_k - H\hat{X}_{k,0} \\
&= [I_r - H(H^T R^{-1} H)^{-1} H^T R^{-1}] Y_k \\
&= [I_r - H(H^T R^{-1} H)^{-1} H^T A^{-T} A^{-1}] A A^{-1} Y_k \\
&= [A - H(H^T R^{-1} H)^{-1} H^T A^{-T}] A^{-1} Y_k \\
&= A[I_r - A^{-1} H(H^T R^{-1} H)^{-1} H^T A^{-T}] A^{-1} Y_k \\
&= A P_H A^{-1} Y_k \qquad (7.2.63) \\
&= \tilde{P}_H Y_k \qquad (7.2.64)
\end{aligned}$$

where

$$P_H = I_r - A^{-1} H(H^T R^{-1} H)^{-1} H^T A^{-T} \tag{7.2.65}$$

Note that P_H is idempotent and symmetric, and thus corresponds to an orthogonal projection, while $\tilde{P}_H$ is idempotent but not symmetric, except if R is diagonal. Nevertheless, $\tilde{P}_H$ satisfies $\tilde{P}_H H = 0$.

The corresponding log-likelihood function under the no change hypothesis is thus

$$\begin{aligned} -\ln p_0(Y_k|\hat{X}_{k,0}) &= \frac{1}{2}(Y_k - H\hat{X}_{k,0})^T R^{-1}(Y_k - H\hat{X}_{k,0}) \\ &= \frac{1}{2}e_k^T R^{-1} e_k \\ &= \frac{1}{2}Y_k^T A^{-T} P_H A^{-1} Y_k \end{aligned} \tag{7.2.66}$$

Similarly, we have

$$-\ln p_\Upsilon(Y_k|\hat{X}_{k,\Upsilon}) = \frac{1}{2}(Y_k - H\hat{X}_{k,\Upsilon} - \Upsilon)^T A^{-T} P_H A^{-1}(Y_k - H\hat{X}_{k,\Upsilon} - \Upsilon) \tag{7.2.67}$$

Thus, the log-likelihood ratio can be written as

$$s_k = \Upsilon^T A^{-T} P_H A^{-1} Y_k - \frac{1}{2}\Upsilon^T A^{-T} P_H A^{-1}\Upsilon \tag{7.2.68}$$

Now, straightforward computations give

$$\begin{aligned} A^{-T} P_H A^{-1} &= (AP_H A^{-1})^T A^{-T} A^{-1}(AP_H A^{-1}) \\ &= (AP_H A^{-1})^T R^{-1}(AP_H A^{-1}) \\ &= \tilde{P}_H^T R^{-1} \tilde{P}_H \end{aligned} \tag{7.2.69}$$

Thus, the log-likelihood ratio can be equivalently rewritten as

$$s_k = \rho^T R^{-1} e_k - \frac{1}{2}\rho^T R^{-1}\rho \tag{7.2.70}$$

where

$$\rho = \tilde{P}_H \Upsilon \tag{7.2.71}$$

is the *signature* of the additive change on the residual e. In other words, the log-likelihood ratio results in nothing but a *correlation* between the innovation and the signature of the change on the innovation. Note that this is the relevant counterpart of (7.2.3).

In the other cases of change, we make use of the following result. From definition (4.1.42) and from (7.2.68), we deduce that the Kullback divergence between the two regression models before and after change is

$$\mathbf{J}(0, \Upsilon) = \Upsilon^T A^{-T} P_H A^{-1}\Upsilon = \rho^T R^{-1}\rho \tag{7.2.72}$$

Note that the matrix $A^{-T} P_H A^{-1}$ is not full rank because of the projection P_H. As we show in subsection 7.2.6, this fact is a central issue of detectability. And thus, before using any of the algorithms described below for a particular change Υ, it is necessary to investigate the rank of this matrix.

7.2.2.2 Case 3 : Known Magnitude but Unknown Direction of Change

We now begin to discuss cases of composite hypotheses. The background for solving these testing problems with nuisance parameters was described in subsection 4.2.8. As discussed above for the corresponding basic problem, the model that we choose in the case of known magnitude and unknown direction of change comes basically from a constant Kullback divergence between the models before and after the change. Therefore, in this case, we assume the following model after change :

$$Y_k = HX_k + V_k + \Upsilon(k, t_0) \tag{7.2.73}$$

where Υ is such that

$$\Upsilon(k,t_0) = \begin{cases} = \Upsilon_0 = 0 & \text{when} \quad k < t_0 \\ \Upsilon \in \boldsymbol{\Upsilon_1} = \{\Upsilon : \Upsilon^T A^{-T} P_H A^{-1} \Upsilon = b^2\} & \text{when} \quad k \geq t_0 \end{cases} \tag{7.2.74}$$

As for the corresponding basic problem, we use the concept of invariant SPRT for solving this change detection problem. Thus, the resulting algorithm is given by (7.2.16), (7.2.17), (7.2.12), and (7.2.20), where (7.2.20) is modified with the aid of the transformation from observations to innovations. It results from the log-likelihood function (7.2.66) that (7.2.20) is replaced by

$$(\chi^k_{k-N_k+1})^2 = (\bar{e}_k)^T R^{-1} \bar{e}_k \tag{7.2.75}$$

$$= (\bar{Y}_k)^T A^{-T} P_H A^{-1} \bar{Y}_k \tag{7.2.76}$$

where $\bar{e}_k = \frac{1}{N_k} \check{e}_k$ is the mean of the N_k last residuals and can be recursively computed as

$$\check{e}_k = \mathbf{1}_{\{g_{k-1}>0\}}\ \check{e}_{k-1} + e_k \tag{7.2.77}$$

7.2.2.3 Case 6 : Known Bounds for the Parameters Before and After Change, but Unknown Direction

As for the corresponding basic problem, the model we choose in the case of known bounds for the parameters and unknown direction of change involves ellipsoids which come from a constant Kullback divergence between the models before and after change. Therefore, in this case, we assume the following model :

$$Y_k = HX_k + V_k + \Upsilon(k,t_0) \tag{7.2.78}$$

where Υ is such that

$$\Upsilon(k,t_0) \in \begin{cases} \boldsymbol{\Upsilon_0} = \{\Upsilon : \Upsilon^T A^{-T} P_H A^{-1} \Upsilon \leq a^2\} & \text{when} \quad k < t_0 \\ \boldsymbol{\Upsilon_1} = \{\Upsilon : \Upsilon^T A^{-T} P_H A^{-1} \Upsilon \geq b^2\} & \text{when} \quad k \geq t_0 \end{cases} \tag{7.2.79}$$

As above, we solve this problem with the aid of the GLR algorithm :

$$\begin{aligned} t_a &= \min\{k \geq 1 : g_k > h\} \\ g_k &= \max_{1 \leq j \leq k} S_j^k \\ S_j^k &= \ln \frac{\sup_{\Upsilon \in \boldsymbol{\Upsilon}_1} \prod_{i=j}^k \sup_X p_\theta(Y_i|X)}{\sup_{\Upsilon \in \boldsymbol{\Upsilon}_0} \prod_{i=j}^k \sup_X p_\theta(Y_i|X)} \end{aligned} \tag{7.2.80}$$

This results in S_j^k, defined in (7.2.39), where χ_j^k is as in (7.2.75).

7.2.2.4 Case 7 : Known θ_0, Dynamic Profile of the Change, and Unknown Magnitude

In this case, we assume the following model of change :

$$Y_k = HX_k + V_k + \nu \Upsilon(k,t_0) \tag{7.2.81}$$

where $\Upsilon(k,t_0)$ is the known dynamic profile of the change and ν its unknown magnitude. The corresponding GLR decision function is

$$g_k = \max_{1 \leq j \leq k} \ln \frac{\sup_\nu \prod_{i=j}^k \sup_X p_{\theta_0 + \nu \Upsilon(k,t_0)}(Y_i|X)}{\prod_{i=j}^k \sup_X p_{\theta_0}(Y_i|X)} \tag{7.2.82}$$

Straightforward computations give

$$g_k = \max_{1\leq j\leq k} \left[\hat{\nu}_k(j) \sum_{i=j}^{k} \Upsilon(i,j)^T A^{-T} P_H A^{-1} Y_i - \frac{\hat{\nu}_k^2(j)}{2} \sum_{i=j}^{k} \Upsilon(i,j)^T A^{-T} P_H A^{-1} \Upsilon(i,j) \right] \tag{7.2.83}$$

where

$$\hat{\nu}_k(j) = \frac{\sum_{i=j}^{k} \Upsilon(i,j)^T A^{-T} P_H A^{-1} Y_i}{\sum_{i=j}^{k} \Upsilon(i,j)^T A^{-T} P_H A^{-1} \Upsilon(i,j)} \tag{7.2.84}$$

is the estimate at time k of the change magnitude, assuming a change at time j. Using the property (7.2.69), we can rewrite g_k and $\hat{\nu}$ in the same manner as s_k in (7.2.70) :

$$g_k = \max_{1\leq j\leq k} \left[\hat{\nu}_k(j) \sum_{i=j}^{k} \rho(i,j)^T R^{-1} e_i - \frac{\hat{\nu}_k^2(j)}{2} \sum_{i=j}^{k} \rho(i,j)^T R^{-1} \rho(i,j) \right] \tag{7.2.85}$$

and

$$\hat{\nu}_k(j) = \frac{\sum_{i=j}^{k} \rho(i,j)^T R^{-1} e_i}{\sum_{i=j}^{k} \rho(i,j)^T R^{-1} \rho(i,j)} \tag{7.2.86}$$

where

$$\rho(i,k) = \tilde{P}_H \Upsilon(i,k) \tag{7.2.87}$$

is the projection of the change Υ on the residual e. Note that we again have, as in (7.2.32), a *correlation* operation between this signature of the change and the residuals. This result is extended to the case of state-space models in subsection 7.2.4.

7.2.3 ARMA Models

Here we investigate changes in an r-dimensional process $(Y_k)_k$ which can be described by a stable ARMA model as

$$Y_k \;=\; \textstyle\sum_{i=1}^{p} A_i Y_{k-i} \;+\; \sum_{j=0}^{q} B_j V_{k-j} \tag{7.2.88}$$

where $(V_k)_k$ is a white noise sequence with covariance matrix R and where $B_0 = I$. We follow [Nikiforov, 1980, Nikiforov, 1983].

To introduce the additive changes, let us consider first the AR case where we assume the following model after change :

$$Y_k \;=\; \textstyle\sum_{i=1}^{p} A_i Y_{k-i} \;+\; V_k \;+\; \Upsilon(k,t_0) \tag{7.2.89}$$

where $\Upsilon(k,t_0)$ is the profile as defined according to the different cases that we investigated in subsection 7.2.1. This can be rewritten as

$$A(z)\, Y_k \;=\; (I - \textstyle\sum_{i=1}^{p} A_i z^{-i})\, Y_k \;=\; V_k \;+\; \Upsilon(k,t_0) \tag{7.2.90}$$

Of course, as in subsection 3.2.4, we assume that the roots of the matrix polynomial on the left side of this equation are outside the unit circle. In the ARMA case, we replace the polynomial transfer function on the left side of (7.2.90) by a rational one, which results in

$$\varepsilon_k = \frac{A(z)}{B(z)}\, Y_k \;=\; \frac{I - \sum_{i=1}^{p} A_i z^{-i}}{\sum_{j=0}^{q} B_j z^{-j}}\; Y_k = V_k + \Upsilon(k,t_0) \tag{7.2.91}$$

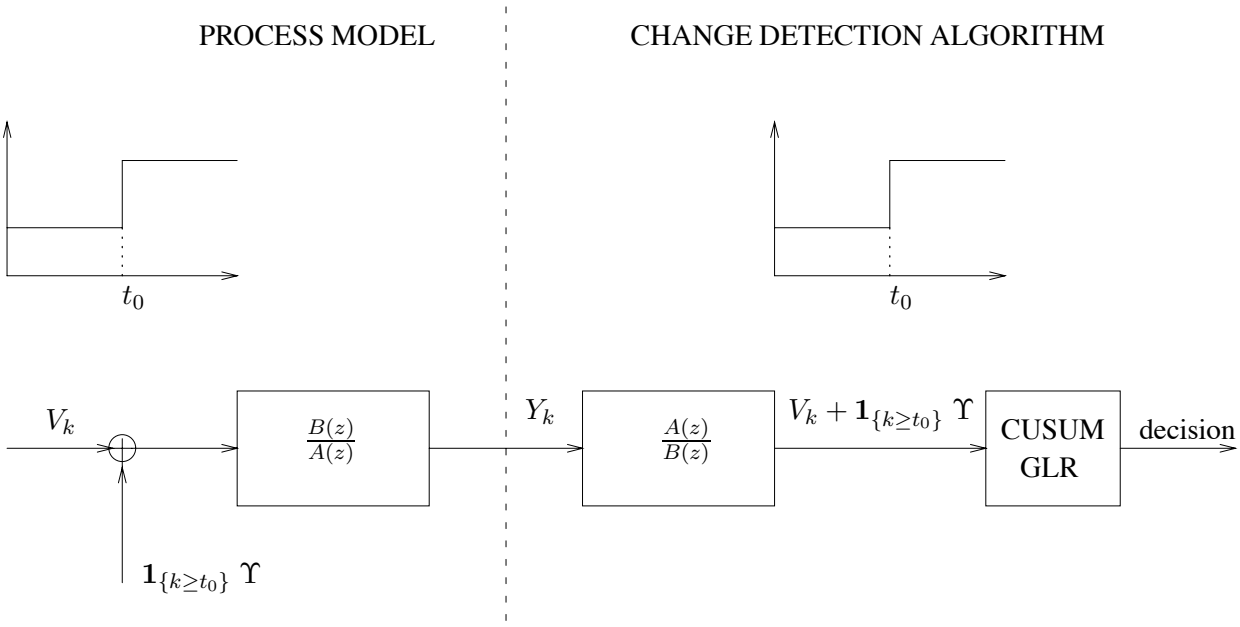

Figure 7.11 Change detection in ARMA models through innovations.

or equivalently

$$Y_k = \frac{B(z)}{A(z)}\ [V_k + \Upsilon(k, t_0)] \tag{7.2.92}$$

which gives

$$Y_k \quad = \quad \sum_{i=1}^{p} A_i Y_{k-i} \quad + \quad \sum_{j=0}^{q} B_j\, [V_{k-j} + \Upsilon(k-j, t_0)] \tag{7.2.93}$$

It is important to note that, in the present case of ARMA models and because of (7.1.13), we can assume that $\theta_0 = 0$ and that $\theta_1 = \Upsilon(k, t_0)$ for $k \geq t_0$. Recall again that, as we explained in section 7.1, in the case of dynamic profiles of changes, we consider only parametrically unknown dynamic profiles.

Now, as stated in section 7.1, we solve this change detection problem using first the log-likelihood function and the transformation from observations to innovations, and then the solution of the relevant corresponding basic problem. Note that the resulting additive change on the innovation is exactly the same as on the input excitation V, namely Υ. This is obvious from (7.2.91) (see also (7.1.13)) and is summarized in figure 7.11. Furthermore, this means basically that we model here additive changes as additive changes on the innovation of the ARMA model. For the state-space models in subsection 7.2.4, we consider additive changes in the state.

For solving the above change detection problems in ARMA models, all the algorithms introduced in subsection 7.2.1 can be used, replacing the shifted observations $Y_k - \theta_0$ by the innovations ε_k defined in (7.1.13). The main reason that we describe change detection algorithms in more detail for regression models than for ARMA models is that the key difficulty in the regression case is the degeneracy of the Gaussian distribution of the residuals, as noted in (7.2.72), which does not occur in the ARMA case.

7.2.4 State-Space Models

In this section, we investigate additive changes in the state or observation equation of a linear time invariant system represented by a state-space model. We investigate the detectability issue from this statistical point

of view in section 7.2.6. The state space-model that we consider here is

$$\left\{ \begin{array}{rcl} X_{k+1} & = & FX_k + GU_k + W_k \\ Y_k & = & HX_k + JU_k + V_k \end{array} \right. \tag{7.2.94}$$

where the state X, the control U, and the observation Y have dimensions n, m, r, respectively, and where $(W_k)_k$ and $(V_k)_k$ are two independent white noise sequences, with covariance matrices Q and R, respectively. As in section 3.2, by using the forward shift operator z, this can be rewritten as

$$Y_k = \left[\begin{array}{ccc} H(zI_n - F)^{-1}G + J & \vdots & H(zI_n - F)^{-1} \quad I_r \end{array} \right] \left(\begin{array}{c} U_k \\ \cdots \\ W_k \\ V_k \end{array} \right) \tag{7.2.95}$$

Thus, let

$$\begin{array}{rcl} \mathcal{T}(z) & = & [\quad \mathcal{T}_U(z) \qquad\qquad\qquad \vdots \quad \mathcal{T}_W(z) \qquad\qquad I_r \quad] \\ & = & [\quad H(zI_n - F)^{-1}G + J \quad \vdots \quad H(zI_n - F)^{-1} \quad I_r \quad] \end{array} \tag{7.2.96}$$

be the transfer function of this system. By transfer function, we mean here a possibly *unknown input transfer function*. This is not the case in many fault detection and diagnosis techniques, where a *known* input is used for inferring about the possible faults.

We first describe models for additive changes, both in state-space and in transfer function representations, and the corresponding types of faults in dynamical systems. Then we discuss the dynamic profile of the resulting change on the innovation. Next, we describe the statistical algorithms that are convenient for the cases 1, 7, and 8 of the basic problem. And finally, we discuss a modified version of the GLR algorithm, which was proven to be of interest in a particular case of the state-space model [Basseville and Benveniste, 1983a], together with the practically important issue of the estimation of noise variances for improving the performances of a change detection algorithm.

7.2.4.1 Additive Changes in State-Space Models

As mentioned before, we consider the following model of changes :

$$\left\{ \begin{array}{rcl} X_{k+1} & = & FX_k + GU_k + W_k + \Gamma\, \Upsilon_x(k, t_0) \\ Y_k & = & HX_k + JU_k + V_k + \Xi\, \Upsilon_y(k, t_0) \end{array} \right. \tag{7.2.97}$$

where Γ and Ξ are matrices of dimensions $n \times \tilde{n}$ and $r \times \tilde{r}$, respectively, and $\Upsilon_x(k, t_0)$ and $\Upsilon_y(k, t_0)$ are the *dynamic* profiles of the assumed changes, of dimensions $\tilde{n} \leq n$ and $\tilde{r} \leq r$, respectively. Neither the gain matrices nor the profiles are necessarily completely known *a priori*. These additive changes can be represented with the aid of figure 6.2 and figure 7.12. As in the case of regression models in subsection 7.2.2, we investigate several types of constraints on these quantities, corresponding to different levels of *a priori* knowledge on the change. This point is discussed later. The instant t_0 is again the unknown change time, so that $\Upsilon_x(k, t_0) = \Upsilon_y(k, t_0) = 0$ for $k < t_0$.

Let us now comment on the relations between the dynamic profiles of changes and the parameter θ of the distribution of the observations Y. As we discussed in section 7.1, for state-space models the question of knowing whether the change vectors Υ_x and Υ_y are constant or not is of no interest, because in both cases the resulting change on the innovation has a dynamic profile. This becomes clear when we compute this signature. Therefore, let us assume that Υ_x *and* Υ_y *are parametric functions of time*, and that the vector of parameters after change is θ.

Example 7.2.1 *To explain this assumption, we consider the following simple case of a slope :*

$$\Upsilon_{x,\theta}(k,t_0) = [\theta^p + \theta^s(k-t_0)]\ \mathbf{1}_{\{k\geq t_0\}} \tag{7.2.98}$$

where

$$\theta = \begin{pmatrix} \theta^p \\ \theta^s \end{pmatrix} \tag{7.2.99}$$

It is useful for drawing some conclusions [Willsky, 1986, Tanaka, 1989] to consider the particular case of a scalar change magnitude :

$$\left\{ \begin{array}{rcccccccc} X_{k+1} & = & FX_k & + & GU_k & + & W_k & + & \nu\,\Upsilon_x(k,t_0) \\ Y_k & = & HX_k & + & JU_k & + & V_k & + & \nu\,\Upsilon_y(k,t_0) \end{array} \right. \tag{7.2.100}$$

where ν is a *scalar* unknown magnitude of changes lying respectively in the directions Υ_x and Υ_y, which are dynamic profiles of failures, of dimensions n and r, respectively.

In the corresponding transfer function, these changes are also additive, as the following equation shows. The model (7.2.97) can be rewritten as

$$\begin{array}{rcl} Y_k & = & \left[\ \mathcal{T}_U(z)\ \vdots\ \mathcal{T}_W(z)\quad I_r\quad \vdots\ \mathcal{T}_\Upsilon(z)\ \right] \begin{pmatrix} U_k \\ \cdots \\ W_k \\ V_k \\ \cdots \\ \Upsilon_x(k,t_0) \\ \Upsilon_y(k,t_0) \end{pmatrix} \\ \text{where}\ \ \mathcal{T}_\Upsilon(z) & = & \left[\ H(zI_n - F)^{-1}\Gamma\ \vdots\ \Xi\ \right] \end{array} \tag{7.2.101}$$

and where $\mathcal{T}_U$ and $\mathcal{T}_W$ are defined in (7.2.96). Let us discuss these failure models, give several examples, and compare them with models used in most investigations about the geometrical methods, which we introduce in section 7.4. First, we observe that (7.2.97) contains (7.2.100) as a particular case. Second, we note that useful examples of changes or failures that can be modeled with the aid of (7.2.100) are a bias on a particular sensor or actuator, as we discuss now. If $\Upsilon_x = 0$ and Υ_y is a vector, the components of which are all zero except for the jth component, which equals one for $k \geq t_0$, then (7.2.100) corresponds to the onset of a bias in the jth component of Y, namely in the jth sensor. Similarly, if $\Upsilon_y = 0$ and Υ_x is a vector, the components of which are all zero except for the jth component, which equals one for $k \geq t_0$, then (7.2.100) corresponds to the onset of a bias in the jth component of U, namely in the jth actuator.

Finally, we see that

$$\begin{array}{rcl} \tilde{n} & = & n \\ \tilde{r} & = & r \\ \Gamma & = & I_n \\ \Xi & = & I_r \\ \Upsilon_x(k,t_0) & = & \nu_x(k,t_0)\Upsilon_x \\ \Upsilon_y(k,t_0) & = & \nu_y(k,t_0)\Upsilon_y \end{array} \tag{7.2.102}$$

lead to the model usually used in the detection filter approach [White and Speyer, 1987], where Υ_x and Υ_y are design failure directions associated with plant or actuator or sensor failures. For example, when Υ_x is chosen to be a specific column G_j of the input matrix G, then

- $\nu_x(k,t_0) = \nu\ \mathbf{1}_{\{k\geq t_0\}}$ corresponds to a constant bias in the jth actuator;

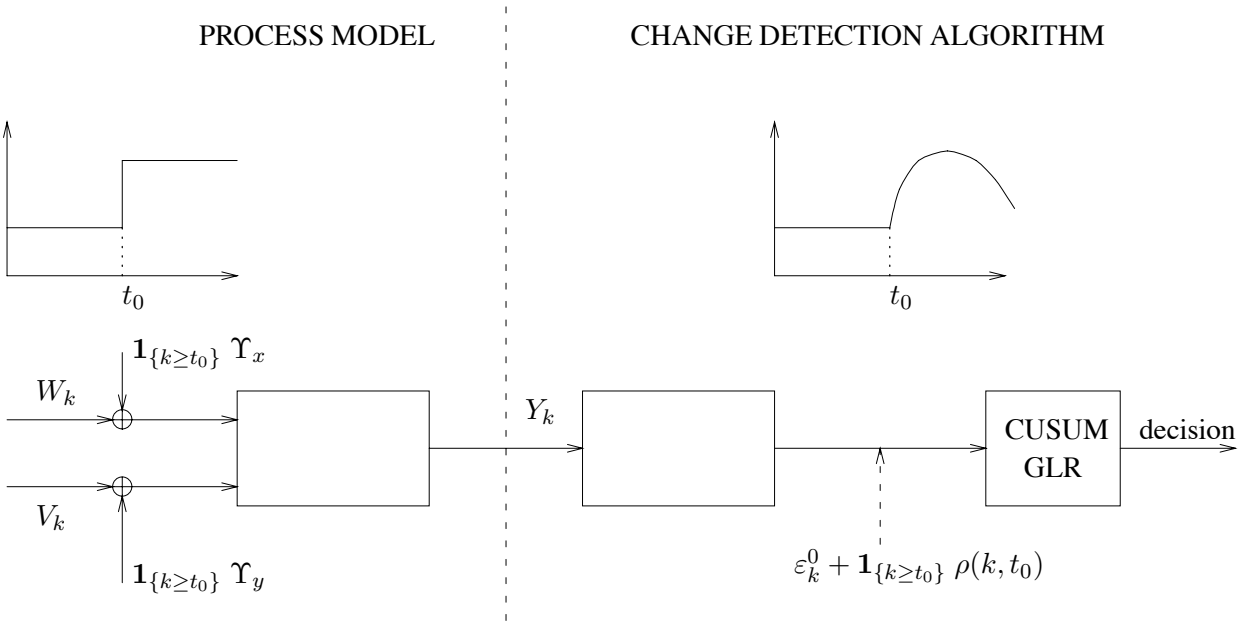

Figure 7.12 Change detection in state-space models through innovations.

- $\nu_x(k,t_0) = -U_k(j)\ \mathbf{1}_{\{k\geq t_0\}}$ corresponds to a complete failure of the jth actuator;
- $\nu_x(k,t_0) = (\nu - U_k(j))\ \mathbf{1}_{\{k\geq t_0\}}$ corresponds to the case where the jth actuator is stacked to a constant value.

Similarly, Υ_y being equal to a column H_i of the output matrix H corresponds to failures of the ith sensor.

It should be noted that, in some cases, it can be useful to model sensor failures as actuators failures by adding to the dynamics of the system, the dynamics of the system generating the additive change Υ_y. This is discussed, for example, in [White and Speyer, 1987, Massoumnia *et al.*, 1989, Wahnon *et al.*, 1991a]. We use this when introducing some geometrical techniques.

It is worth emphasizing a key issue about failure models. In some circumstances, algorithms designed with the aid of the model (7.2.97) for *additive* changes may also detect changes in the *dynamics* of the system (7.2.94); see, for example, discussions in [Willsky, 1986, Massoumnia, 1986, White and Speyer, 1987]. This means only that a given change detection algorithm, assuming a particular type of change, can be tried in practice for the detection of *any* other type change. Sometimes, this approach works, but it should be clear that it is not the best way of using the available *a priori* information about the possible changes!

As we explained in section 7.1, we solve the additive change detection problem by first using the *transformation from observations to innovations*, which can be achieved with the aid of a Kalman filter, and then solving the relevant corresponding basic problem. Therefore, we first investigate the profile of the change in the innovation process which results from the model of change (7.2.97). Let us recall that, even though there exists an equivalence between state-space and ARMA models, as we explained in section 3.2, we encounter here a new key issue concerning the effect on the innovation of an additive change on the model. The main reason for this comes from the dynamics of the system (7.2.94) and of the Kalman filter, as depicted in figure 7.12 and as opposed to figure 7.11. Let us now explain this carefully.

7.2.4.2 Signature of the Change on the Innovation

We first show on a simple example that a step change on the input sequence results, on the innovation, in a change with a *dynamic profile*. Then we derive the recursive formulas for computing this *signature* and give the analytic expressions - both in the time domain and using transfer functions - of this signature in the general case (7.2.97).

In the case of a known ARMA model, it is possible to transform the observations into a sequence of innovations using the whitening filter having a transfer function that is exactly the inverse of the transfer function associated with the model. Thus, the effect on the innovation of a change on the white noise input sequence is exactly the same change. If the change on the input sequence is a step, the change on the innovation is also a step, with exactly the same magnitude, but shifted in time. In the case of a state-space model, the Kalman filter produces an innovation sequence but cannot cancel the dynamics of the system as far as the step jumps on the two white noise sequences $(W_k)_k$ and $(V_k)_k$ are concerned. Let us show this with a simple example, and then discuss this point more formally.

Example 7.2.2 *Consider the following first-order state-space model :*

$$\left\{ \begin{array}{rcccccc} x_{k+1} & = & \alpha x_k & + & w_k & + & \nu_x \, \mathbf{1}_{\{k \geq t_0\}} \\ y_k & = & x_k & + & v_k & + & \nu_y \, \mathbf{1}_{\{k \geq t_0\}} \end{array} \right. \tag{7.2.103}$$

Let us rewrite this model in an ARMA form

$$y_k = \frac{w_k + \nu_x \, \mathbf{1}_{\{k \geq t_0\}}}{1 - \alpha z^{-1}} + v_k + \nu_y \, \mathbf{1}_{\{k \geq t_0\}} \tag{7.2.104}$$

or equivalently

$$\begin{aligned} (1 - \alpha z^{-1}) y_k &= w_k + v_k - \alpha v_{k-1} + (\nu_x + \nu_y) \left(1 - \frac{\nu_y}{\nu_x + \nu_y} \alpha z^{-1} \right) \mathbf{1}_{\{k \geq t_0\}} \\ &= K(1 - \beta z^{-1}) \varepsilon_k + (\nu_x + \nu_y) \left(1 - \frac{\nu_y}{\nu_x + \nu_y} \alpha z^{-1} \right) \mathbf{1}_{\{k \geq t_0\}} \end{aligned} \tag{7.2.105}$$

where ε_k is a white noise sequence with variance 1. Now, let us consider the no-change ARMA model corresponding to this last equation, and investigate the additive changes as modeled in subsection 7.2.3 :

$$\begin{aligned} (1 - \alpha z^{-1}) y_k &= K(1 - \beta z^{-1})(\varepsilon_k + \nu \, \mathbf{1}_{\{k \geq t_0\}}) \\ &= K(1 - \beta z^{-1}) \varepsilon_k + K(1 - \beta z^{-1}) \nu \, \mathbf{1}_{\{k \geq t_0\}} \end{aligned} \tag{7.2.106}$$

The comparison between (7.2.105) and (7.2.106) shows that if

$$\beta = \frac{\nu_y}{\nu_x + \nu_y} \alpha \tag{7.2.107}$$

then these two models are equivalent for additive changes, up to a scale factor on the change magnitude. This condition holds when

$$\frac{\nu_y^2}{(\nu_x + \nu_y)^2} = \frac{\sigma_v^2}{\sigma_w^2 + \sigma_v^2} \tag{7.2.108}$$

Therefore, in all the other cases, the dynamics of the excitation and of the change in (7.2.105) are different, and thus the stepwise additive changes in (7.2.103) result in an additive change with a different dynamic profile on the innovation.

Now, let us derive the **recursive equations and explicit formulas** for the dynamic profile of the signature of the change on the innovation in the general case. It should be clear that here innovation is the output of the Kalman filter corresponding to the model (7.2.94) without change. Referring to the terminology introduced in subsection 3.1.2, this output process is exactly the innovation process before the change and is the residual process after the change. For simplicity, we keep the name innovation, which should not introduce any confusion.

As in section 3.2, the Kalman filter for estimating the state X of the model (7.2.94) can be summarized in the following manner. The one-step ahead prediction, the innovation ε_k, and the estimated state are given in (3.2.19) and the Kalman gain K_k, the state estimation error covariance $P_{k|k}$, and the covariance of the innovation Σ_k are given in (3.2.20). The linear feature of the model (7.2.94) and the additive effect of the change $(\Gamma\Upsilon_x, \Xi\Upsilon_y)$ in (7.2.97) lead to the following decomposition of the state, its estimate, and the innovation :

$$\begin{array}{rcl} X_k &=& X_k^0 + \alpha(k,t_0) \\ \hat{X}_{k|k} &=& \hat{X}_{k|k}^0 + \beta(k,t_0) \\ \varepsilon_k &=& \varepsilon_k^0 + \rho(k,t_0) \end{array} \tag{7.2.109}$$

where the exponent 0 is for the quantities corresponding to the unchanged model (7.2.94) and where the last term of each equation represents the effect of a change $(\Gamma\Upsilon_x, \Xi\Upsilon_y)$ occurring at time $t_0 \leq k$. The functions α, β, and ρ are (pre-)computed with the aid of the following recursions :

$$\begin{array}{rcl} \alpha(k,t_0) &=& F\alpha(k-1,t_0) + \Gamma\Upsilon_x(k-1,t_0) \\ \beta(k,t_0) &=& (I-K_kH)F\beta(k-1,t_0) + K_k\left[H\alpha(k,t_0) + \Xi\Upsilon_y(k,t_0)\right] \\ &=& F\beta(k-1,t_0) + K_k\rho(k,t_0) \\ \rho(k,t_0) &=& H\left[\alpha(k,t_0) - F\beta(k-1,t_0)\right] + \Xi\Upsilon_y(k,t_0) \end{array} \tag{7.2.110}$$

with the initial conditions

$$\begin{array}{rcl} \alpha(t_0,t_0) &=& 0 \\ \beta(t_0-1,t_0) &=& 0 \end{array}$$

Note that the signature $\rho(k,t_0)$ of the change on the innovation depends upon both k and t_0 during the transient behavior of the Kalman filter. When the steady-state behavior is reached, this signature depends only upon $k-t_0$.

The closed-form expressions of α, β, ρ assuming the steady-state behavior of the Kalman filter, and using both time domain and transfer function representations, are given in the appendix to this section. We obtain

$$\begin{array}{rcll} \rho(k,t_0) &=& +\sum_{i=0}^{k-t_0-1} & H\bar{F}^i\Gamma\Upsilon_x(k-i-1,t_0) \\ && -\sum_{i=0}^{k-t_0-1} & H\bar{F}^iFK\Xi\Upsilon_y(k-i-1,t_0) \\ && + & \Xi\Upsilon_y(k,t_0) \end{array} \tag{7.2.111}$$

Using transfer function notation, this can be rewritten as

$$\rho(k,t_0) = \mathcal{K}_x(z)\Upsilon_x(k,t_0) + \mathcal{K}_y(z)\Upsilon_y(k,t_0) \tag{7.2.112}$$

$$\text{where } \mathcal{K}_x(z) = \sum_{i=0}^{k-t_0-1} H\bar{F}^i\Gamma z^{-i-1}$$

$$\mathcal{K}_y(z) = -\sum_{i=0}^{k-t_0-1} H\bar{F}^iFK\Xi z^{-i-1} + \Xi$$

Straightforward computations lead to

$$\begin{aligned} \mathcal{K}_x(z) &= H(zI_n - \bar{F})^{-1}(I_n - \bar{F}^{k-t_0} z^{-k+t_0})\Gamma \\ \mathcal{K}_y(z) &= -H(zI_n - \bar{F})^{-1}(I_n - \bar{F}^{k-t_0} z^{-k+t_0})FK\Xi + \Xi \end{aligned} \tag{7.2.113}$$

which for k large asymptotically simplify into

$$\begin{aligned} \mathcal{K}_x(z) &= H(zI_n - \bar{F})^{-1}\Gamma \\ \mathcal{K}_y(z) &= \left[I_r - H(zI_n - \bar{F})^{-1}FK\right]\Xi \end{aligned} \tag{7.2.114}$$

In summary, the innovation ε_k (output of the Kalman filter) has the following distribution :

$$\begin{aligned} \mathcal{L}(\varepsilon_k) &= \mathcal{N}(0, \Sigma_k) \qquad \text{when no change occurs} \\ \mathcal{L}(\varepsilon_k) &= \mathcal{N}[\rho(k, t_0), \Sigma_k] \quad \text{after change} \end{aligned} \tag{7.2.115}$$

We now investigate several cases of the basic problem.

7.2.4.3 Case 1 : Known Parameter After Change

The case of known parameters before and after change is solved with the aid of the CUSUM algorithm. Therefore, the relevant algorithm here is given by formula (7.2.3), where $Y_i - \theta_0$ should be replaced by the innovation $\varepsilon_i, \theta_1 - \theta_0$ by ρ, and where Σ is the time-varying estimated covariance matrix of ε.

This gives

$$t_a = \min\{k \geq 1 : g_k \geq h\} \tag{7.2.116}$$

$$g_k = \max_{1 \leq j \leq k} S_j^k \tag{7.2.117}$$

$$\begin{aligned} S_j^k &= \ln \frac{\prod_{i=j}^k p_{\rho(i,j)}(\varepsilon_i)}{\prod_{i=j}^k p_0(\varepsilon_i)} \\ &= \sum_{i=j}^k \rho^T(i,j)\Sigma_i^{-1}\varepsilon_i - \frac{1}{2}\sum_{i=j}^k \rho^T(i,j)\Sigma_i^{-1}\rho(i,j) \end{aligned} \tag{7.2.118}$$

Let us comment about the characteristic features of this particular CUSUM algorithm. In the present case of a dynamic profile, after the change, the parameter of the distribution of the innovations does vary with time, and thus the increments of the decision function are not identically distributed. The resulting CUSUM algorithm can lead to difficulties when the time-varying parameter $\rho(i,j)$ becomes equal to zero (or more generally to the value of the parameter before change), and, moreover, nothing is known about its properties.

Formula (7.2.118) can be rewritten in the following more recursive form :

$$g_k = (\mathcal{S}_k)^+ \tag{7.2.119}$$

$$N_k = N_{k-1}\, \mathbf{1}_{\{g_{k-1}>0\}} + 1 \tag{7.2.120}$$

$$\begin{aligned} \mathcal{S}_k &= \mathcal{S}_{k-1}\, \mathbf{1}_{\{g_{k-1}>0\}} \\ &\quad + \rho^T(k, k-N_k+1)\Sigma_k^{-1}\varepsilon_k - \frac{1}{2}\rho^T(k, k-N_k+1)\Sigma_k^{-1}\rho(k, k-N_k+1) \end{aligned} \tag{7.2.121}$$

Note, however, that ρ cannot be computed in a completely recursive way until the steady-state behavior of the Kalman filter is reached, as stated before.

In (7.2.118), the quantity

$$\mathbf{J}_{k,j} = \sum_{i=j}^{k} \rho^T(i,j)\Sigma_i^{-1}\rho(i,j) \tag{7.2.122}$$

is the *Kullback divergence* between the two joint distributions of the innovation sequence $(\varepsilon_i)_{i=j,\ldots,k}$ given in (7.2.115), and is used when discussing the detectability issue after. Furthermore, as we noted in subsection 7.2.1, the basic computation on which the detection is based is the *correlation between the innovations ε of the Kalman filter and the signatures of the changes in (7.2.97) on these innovations*. Finally, the estimation of the change time is achieved through

$$\hat{t}_0 = \arg \max_{1\leq j\leq t_a} S_j^{t_a} \tag{7.2.123}$$

This algorithm is valid not only in the case of known constant parameter θ_1 after change, but also in the more general case of known magnitude and dynamic profiles of change $\Upsilon_x(k,t_0)$ and $\Upsilon_y(k,t_0)$.

7.2.4.4 Case 7 : Known θ_0 and Dynamic Profile of the Change, but Unknown Magnitude

As for the basic problem, when the change magnitude ν is unknown, as in (7.2.100), it is estimated by maximizing the log-likelihood ratio. Therefore, the relevant algorithm here is given by formulas (7.2.46) and (7.2.47), where $Y_i - \theta_0$ should be replaced by the innovation ε_i, Υ by ρ, and Σ is the time-varying estimated covariance matrix of ε. This gives

$$\begin{aligned} g_k &= \max_{1\leq j\leq k} \sup_{\nu} S_j^k \\ \sup_{\nu} S_j^k &= \hat{\nu}_k(j) \sum_{i=j}^{k} \tilde{\rho}^T(i,j)\Sigma_i^{-1}\varepsilon_i - \frac{\hat{\nu}_k^2(j)}{2} \sum_{i=j}^{k} \tilde{\rho}^T(i,j)\Sigma_i^{-1}\tilde{\rho}(i,j) \qquad (7.2.124) \\ &= \frac{1}{2} \frac{\left(\sum_{i=j}^{k} \tilde{\rho}^T(i,j)\Sigma_i^{-1}\varepsilon_i\right)^2}{\sum_{i=j}^{k} \tilde{\rho}^T(i,j)\Sigma_i^{-1}\tilde{\rho}(i,j)} \qquad (7.2.125) \end{aligned}$$

where

$$\hat{\nu}_k(j) = \frac{\sum_{i=j}^{k} \tilde{\rho}^T(i,j)\Sigma_i^{-1}\varepsilon_i}{\sum_{i=j}^{k} \tilde{\rho}^T(i,j)\Sigma_i^{-1}\tilde{\rho}(i,j)} \tag{7.2.126}$$

is the estimate of the change magnitude at time k, assuming a change at time j. Note that, with respect to case 1, in these formulas we assume that ρ is of the form $\nu\tilde{\rho}$ because of the particular model (7.2.100) that we consider here.

As we explained in subsection 7.2.1, the change time t_0 is estimated with the aid of maximum likelihood estimation, which leads to an exhaustive search of this maximum for all possible past (i.e., before k) time instants. In order not to increase linearly the size of this search, t_0 is estimated by looking for the maximum value of S inside a finite window of fixed size M :

$$\hat{t}_{0_k} = \arg \max_{k-M+1\leq j\leq k} S_j^k \tag{7.2.127}$$

The underlying intuitive idea is that we assume that older changes have already been detected. It is worth emphasizing that, even though the search for the change time is constrained in time, the resulting algorithm

is *not a finite horizon technique*, basically because the likelihood ratio itself is recursively computed with the aid of *all* past information.

The change magnitude estimate is finally

$$\hat{\nu}_k = \hat{\nu}_k(\hat{t}_{0_k}) \tag{7.2.128}$$

for $k = t_a$. The distributions of $\hat{\nu}_k$ and S_j^k are

$$\begin{array}{rcl} \mathcal{L}(\hat{\nu}_k) & = & \mathcal{N}(\nu, \mathbf{J}_{k,t_0}^{-1}) \\ \mathcal{L}(S_j^k) & = & \chi'^2(1, \mathbf{J}_{k,j}) \end{array} \tag{7.2.129}$$

In other words, the log-likelihood ratio S is a χ^2 variable with noncentrality parameter $\mathbf{J}$. We use this fact when considering the detectability issue.

7.2.4.5 Case 8 : Unknown Parameter After Change

When both the change magnitude and direction are unknown, as in (7.2.97), the relevant algorithm is again the GLR algorithm, as described in (7.2.50), where $\bar{Y}$ should be replaced by $\bar{\varepsilon}$ and Σ is again the covariance matrix of the innovation.

The algorithm is thus

$$g_k = \max_{1 \le j \le k} \sup_{\Upsilon} S_j^k \tag{7.2.130}$$

$$\sup_{\Upsilon} S_j^k = \left[\sum_{i=j}^{k} \tilde{\rho}^T(i,j) \Sigma_i^{-1} \varepsilon_i \right]^T \left[\sum_{i=j}^{k} \tilde{\rho}^T(i,j) \Sigma_i^{-1} \tilde{\rho}(i,j) \right]^{-1} \left[\sum_{i=j}^{k} \tilde{\rho}^T(i,j) \Sigma_i^{-1} \varepsilon_i \right]$$

The estimate of the change direction at time k, assuming a change at time j, is

$$\hat{\Upsilon}_k(j) = \left[\sum_{i=j}^{k} \tilde{\rho}^T(i,j) \Sigma_i^{-1} \tilde{\rho}(i,j) \right]^{-1} \left[\sum_{i=j}^{k} \tilde{\rho}^T(i,j) \Sigma_i^{-1} \varepsilon_i \right] \tag{7.2.131}$$

Note that, with respect to case 1, in these formulas we assume that ρ is of the form $\tilde{\rho}\Upsilon$. Moreover, the quantity

$$\mathbf{J}_{k,j} = \Upsilon^T \left[\sum_{i=j}^{k} \tilde{\rho}^T(i,j) \Sigma_i^{-1} \tilde{\rho}(i,j) \right] \Upsilon \tag{7.2.132}$$

$$= \sum_{i=j}^{k} \rho^T(i,j) \Sigma_i^{-1} \rho(i,j) \tag{7.2.133}$$

is the *Kullback divergence* between the two joint distributions of the innovation sequence $(\varepsilon_i)_{i=j,\ldots,k}$ given in (7.2.115). This divergence is again the noncentrality parameter of the distribution of the cumulative sum S_j^k after change.

In the present case, the estimate of the change time is (7.2.127) as in the previous case. Then, the final change magnitude estimate is

$$\hat{\Upsilon}_k = \hat{\Upsilon}_k(\hat{t}_{0_k}) \tag{7.2.134}$$

Actually, this and the previous cases were investigated and their GLR solution was first derived in a completely recursive form in [Willsky and Jones, 1976]. Recall that this algorithm is made of several steps :

- detection of the change;
- estimation of the change time and magnitude;
- updating of the initial state and error covariance estimates for the Kalman filter, using the change magnitude estimate.

The first two steps are basic in GLR methodology; the third is aimed at improving the tracking ability of the Kalman filter in the presence of abrupt changes in the state X (see the discussion in section 2.5).

The drawback of this algorithm is that the choice of threshold h in (7.2.116) may be critical, in the sense that the number of resulting alarms may be sensitive to this choice, as is shown for a particular application in [Basseville and Benveniste, 1983a]. Moreover this threshold also depends upon the window size M used in (7.2.127) for maximization over t_0. These are the main motivations for the derivation of the modified version of the GLR algorithm introduced in [Basseville and Benveniste, 1983a], which we explain in the next example.

Finally, after the detection of a change, the reason for updating the initial estimates is to give to the Kalman filter more appropriate initial values after the detection of a change than the initial values given at the beginning of the processing, now using all the past information about the processed signal as summarized by the estimated change times and magnitudes. One possible solution, which is given in [Willsky and Jones, 1976], consists of

$$\hat{X}_{k|k,\text{update}} = \hat{X}^0_{k|k} + \left[\alpha(k,\hat{t}_{0_k}) - \beta(k,\hat{t}_{0_k})\right] \hat{\Upsilon}_k \tag{7.2.135}$$

for the estimation of the state variables, and of

$$P_{k|k,\text{update}} = P^0_{k|k} + \left[\alpha(k,\hat{t}_{0_k}) - \beta(k,\hat{t}_{0_k})\right] \left[\sum_{i=j}^{k} \tilde{\rho}^T(i,j)\Sigma_i^{-1}\tilde{\rho}(i,j)\right]^{-1} \left[\alpha(k,\hat{t}_{0_k}) - \beta(k,\hat{t}_{0_k})\right]^T$$

for the covariance matrix of the state estimation error. Recall that α and β are the signatures of the change on the state and state estimate, respectively; they are computed in the appendix to this section. With respect to the discussion in chapter 8 of different possible ways of generating changes, this updating scheme assumes the first method. Other updating schemes are investigated in [Caglayan and Lancraft, 1983].

This overall algorithm - namely filter, detection, estimation, updating - has been successfully used in a variety of applications, such as sensor failure detection in aircraft [Deckert *et al*., 1977], rhythm analysis in ECG signals [Gustafson *et al*., 1978], monitoring of road traffic density [Willsky *et al*., 1980], tracking of maneuvering targets [Korn *et al*., 1982], and geophysical signal processing [Basseville and Benveniste, 1983a].

Example 7.2.3 (Modified GLR algorithm). *The above-mentioned drawback of the GLR algorithm, namely the coupling effect between threshold h and window size M, frequently arises in practical applications. For this reason, a modified decision function is proposed in [Basseville and Benveniste, 1983a] for a particular state-space model. The decision function is no longer the likelihood ratio as before, but a smoothed version of the change magnitude estimate, which was observed to be quite accurate on real data, even when the change actually occurs in much more than one time step. The resulting algorithm turns out to act as a* low-pass filter everywhere except at the change times.

The chosen model is a state-space model of dimension 2, namely a model of a constant slope perturbed by noise, on which changes on the mean level can occur. The state-space model is

$$\begin{cases} x_{k+1} &= x_k + \mu_k + w^1_k + \nu\, \mathbf{1}_{\{k\geq t_0\}} \\ \mu_{k+1} &= \mu_k + w^2_k \\ y_k &= x_k + v_k \end{cases} \tag{7.2.136}$$

where $(w_k^1)_k$, $(w_k^2)_k$, and $(v_k)_k$ are zero mean independent white Gaussian noises, with respective variances q^1, q^2, and σ^2, which are all unknown parameters. The noises w_k^1 and w_k^2 allow the Kalman filter to track the slow fluctuations of the signal with respect to the constant slope model. It is known [Bohlin, 1977] that q^1 and q^2 are much more difficult to identify than σ^2. On the other hand, $q^i\sigma^2 \quad (i = 1, 2)$ is known to be a kind of forgetting factor for the Kalman filter which computes the innovations. Therefore, q^1 and q^2 are chosen a priori and σ^2 is estimated on-line, with the aid of the following estimate :

$$\hat{\sigma}^2_{k+1} = \frac{k-2}{k-1}\,\hat{\sigma}^2_k \;\mathbf{1}_{\{k>2\}} + \frac{k}{(k+1)(k+2)}\,(y_{k+1} - \hat{x}_k)^2 \tag{7.2.137}$$

It is worth emphasizing that all variance estimates are not equivalent in this framework. The key advantage of using $(y_{k+1} - \hat{x}_k)^2$ instead of $(y_k - \hat{x}_k)^2$ or even $(y_{k+1} - \hat{x}_k - \hat{\mu}_k)^2$ is to incorporate the local slope effect, and thus to increase σ^2 - and consequently decrease the Kalman gain - in the "noisy" slope segments. Actually, any underestimation of noise variances is undesirable. The reader is referred to [Mehra, 1970] for further discussion of this issue of variance estimation.

The GLR algorithm is then computed for this particular model (7.2.136) with these choices of a priori fixed or estimated variances. The modified GLR algorithm is based upon another decision function, which is a smoothed version of the change magnitude estimate. Then we have to choose a minimum magnitude of change ν_m to be detected - exactly as in section 2.2 - and a threshold h. This allows us to obtain a significantly better decoupling between these parameters and the size M of the time window inside which the change time is estimated, and a lower sensitivity of the detector with respect to the choice of the threshold.

Let $\bar{\nu}_k$ be the smoothed version of the change magnitude estimate, inside a time window of length p $(p < M)$:

$$\bar{\nu}_k = \frac{1}{I_k} \sum_{j=k-I_k+1}^{k} \hat{\nu}_j \tag{7.2.138}$$

where

$$I_k = p\;\mathbf{1}_{\{k>p+1\}} + (k-1)\;\mathbf{1}_{\{k\leq p+1\}} \tag{7.2.139}$$

The empirical variance of $\hat{\nu}_k$ is computed with the aid of

$$\varsigma_k = \frac{1}{I_k - 1} \sum_{j=k-I_k+1}^{k} (\hat{\nu}_j - \bar{\nu}_k)^2 \tag{7.2.140}$$

In practice, ς_k has to be bounded from below. The decision rule is based upon the following test between $\{\bar{\nu}_k < \nu_m\}$ and $\{\bar{\nu}_k \geq \nu_m\}$:

$$\frac{I_k(\bar{\nu}_k - \nu_m)^2}{\varsigma_k} \underset{\mathbf{H}_0}{\overset{\mathbf{H}_1}{\gtrless}} h \tag{7.2.141}$$

where ν_m and h are positive quantities to be chosen.

When $\mathbf{H}_1$ is decided in $k = t_a$, the algorithm gives, as the original algorithm, the estimates

$$\hat{t}_0 = \hat{t}_{0_k}, \qquad \hat{\nu} = \hat{\nu}_k(\hat{t}_{0_k}) \tag{7.2.142}$$

and the Kalman filter is updated as before.

The decision rules (7.2.116)-(7.2.124) and (7.2.141) are both quadratic in $\hat{\nu}_k$. For the particular model and for the considered application to geophysical signals, however, their performance is significantly different, essentially in terms of robustness with respect to the choice of the parameters.

7.2.5 Statistical Decoupling for Diagnosis

We now address the problem of diagnosis or isolation, namely the problem of deciding - once a change has been detected - which one out of a set of possible changes actually occurred, using a statistical approach. We confine our discussion to the case of additive changes in state-space models.

Let us first emphasize that we do *not* address the difficult problem of diagnosis in the framework of sequential statistical inference and hypotheses testing described in chapter 4 and underlying the present section. The key reason for this is that the sequential multiple decision theory is not complete. Appropriate choice of criterion and design of decision functions are unsolved questions. For example, allowing an additional delay after detection for improving the diagnosis leads to a criterion for which optimal decision rules are difficult to derive analytically. Rather, we take here an off-line point of view for designing decision functions, which result from a statistical decoupling criterion and which can be implemented on-line for solving diagnosis problems. This subsection is thus an exception with respect to the main lines of this book - namely on-line algorithms - but is included here because of its ability to establish bridges between the statistical and geometrical points of view for additive change detection and diagnosis in state-space models, as we discuss in section 7.5.

In this subsection, we first show that the *off-line* statistical decoupling of additive changes in a *dynamic* state-space system reduces to a *static* statistical decoupling problem in a regression model, which is nothing but a hypotheses testing problem with nuisance parameters. We prove this reduction to a static framework for the detection problem, which is sufficient for the diagnosis problem as well. Then we describe two different but equivalent solutions to the statistical decoupling problem. We show that the first step of the solution to this static *statistical* decoupling problem can be implemented as a transformation of the observations that uses only the *deterministic* part of the system, and not the statistics of the noises. The link between this transformation and a standard geometrical decoupling technique is investigated in subsection 7.5.3. But, it is important to note that here we design a decision rule, based upon these residuals, which *does* include the statistics of the noises.

7.2.5.1 Off-line Detection Reduces to a Static Detection Problem

Let us thus consider the following *detection* problem in a *dynamic* system. We assume that the unfailed model is

$$\left\{ \begin{array}{rcl} X_{k+1} & = & FX_k + GU_k + W_k \\ Y_k & = & HX_k + JU_k + V_k \end{array} \right. \tag{7.2.143}$$

where the dimensions are as in (7.1.6), and that the failed model is

$$\left\{ \begin{array}{rcl} X_{k+1} & = & FX_k + GU_k + W_k + \Gamma\, \Upsilon(k) \\ Y_k & = & HX_k + JU_k + V_k \end{array} \right. \tag{7.2.144}$$

Let us first show that, when using an observation sample of size N, the off-line dynamic detection problem reduces to a static detection problem. Using the same computations as in subsection 3.2.2, we rewrite the set of N successive equations in the following manner :

$$\mathcal{Y}_1^N = \mathcal{O}_N\, X_1 + \mathcal{J}_N(G, J)\, \mathcal{U}_1^N + \mathcal{J}_N(I_n, 0)\, \mathcal{W}_1^N + \mathcal{V}_1^N + \mathcal{J}_N(\Gamma, 0)\, \Psi_1^N \tag{7.2.145}$$

where $\mathcal{O}_N$ is the observability matrix, and $\mathcal{J}_N(G,J)$ is the block Toeplitz matrix associated with the impulse response of the system :

$$\mathcal{J}_N(G,J) = \begin{pmatrix} J & \dots & \dots & \dots & \dots & \dots \\ HG & J & \dots & 0 & \dots & \dots \\ HFG & HG & J & \dots & \dots & \dots \\ HF^2G & HFG & HG & J & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ HF^{N-2}G & \dots & \dots & HFG & HG & J \end{pmatrix} \tag{7.2.146}$$

and thus is a lower triangular matrix. Finally,

$$\Psi_1^N = \begin{pmatrix} \Upsilon(1) \\ \Upsilon(2) \\ \vdots \\ \Upsilon(N) \end{pmatrix} \tag{7.2.147}$$

which reduces to $\Psi_1^N = \mathbb{1}_N \otimes \Upsilon$ when Υ is constant.

We rewrite (7.2.145) as

$$\tilde{\mathcal{Y}}_1^N = \mathcal{O}_N\, X_1 + \tilde{\mathcal{V}}_1^N + \mathcal{J}_N(\Gamma,0)\, \Psi_1^N \tag{7.2.148}$$

where

$$\tilde{\mathcal{Y}}_1^N = \mathcal{Y}_1^N - \mathcal{J}_N(G,J)\, \mathcal{U}_1^N \tag{7.2.149}$$

and where the noise

$$\tilde{\mathcal{V}}_1^N = \mathcal{J}_N(I_n,0)\, \mathcal{W}_1^N + \mathcal{V}_1^N \tag{7.2.150}$$

has the covariance matrix

$$\begin{aligned} \mathcal{R}_N &= \mathcal{J}_N(I_n,0)\, (I_N \otimes Q)\, \mathcal{J}_N^T(I_n,0) + I_N \otimes R \\ &= \check{\mathcal{J}}_N(QH^T,0) + \check{\mathcal{J}}_N^T(QH^T,0) + I_N \otimes R \end{aligned} \tag{7.2.151}$$

In this formula, we use the notation

$$\check{\mathcal{J}}_N = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \mathcal{J}_{N-1} & \\ 0 & & & \end{pmatrix} \tag{7.2.152}$$

Note that, when the transfer function of the system is invertible, $\mathcal{R}_N$ is positive definite.

The key issue in (7.2.148) is that X_1 is *independent* of $\tilde{\mathcal{V}}_1^N$, basically because of the definition of $\mathcal{W}_1^N$ and $\mathcal{V}_1^N$, and because the first block-line of $\mathcal{J}_N(I_n,0)$ in (7.2.146) is zero. Without loss of generality, we shall assume X_1 to be known. When X_1 is unknown, we replace it by its estimate. The least-squares estimation of X_1 is discussed in subsection 7.4.2. Therefore, from now on we consider $\bar{Y} = \tilde{\mathcal{Y}}_1^N - \mathcal{O}_N\, X_1$. The off-line detection problem then reduces to the following *static statistical detection problem* :

$$\begin{aligned} \bar{Y} &\sim \mathcal{N}(\bar{\mu}, \Sigma) \\ \Sigma &= \mathcal{R}_N = \mathrm{cov}(\tilde{\mathcal{V}}_1^N) \\ \mathbf{H}_0 = \{\bar{\mu} : \bar{\mu} = 0\} &\quad \text{and} \quad \mathbf{H}_1 = \{\bar{\mu} : \bar{\mu} = M\mu \neq 0\} \end{aligned} \tag{7.2.153}$$

where

$$M\mu = \mathcal{J}_N(\Gamma, 0)\ \Psi_1^N \tag{7.2.154}$$

Note that we can write

$$\mathcal{J}_N(\Gamma, 0)\ \Psi_1^N = \tilde{\mathcal{J}}_N(\Gamma)\ \tilde{\Psi}_1^N \tag{7.2.155}$$

where

$$\tilde{\mathcal{J}}_N(G) = \begin{pmatrix} HG & \dots & \dots & \dots & \dots & \dots \\ 0 & HG & 0 & \dots & 0 & \dots \\ 0 & HFG & HG & 0 & \dots & \dots \\ 0 & HF^2G & HFG & HG & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & HF^{N-2}G & \dots & \dots & HFG & HG \end{pmatrix} \tag{7.2.156}$$

is a full column rank matrix, and where

$$\tilde{\Psi}_1^N = \begin{pmatrix} 0 \\ \Upsilon(1) \\ \vdots \\ \Upsilon(N-1) \end{pmatrix} \tag{7.2.157}$$

In (7.2.153), we can thus assume that M *is a full column rank matrix*. The dimensions of $\bar{Y}$ and μ are $\bar{r} = Nr$ and $\bar{n} = N\tilde{n}$, respectively, $\tilde{n}$ being the dimension of the change vector Υ. We assume here that $\tilde{n} \leq r$, namely that *the dimension of the change vector is less than or equal to the dimension of the measurements*. We shall further discuss this dimension issue due to the off-line point of view later.

In subsection 4.2.6, we describe the GLR solution to the hypotheses testing problem (7.2.153) concerning the mean of an independent Gaussian sequence. The present case of a regression model with Gaussian excitation was discussed in subsection 7.2.2, and the GLR algorithm consists of computing the generalized log-likelihood ratio S_N :

$$\begin{aligned} S_N &= \ln \frac{\sup_\mu p_\mu(\bar{Y})}{p_0(\bar{Y})} \\ &= \ln p(\bar{Y}|\hat{\mu}) - \ln p(\bar{Y}|0) \end{aligned} \tag{7.2.158}$$

It results from the computation (7.2.66) of the log-likelihood function in a regression model that

$$2S_N = \bar{Y}^T \Sigma^{-1} M (M^T \Sigma^{-1} M)^{-1} M^T \Sigma^{-1} \bar{Y} \tag{7.2.159}$$

which has a χ^2 distribution, with a number of degrees of freedom equal to $\bar{n}$ (the size of μ) and a noncentrality parameter under $\mathbf{H}_1$ given by

$$\lambda = \mu^T M^T \Sigma^{-1} M \mu \tag{7.2.160}$$

7.2.5.2 Statistical Isolation in a Static System

Let us now go back to the *diagnosis* problem in the dynamical system. We want to build a statistical decision function based upon an observation sample of fixed size N, which is able to detect an additive change $\Gamma_1\ \Upsilon_1(k)$ as in (7.2.144), while being insensitive to another additive change $\Gamma_2\ \Upsilon_2(k)$. We will call this problem off-line statistical diagnosis, or equivalently off-line *dynamic statistical decoupling (or isolation) problem*. Note that the isolation of one failure among a set of κ possible failures can be deduced from the solution of this problem in a straightforward manner.

Now we use two different problem statements and investigate three possible solutions to this isolation problem. We also show that these three solutions are equivalent. First we state the isolation or diagnosis problem as an hypotheses testing problem between

$$\mathbf{H}_0 \; : \; \left\{ \begin{array}{rcccccccc} X_{k+1} & = & FX_k & + & GU_k & + & W_k & + & \Gamma_2 \, \tilde{\Upsilon}_2(k) \\ Y_k & = & HX_k & + & JU_k & + & V_k & & \end{array} \right. \tag{7.2.161}$$

and

$$\mathbf{H}_1 \; : \; \left\{ \begin{array}{rcccccccccc} X_{k+1} & = & FX_k & + & GU_k & + & W_k & + & \Gamma_1 \, \Upsilon_1(k) & + & \Gamma_2 \, \Upsilon_2(k) \\ Y_k & = & HX_k & + & JU_k & + & V_k & & & & \end{array} \right. \tag{7.2.162}$$

Note that we assume the presence of the *nuisance* change $\Gamma_2 \, \Upsilon_2(k)$ under *both* hypotheses $\mathbf{H}_0$ and $\mathbf{H}_1$. Otherwise, the probability of false alarms would be biased because it would not take into account the presence of the nuisance change. Note also that the direction of the nuisance change is not necessarily the same for both hypotheses.

The equivalence between off-line detection in a dynamical system and a static detection problem implies that this diagnosis problem is equivalent to the following *static* detection problem :

$$\begin{array}{rcl} \bar{Y} & \sim & \mathcal{N}(\bar{\mu}, \Sigma) \\ \Sigma & = & \mathcal{R}_N \\ \mathbf{H}_0 = \{\bar{\mu} \; : \bar{\mu} = M_2 \tilde{\mu}_2\} \quad \text{and} & & \mathbf{H}_1 = \{\bar{\mu} \; : \bar{\mu} = M\mu = M_1\mu_1 + M_2\mu_2\} \end{array} \tag{7.2.163}$$

where, for $i = 1, 2$,

$$M_i = \mathcal{J}_N(\Gamma_i, 0) \tag{7.2.164}$$

and

$$\mu_i = (\Psi_i)_1^N = \left(\begin{array}{c} \Upsilon_i(1) \\ \Upsilon_i(2) \\ \vdots \\ \Upsilon_i(N) \end{array} \right) \quad \text{and :} \quad \tilde{\mu}_2 = (\tilde{\Psi}_2)_1^N = \left(\begin{array}{c} \tilde{\Upsilon}_2(1) \\ \tilde{\Upsilon}_2(2) \\ \vdots \\ \tilde{\Upsilon}_2(N) \end{array} \right) \tag{7.2.165}$$

In (7.2.163), the informative parameter is μ_1 and the nuisance parameter is μ_2. The dimensions are as before : the dimension of $\bar{Y}$ is $\bar{r} = Nr$, and the dimension of μ_i is $\bar{n}_i = N\tilde{n}_i$, where $\tilde{n}_i$ is the dimension of the change vector $\Upsilon_i, \; (i = 1, 2)$. As before, the matrices M_i $(i = 1, 2)$ can be assumed to be full rank. We assume furthermore that

$$M = [\; M_1 \quad M_2 \;] \tag{7.2.166}$$

is a *full column rank matrix*.

In subsection 4.2.8, we discuss the hypotheses testing problem concerning the mean of an independent Gaussian sequence in the presence of a nuisance parameter. We show that there exist two possible approaches to this problem, namely the minmax approach and the GLR approach for both the nuisance and informative parameters, and that these two solutions are equivalent. For the *regression* model, we first use the GLR approach. Then we use the results given in subsection 4.2.8 concerning the derivation of the minmax algorithm and the equivalence between the GLR and minmax approaches. Therefore, we give here only the algorithm resulting from this minmax approach, and not its derivation. Finally, we show how it is possible to use a typical system theory approach to the decoupling problem *in connection with* a statistical decision function, and show the equivalence of the resulting algorithm with the GLR and thus with the minmax algorithms.

GLR solution The GLR solution to the hypotheses testing problem (7.2.163) consists of computing the log-likelihood ratio :

$$S_N = \ln p(\bar{Y}|\hat{\mu}_2, \hat{\mu}_1) - \ln p(\bar{Y}|\hat{\tilde{\mu}}_2, 0) \tag{7.2.167}$$

It results from (7.2.159) that

$$\begin{aligned} 2S_N &= -\bar{Y}^T \left[\Sigma^{-1} - \Sigma^{-1}M(M^T\Sigma^{-1}M)^{-1}M^T\Sigma^{-1}\right]\bar{Y} \\ &\quad +\bar{Y}^T \left[\Sigma^{-1} - \Sigma^{-1}M_2(M_2^T\Sigma^{-1}M_2)^{-1}M_2^T\Sigma^{-1}\right]\bar{Y} \\ &= \bar{Y}^T\Sigma^{-1}\left[M(M^T\Sigma^{-1}M)^{-1}M^T - M_2(M_2^T\Sigma^{-1}M_2)^{-1}M_2^T\right]\Sigma^{-1}\bar{Y} \end{aligned} \tag{7.2.168}$$

Using the expression

$$M^T\Sigma^{-1}M = \begin{pmatrix} M_1^T\Sigma^{-1}M_1 & M_1^T\Sigma^{-1}M_2 \\ M_2^T\Sigma^{-1}M_1 & M_2^T\Sigma^{-1}M_2 \end{pmatrix} \tag{7.2.169}$$

and the formula for the inverse of a partitioned matrix, straightforward but long computations lead to

$$2S_N = \bar{Y}^T\left[\bar{P}_2M_1(M_1^T\bar{P}_2M_1)^{-1}M_1^T\bar{P}_2\right]\bar{Y} \tag{7.2.170}$$

where

$$\bar{P}_2 = \Sigma^{-1}\left[\Sigma - M_2(M_2^T\Sigma^{-1}M_2)^{-1}M_2^T\right]\Sigma^{-1} \tag{7.2.171}$$

The matrix $\bar{P}_2$ is such that

$$\begin{aligned} \bar{P}_2M_2 &= 0 \\ \text{rank } (\bar{P}_2) &= Nr - \bar{n}_2 = N(r - \tilde{n}_2) \end{aligned} \tag{7.2.172}$$

In order that (7.2.170) can be computed, namely $M_1^T\bar{P}_2M_1$ is invertible, we need that

$$\text{rank } (M_1^T\bar{P}_2M_1) = N\tilde{n}_1 \tag{7.2.173}$$

A *necessary* condition for the last equality to hold true is that $\tilde{n}_1 + \tilde{n}_2 \leq r$. This condition is *not sufficient*, and thus the rank of $M_1^T\bar{P}_2M_1$ must be checked in each situation. Actually, we recover here the intuitively obvious fact that if there exists a masking effect of the change of interest by the nuisance change, then we cannot isolate these two changes using this statistical approach. Furthermore, the noncentrality parameter of the χ^2-test (7.2.170) is

$$\lambda = \mu_1^TM_1^T\bar{P}_2M_1\mu_1 \tag{7.2.174}$$

which is independent of μ_2 and $\tilde{\mu}_2$, as desired.

Note that the positivity of this noncentrality parameter provides us with a condition of detectability of the change μ_1 in the presence of the nuisance change μ_2.

Minmax approach The minmax approach to hypotheses testing problems in the presence of nuisance parameters described in subsection 4.2.8 leads to the following choice of transformation of the observations :

$$\mathcal{A} = \left[\begin{array}{c|c} I_{\bar{n}_1} & -M_1^T\Sigma^{-1}M_2(M_2^T\Sigma^{-1}M_2)^{-1} \end{array}\right] M^T\Sigma^{-1} \tag{7.2.175}$$

Note that

$$\mathcal{A} = M_1^T\bar{P}_2 \tag{7.2.176}$$

where $\bar{P}_2$ is defined in (7.2.171). This matrix $\mathcal{A}$ satisfies :

$$\mathcal{A}\, M_2 = 0 \tag{7.2.177}$$

For this particular choice, the transformed observations $\bar{Y}^* = \mathcal{A}\bar{Y}$ satisfy

$$\bar{Y}^* \sim \mathcal{N}(\bar{P}\mu_1, \bar{P}) \tag{7.2.178}$$

where

$$\bar{P} = M_1^T \bar{P}_2 M_1 \tag{7.2.179}$$

In particular, its mean value depends only upon μ_1. Thus, the isolation decision function reduces to

$$2S_N^* = (\bar{Y}^*)^T \bar{P}^{-1} \bar{Y}^* \tag{7.2.180}$$

which is a χ^2-test, with a number of degrees of freedom equal to $\bar{n}_1$ (the size of μ_1) and a noncentrality parameter under $\mathbf{H}_1$ given by

$$\lambda^* = \mu_1^T \bar{P} \mu_1 \tag{7.2.181}$$

which is, of course, independent of μ_2.

Now, let us discuss the rank issues. In order that the decision rule (7.2.180) can be computed, we need that

$$\text{rank}\bar{P} = N\tilde{n}_1 \tag{7.2.182}$$

namely this matrix should be full rank, exactly as in the GLR approach. Note that this condition is equivalent to the full rank condition concerning the transformation matrix $\mathcal{A}$. Again, this rank issue will condition the feasibility of the isolation of the two changes.

Note that the minmax χ^2-test (7.2.180) can be written as

$$\begin{aligned} 2S_N^* &= \bar{Y}^T \mathcal{A}^T \bar{P}^{-1} \mathcal{A} \bar{Y} \\ &= \bar{Y}^T \left[\bar{P}_2 M_1 (M_1^T \bar{P}_2 M_1)^{-1} M_1^T \bar{P}_2 \right] \bar{Y} \end{aligned} \tag{7.2.183}$$

and the noncentrality parameter as

$$\begin{aligned} \lambda^* &= \mu_1^T \bar{P} \mu_1 \\ &= \mu_1^T M_1^T \bar{P}_2 M_1 \mu_1 \end{aligned} \tag{7.2.184}$$

which are exactly as in (7.2.170) and (7.2.174), respectively. In other words, we find that the GLR solution is exactly the same as the minmax solution, as in subsection 4.2.8.

Using a geometrical decoupling in a statistical framework We now use another straightforward method to cancel the effect of the nuisance change Υ_2 in (7.2.163). Let $\mathcal{A}$ be any maximal full row rank matrix such that

$$\mathcal{A}\, M_2 = 0 \tag{7.2.185}$$

Two such matrices are related through premultiplication with an invertible matrix. In (7.2.175), we exhibit one possible choice for $\mathcal{A}$, and we discuss other possible choices after.

It is of key importance to ensure that such a matrix $\mathcal{A}$ does not kill part of the information related to the change of interest Υ_1, namely that the condition

$$\text{ker}\, \mathcal{A}M_1 = \text{ker}\, M_1 \tag{7.2.186}$$

holds. It can be shown that this condition is exactly the same as the rank conditions in the two previously described statistical approaches. For this purpose, we assume that $M = [\ M_1 \quad M_2\ ]$ is a *full column rank*

matrix. In the case of unknown dynamic profiles of changes in (7.2.162), a condition for this to be true is that $\tilde{n}_1 + \tilde{n}_2 \leq r$ and rank$[\ H\ \Gamma_1 \quad H\ \Gamma_2\ ] = r$ [Wahnon *et al.*, 1991a].

Consider the transformed observation given by

$$\bar{Y}^* = \mathcal{A}\,\bar{Y} \tag{7.2.187}$$

This transformed observation satisfies

$$\begin{array}{rcl} \bar{Y}^* & \sim & \mathcal{N}(\bar{\mu}^*, \Sigma^*) \\ \Sigma^* & = & \mathcal{A}\Sigma\mathcal{A}^T \\ \mathbf{H}_0 = \{\bar{\mu}^*\ : \bar{\mu}^* = 0\} & \text{and} & \mathbf{H}_1 = \{\bar{\mu}^*\ : \bar{\mu}^* = \mathcal{A}M_1\mu_1\} \end{array} \tag{7.2.188}$$

Because of (7.2.159) and (7.2.160), the resulting χ^2 test is

$$2S_N^* = (\bar{Y}^*)^T(\mathcal{A}\Sigma\mathcal{A}^T)^{-1}\mathcal{A}M_1\left[M_1^T\mathcal{A}^T(\mathcal{A}\Sigma\mathcal{A}^T)^{-1}\mathcal{A}M_1\right]^{-1}M_1^T\mathcal{A}^T(\mathcal{A}\Sigma\mathcal{A}^T)^{-1}\bar{Y}^* \tag{7.2.189}$$

which has a number of degrees of freedom equal to $\bar{n}_1$ (the size of μ_1), and a noncentrality parameter equal to

$$\lambda^* = \mu_1^T M_1^T \mathcal{A}^T(\mathcal{A}\Sigma\mathcal{A}^T)^{-1}\mathcal{A}M_1\mu_1 \tag{7.2.190}$$

which is independent of μ_2 and $\tilde{\mu}_2$. Thus, this is another relevant isolation decision function.

Again, the issue of the rank of the matrices that have to be inverted in (7.2.189) is crucial. We must check that the matrix $\mathcal{A}\Sigma\mathcal{A}^T$ is full rank, and then, using the previous geometrical decoupling assumption (7.2.186), that the matrix $M_1^T\mathcal{A}^T(\mathcal{A}\Sigma\mathcal{A}^T)^{-1}\mathcal{A}M_1$ is also invertible. When $\mathcal{A}$ is chosen to be $\mathcal{A} = \bar{P}_2$, we recover the GLR test (7.2.170) because in this case $\mathcal{A}\Sigma\mathcal{A}^T = \mathcal{A}$.

Equivalence between the GLR algorithm and the mixed geometrical/statistical approach The important point is that the χ^2-test (7.2.189) is left *unchanged* when $\mathcal{A}$ is premultiplied by any invertible matrix, and thus does *not* depend upon the choice of $\mathcal{A}$ in (7.2.185). Thus, the key design issue as far as statistical diagnosis is concerned lies in justifying the choice of $\mathcal{A}$.

Any solution to equation (7.2.185) is convenient; thus, any solution that involves the single *deterministic* part of the system, and thus does *not* depend upon the covariance Σ, is of interest. In other words, $\mathcal{A}$ can be chosen to be any full row rank matrix, the rows of which span the left null space of M_2. It turns out that there exist *standard* and efficient algorithms for computing such a matrix [White and Speyer, 1987, Massoumnia *et al.*, 1989].

This result means the following. The static statistical decoupling problem can be solved with the aid of a transformation of observations which is independent of the statistics of the signal. The only things needed are to find a full row rank matrix $\mathcal{A}$ satisfying (7.2.185), and *to check the above mentioned invertibility conditions*. Note, however, that the decision rule, based upon these transformed observations, uses the statistics of the signal.

Discussion: dimensional issues Because of the dimensions of the involved matrices, and even though the M_i are lower triangular matrices, the feasibility of such an off-line approach may be questionable. We refer the reader to [Wahnon *et al.*, 1991a] for a suboptimal on-line implementation of this diagnosis algorithm. Furthermore, robustness problems can arise with this approach, basically because M_2 and thus $\mathcal{A}$ depend upon the dynamics of the system, which cannot be perfectly known. However, using a descriptor system representation of this detection and diagnosis problem, it is possible [Wahnon *et al.*, 1991b, Benveniste *et al.*, 1993] to considerably reduce the dimensions of the involved matrices and at the same time

to gain in robustness by using $\mathcal{A}$ matrices which do not depend upon the dynamics of the system. We do not discuss further this issue here.

The fact that statistical decoupling includes as its first step a transformation that involves only the single deterministic part of the system, is used for discussing the connection between the statistical and geometrical points of view for diagnosis in subsection 7.5.3.

7.2.6 Statistical Detectability

In this subsection, we discuss the detectability issue within the framework of the statistical approach, and relate it to the concept of mutual information between two distributions. In subsection 7.4.4, we investigate the geometrical point of view and the connection between the two resulting detectability definitions.

We investigate the detectability issue in the three types of models that we consider here for additive changes, namely the regression, ARMA, and state-space models, and using the detectability definition that we gave in the introduction to part II in terms of Kullback information, or equivalently in terms of Kullback divergence in the present case of additive changes in a Gaussian distribution. We first complete the definition of detectability that we gave in section 6.3, adding comments on composite hypotheses and robust detectability. Then, for regression and ARMA models, we basically discuss the detectability of an ideal step change, namely of a change with a known constant profile. For state-space models, we discuss the more complex issue of the detectability of a change having a dynamic profile, because this type of change is the basic one when considering the innovations provided by a Kalman filter, as discussed before.

7.2.6.1 Detectability Definitions

Two different types of detectability are of interest. The first concerns the ideal situation of known parameters before and after change, in which the detection algorithm is tuned with the aid of the true model parameters. The second is of interest from a practical point of view, and is related to the issue of robustness of an algorithm tuned with parameter values that are distinct from the true values. We distinguish these two cases now. Let us first recall the detectability definition that we gave in chapter 6.

Definition 7.2.1 (Statistical detectability). *Consider a change from a distribution p_{θ_0} to a distribution p_{θ_1}. Let s be the log-likelihood ratio*

$$s(y) = \ln \frac{p_{\theta_1}(y)}{p_{\theta_0}(y)} \tag{7.2.191}$$

and $\mathbf{K}(\theta_1, \theta_0)$, which is defined by

$$\mathbf{K}(\theta_1, \theta_0) = \mathbf{E}_{\theta_1}[s(Y)] \geq 0 \tag{7.2.192}$$

be the Kullback information. The change is said to be detectable if the Kullback information satisfies

$$\mathbf{K}(\theta_1, \theta_0) > 0 \tag{7.2.193}$$

Recall that, in the case of a *random process*, the information $\mathbf{K}$ is defined as a limit value when the sample size goes to infinity. Note that this definition of detectability includes the case of mutually singular distributions, for which $\mathbf{K}(\theta_1, \theta_0) = +\infty$. An example of this degenerate situation is investigated later.

Let us comment further on this definition. Because in the parametric case the Kullback information is zero only when the two parameter values are equal, one could argue that this definition is equivalent to the much simpler statement that any change between two different parameter values is detectable. But the problem of detectability is strongly related to the problem of parameterization. For example, in the multidimensional Gaussian case with constant covariance matrix, the mean vector can be a complex function

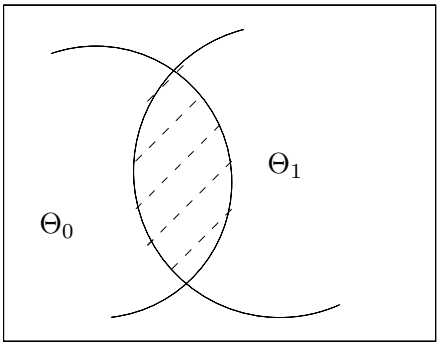

Figure 7.13 Intersecting sets of parameters and detectability.

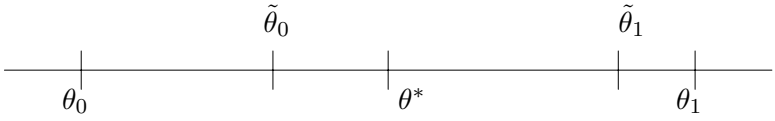

Figure 7.14 Robust detectability in the scalar case.

of the parameter θ of interest. In this case, it can be a *nontrivial* problem to detect a change, not in the mean, but in the parameter θ.

Again in the ideal case of known parameters before and after change, let us define the detectability of changes between two *composite* hypotheses.

Definition 7.2.2 *A change from a family of distributions $\mathcal{P}_0 = \{\mathbf{P}_\theta\}_{\theta\in\Theta_0}$ to a family of distributions $\mathcal{P}_1 = \{\mathbf{P}_\theta\}_{\theta\in\Theta_1}$ is said to be detectable if*

$$\inf_{\theta_0\in\Theta_0,\theta_1\in\Theta_1} \mathbf{K}(\theta_1,\theta_0) > 0 \tag{7.2.194}$$

The intuitive meaning of this definition is depicted in figure 7.13. When the parameter sets intersect, then it is impossible to discriminate the two probability measures when they both belong to the intersecting subset.

We now give a possible definition of detectability in the case of an algorithm tuned with nonexact parameter values. Referring to (6.3.5), the natural extension of the previous definition is as follows.

Definition 7.2.3 (Robust detectability). *A change from $\tilde{\theta}_0$ to $\tilde{\theta}_1$ is said to be detectable by a statistic s if*

$$\mathbf{E}_{\tilde{\theta}_1}[s(Y)] - \mathbf{E}_{\tilde{\theta}_0}[s(Y)] > 0 \tag{7.2.195}$$

When the decision rule is based upon the log-likelihood ratio $s(Y) = \ln \frac{p_{\theta_1}(Y)}{p_{\theta_0}(Y)}$ computed with the aid of the *assumed* parameter values θ_0 and θ_1, this results in

$$\int p_{\tilde{\theta}_1}(y) \ln \frac{p_{\theta_1}(y)}{p_{\theta_0}(y)}\, dy - \int p_{\tilde{\theta}_0}(y) \ln \frac{p_{\theta_1}(y)}{p_{\theta_0}(y)}\, dy > 0 \tag{7.2.196}$$

Let us discuss this condition in the case of a *scalar parameter*, as depicted in figure 7.14. It results from section 4.2 that there exists a parameter value θ^* such that

$$\mathbf{E}_{\theta^*}(s_i) = 0 \tag{7.2.197}$$

where $s_i = s(y_i)$. In the Gaussian case, we have simply

$$\theta^* = \frac{\theta_0 + \theta_1}{2} \tag{7.2.198}$$

For the robust detectability condition to be fulfilled, it is necessary that $\tilde{\theta}_0$ and $\tilde{\theta}_1$ lie on each side of θ^*. We continue our discussion of this question in chapter 10 when we investigate the tuning issues.

Note that in the case of an error in the assumed change direction, namely when the algorithm is tuned for a change from θ_1 to θ_0, the left side of (7.2.196) is negative, and thus the change is not detectable.

7.2.6.2 Regression Models

As in subsection 7.2.2, we consider here the sensor failure model :

$$Y_k = HX_k + V_k + \Upsilon \, \mathbf{1}_{\{k \geq t_0\}} \tag{7.2.199}$$

where X and Y have dimensions n and $r > n$, respectively, Υ is known, and the white noise V has covariance matrix $R = AA^T$. We show in subsection 7.2.2 that the Kullback divergence is

$$\mathbf{J}(0, \Upsilon) = \Upsilon^T A^{-T} P_H A^{-1} \Upsilon = \Upsilon^T \tilde{P}_H^T R^{-1} \tilde{P}_H \Upsilon \tag{7.2.200}$$

where P_H is the projection matrix defined in (7.2.65).

Note that this is also the Kullback divergence for the *transformed problem* on $e_k = \tilde{P}_H Y_k$. The identity of the Kullback divergence in the initial and transformed problems can be checked directly. The transformed problem is concerned with the detection of the change, from 0 to $\tilde{P}_H \Upsilon$, in the mean of a Gaussian distribution with covariance matrix $\tilde{P}_H R \tilde{P}_H^T = A P_H A^T = \tilde{P}_H R$. Let $\tilde{P}_H R = BDB^T$ be the eigen-decomposition of this covariance matrix. Then, from (4.1.91), we deduce that the divergence in the transformed problem is

$$\Upsilon^T \tilde{P}_H^T B D^{-1} B^T \tilde{P}_H \Upsilon \tag{7.2.201}$$

which is equal to the initial divergence because $BD^{-1}B^T = R^{-1}$.

Let us discuss the problem of the detectability of changes between composite hypotheses, as defined in (7.2.194), in the case of a regression model. We consider here that $\theta_0 = 0$ and

$$\theta_1 = \Upsilon \in \mathbf{\Upsilon_1} = \{\Upsilon \ : \|\Upsilon\| \geq \epsilon > 0\} \tag{7.2.202}$$

The rank of the matrix $\tilde{P}_H^T R^{-1} \tilde{P}_H$ is less than the number of components in Υ. For this reason, from (7.2.200), we deduce that it is impossible to detect all the changes Υ for which $\|\Upsilon\| \geq \epsilon > 0$. In other words, in (7.2.194),

$$\inf_{\Upsilon \in \mathbf{\Upsilon}_1} \mathbf{K}(\Upsilon, 0) = 0 \tag{7.2.203}$$

Let us thus discuss the maximal number κ of nonzero components in the change vectors Υ that can be detected. This depends upon the rank of $\tilde{P}_H$, and thus upon the rank of P_H. The rank of the latter matrix is equal to $r-n$. Because this rank is $r-n$, the maximum number of nonzero components of the change vector Υ such that the detectability condition holds is $r - n$. Furthermore, if the number of nonzero components in Υ is $\kappa \leq r - n$, then a necessary and sufficient condition for the detectability of such a fault is that all the main minors with order $l \leq \kappa$ of the matrix $A^{-T} P_H A^{-1}$ should be strictly positive.

Example 7.2.4 *Consider the case where $r = 3, n = 2$, $R = A = I_3$ and*

$$H = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \tag{7.2.204}$$

Then the projection matrix

$$P_H = \frac{1}{2}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} \tag{7.2.205}$$

is of rank $r - n = 1$, *and the Kullback divergence for any change vector* $\Upsilon^T = \begin{pmatrix} \gamma_1 & \gamma_2 & \gamma_3 \end{pmatrix}$ *is*

$$\mathbf{J} = \Upsilon^T P_H \Upsilon = \frac{1}{2}(\gamma_1 - \gamma_3)^2 \tag{7.2.206}$$

Thus, if we consider the detectability of changes between simple *hypotheses, then any change vector having different first and third components is detectable, and any change vector having equal first and third components is nondetectable*.

Let us discuss the detectability of changes between composite *hypotheses, as before the example. It is clear that the first and third main minors of* P_H *are positive (equal to 1), but the second main minor is equal to zero. Therefore, the above-mentioned condition is not fulfilled and the maximum number of nonzero components in a change* Υ *that can be detected is equal to zero. Nevertheless, all possible change vectors with nonzero first or third component are detectable. In other words, we find that the change vectors* $\Upsilon_1^T = \begin{pmatrix} \gamma_1 & 0 & 0 \end{pmatrix}$ *and* $\Upsilon_3^T = \begin{pmatrix} 0 & 0 & \gamma_3 \end{pmatrix}$ *are detectable, but not the change vector* $\Upsilon_2^T = \begin{pmatrix} 0 & \gamma_2 & 0 \end{pmatrix}$, *which is in fact a kind of diagnosability condition*.

7.2.6.3 ARMA Models

According to the discussion we had in subsection 7.2.3, the additive changes that we consider in ARMA models are basically additive changes on the innovation sequence, as shown in (7.2.91). From formula (4.1.90) giving the Kullback divergence between two Gaussian variables having different mean vectors, we deduce that a change Υ on an ARMA model as in (7.2.93) is detectable if and only if

$$\Upsilon^T R^{-1} \Upsilon > 0 \tag{7.2.207}$$

Thus, any nonzero change vector is detectable when the covariance matrix of the input excitation is positive definite.

7.2.6.4 State-Space Models

In this case, the detectability definition is less obvious, because the actual change on the innovations is no longer a step as before, but a *dynamic profile*, even if the additive changes on the state and observation equations are steps, as we discussed before. Therefore, we investigate the detectability of a change in the most general case of dynamic profiles for additive changes on the states and observations as well.

We first discuss several possible detectability definitions in this case, all based upon the notion of Kullback divergence. Then we derive and analyze a closed-form expression for one of them, using the closed-form expressions of the signature ρ of the change on the innovations that we derived in subsection 7.2.4.

Detectability of profiles We still assume the failure model (7.2.97) :

$$\left\{ \begin{array}{rcl} X_{k+1} & = & FX_k + GU_k + W_k + \Gamma\, \Upsilon_x(k, t_0) \\ Y_k & = & HX_k + JU_k + V_k + \Xi\, \Upsilon_y(k, t_0) \end{array} \right. \tag{7.2.208}$$

According to our previous discussion and the GLR framework resulting in formula (7.2.129), the definition of the detectability at time k of a failure occurring at time t_0 is in terms of the Kullback divergence in the

transformed problem :

$$\mathbf{J}_{k,t_0} = \sum_{j=t_0}^{k} \rho^T(j,t_0)\ \Sigma_j^{-1}\ \rho(j,t_0) > 0 \tag{7.2.209}$$

where the failure signature ρ is defined in (7.2.110) and computed in (7.2.111)-(7.2.112) assuming that the steady-state behavior of Kalman filter is reached.

Now, let us discuss relevant conditions to be requested from this divergence to define the detectability of a change. The first definition would consist of saying that a change from (7.2.94) to (7.2.208) occurring at time t_0 is detectable if

$$\forall k \geq t_0 + l,\ \ \mathbf{J}_{k,t_0} \geq \epsilon > 0 \tag{7.2.210}$$

where l is the observability index of the system. In the case of a single change Υ_y in the observation equation (namely, when $\Upsilon_x = 0$), we have $l = 0$. The second possible definition is suggested in [Tanaka, 1989] and considers a change to be detectable if the Kullback divergence strictly increases with time; in other words, if

$$\forall k \geq t_0 + l,\ \ \tilde{\mathbf{J}}_{k,t_0} = \mathbf{J}_{k,t_0} - \mathbf{J}_{k-1,t_0} \geq \epsilon > 0 \tag{7.2.211}$$

The third possible definition is an "average" of the previous definitions and considers the positivity of the following quantity :

$$\breve{\mathbf{J}}_{k,t_0} = \frac{1}{k-t_0+1} \sum_{i=t_0}^{k} \tilde{\mathbf{J}}_{i,t_0} = \frac{1}{k-t_0+1}\ \mathbf{J}_{k,t_0} \tag{7.2.212}$$

From now on, we use definition (7.2.211) for the following reasons. The basic intuitive motivation for selecting (7.2.211) is simply to consider that any new observation must bring new information, which is exactly the motivation underlying any statistical inference method. From this point of view, (7.2.211) implies that, among the four possible behaviors of the function of time $\mathbf{J}_{k,t_0}$, which are depicted in figure 7.15, only the first upper behavior is convenient for establishing the detectability of a given change with dynamic profile. It is intuitively obvious that the second upper behavior could be admissible as well, but this is less simple to condense in a criterion, and it is not included in (7.2.211). Finally, the changes giving rise to the two lower curves in this figure are obviously much less likely to be detected. An additional interest of this detectability definition is that it results in the same conditions about the system and the changes as other intuitive or geometrical definitions. This is explained in section 7.5.

It is important to note that, following the logic of this book, we basically discuss here the detectability of changes that have an *infinite* duration. For changes Υ with a finite duration, such as impulses, other definitions of detectability should be stated.

According to the comments made before, we use the positivity of the increment in the divergence as a detectability index :

$$\tilde{\mathbf{J}}_k = \rho^T(k,t_0)\Sigma^{-1}\rho(k,t_0) \tag{7.2.213}$$

where ρ is given in (7.2.111) or (7.2.112) and Σ is the steady-state value of Σ_k given in (3.2.19).

Detectability of step profiles Let us first emphasize that, in the general case of an unknown dynamic profile $\Upsilon_x(k,t_0)$, the dimension of the parameter vector θ_1 after change is proportional to k. For off-line detection, this is not a problem because this dimension is fixed. For on-line detection, this results in difficulties for defining the detectability. For this reason, to simplify the detectability discussion, from now on we consider changes with *step* profiles, namely

$$\begin{aligned} \Upsilon_x(k,t_0) &= \Upsilon_x\ \mathbf{1}_{\{k\geq t_0\}} \\ \Upsilon_y(k,t_0) &= \Upsilon_y\ \mathbf{1}_{\{k\geq t_0\}} \end{aligned} \tag{7.2.214}$$

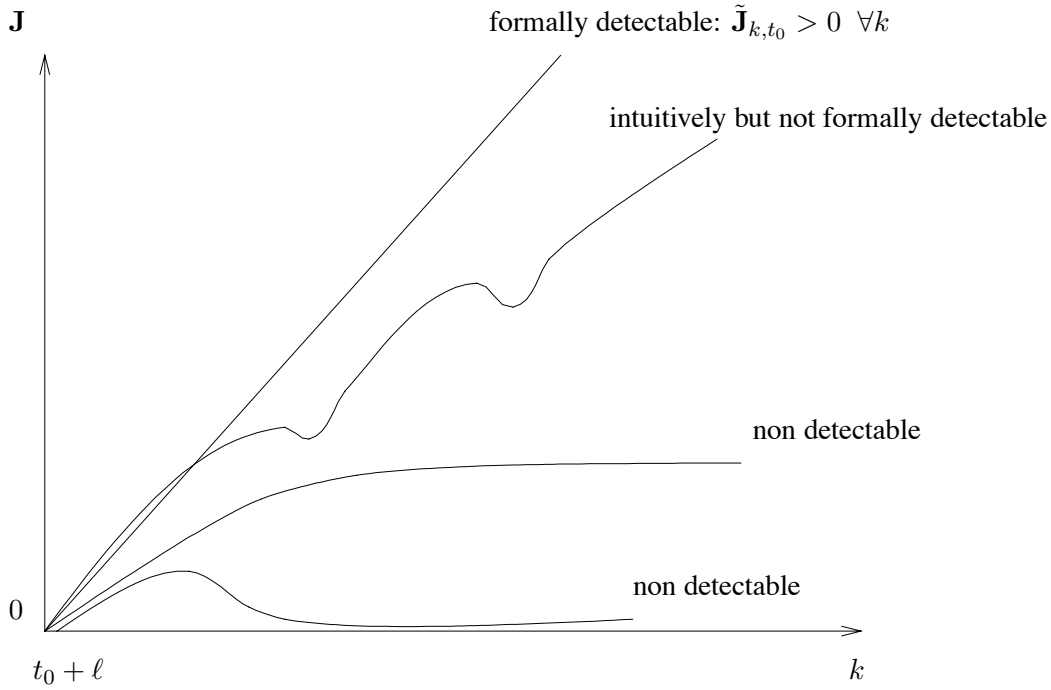

Figure 7.15 Detectability and Kullback divergence.

where Υ_x and Υ_y are constant vectors. A different investigation of the detectability of changes for other profiles can be found in [Tanaka and Müller, 1990].

In the case of step profiles, the signature ρ simplifies into

$$\rho(k, t_0) = H \left(\sum_{i=0}^{k-t_0-1} \bar{F}^i \right) \Gamma \ \Upsilon_x - H \left(\sum_{i=0}^{k-t_0-1} \bar{F}^i \right) FK \ \Xi \ \Upsilon_y + \Xi \ \Upsilon_y \tag{7.2.215}$$

Single change Υ_x In this case, we deduce that

$$\tilde{\mathbf{J}}_k = \Upsilon_x^T \Gamma^T \left(\sum_{i=0}^{k-t_0-1} \bar{F}^i \right)^T H^T \Sigma^{-1} H \left(\sum_{i=0}^{k-t_0-1} \bar{F}^i \right) \Gamma \Upsilon_x \tag{7.2.216}$$

When $k \to \infty$, this reduces to

$$\tilde{\mathbf{J}}_k = \Upsilon_x^T \Gamma^T (I_n - \bar{F})^{-T} H^T \Sigma^{-1} H (I_n - \bar{F})^{-1} \Gamma \Upsilon_x \tag{7.2.217}$$

From this, we deduce the following results. First, if the change gain Γ is such that the matrix $H(I_n - \bar{F})^{-1}\Gamma$ is full rank, then any nonzero change vector Υ_x is detectable. Second, if the change gain Γ and the change vector Υ_x are such that

$$H \left(I_n - \bar{F} \right)^{-1} \Gamma \Upsilon_x = 0 \tag{7.2.218}$$

then the divergence is saturated when k increases, and thus the corresponding change is not detectable. This generalizes the result in [Tanaka, 1989], where it is shown that, in the case of scalar magnitude, namely when $\tilde{n} = n$ and $\Gamma = \nu I_n$, if $(I_n - F)$ is nonsingular and H is full rank, then the set of nondetectable change

directions satisfying (7.2.218) is exactly the null space of $H(I_n - F)^{-1}$. For this result, it is sufficient to remark that

$$H(I_n - \bar{F})^{-1} = \left[I_r + HF(I_n - F)^{-1}K\right]^{-1} H(I_n - F)^{-1} \tag{7.2.219}$$

using the fact that $\bar{F}$ is stable.

Single change Υ_y In the case of a single sensor failure, we have that

$$\tilde{\mathbf{J}}_k = \Upsilon_y^T \Xi^T \left[I_r - H\left(\sum_{i=0}^{k-t_0-1} \bar{F}^i\right) FK\right]^T \Sigma^{-1} \left[I_r - H\left(\sum_{i=0}^{k-t_0-1} \bar{F}^i\right) FK\right] \Xi \Upsilon_y \tag{7.2.220}$$

When $k \to \infty$, this reduces to

$$\tilde{\mathbf{J}}_k = \Upsilon_y^T \Xi^T \left[I_r - H(I_n - \bar{F})^{-1}FK\right]^T \Sigma^{-1} \left[I_r - H(I_n - \bar{F})^{-1}FK\right] \Xi \Upsilon_y \tag{7.2.221}$$

From this, we deduce the following results. First, if the change gain Ξ is such that the matrix $\left[I_r - H(I_n - \bar{F})^{-1}FK\right] \Xi$ is full rank, then any nonzero change vector Υ_y is detectable. Second, if the change gain Ξ and the change vector Υ_y are such that

$$\left[I_r - H(I_n - \bar{F})^{-1}FK\right] \Xi \Upsilon_y = 0 \tag{7.2.222}$$

then the divergence is saturated when k increases, and thus the corresponding change is not detectable. This generalizes the result in [Tanaka, 1989], where it is shown that, in the case of scalar magnitude, namely when $\tilde{r} = r$ and $\Xi = \nu I_r$, if $(I_n - F)$ is nonsingular and H is full rank, then the set of nondetectable changes is empty. In other words, in this case, all sensor failures are detectable. This result is based upon the following identity :

$$\left[I_r - H(I_n - \bar{F})^{-1}FK\right] H = H(I_n - \bar{F})^{-1}(I_n - F) \tag{7.2.223}$$

and again the fact that $\bar{F}$ is stable. This result is investigated further in subsection 7.5.4 when we compare statistical and geometrical detectability definitions. More precisely, it has to be compared with (7.4.48) and (7.4.49). Finally, we should mention more recent investigations in [Tanaka and Müller, 1992].

Example 7.2.5 (Degenerate cases). *Let us conclude our discussion of statistical detectability with two degenerate examples of a change between mutually singular distributions. In this case, the log-likelihood ratio does not exist. Let us extend the definition of the Kullback information to this case, assigning to it an* infinite *value. This allows us to recover the intuitively obvious fact that a change between mutually singular distributions is the most easily detectable, because it is completely deterministic. One simple example of this situation is as follows :*

$$\begin{aligned} X_k &= \begin{pmatrix} v_k \\ v_k \end{pmatrix} \\ R_\theta &= \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \\ Y_k &= R_\theta X_k \end{aligned} \tag{7.2.224}$$

Then two different values of θ lead to mutually singular distributions.

A practically more interesting example for state-space models is the following. Let us assume that in (7.2.208) the state noise vector is of dimension $\bar{n}$ lower than the dimension of the state itself, and that the additive change occurs in the $\underline{n} = n - \bar{n}$ remaining components, namely that

$$X_{k+1} = FX_k + GU_k + \begin{pmatrix} I_{\bar{n}} & 0 \\ 0 & 0 \end{pmatrix} W_k + \begin{pmatrix} 0 \\ I_{\underline{n}} \end{pmatrix} \Upsilon_x(k, t_0) \tag{7.2.225}$$

Then the change occurs on the noise-free part of the system, and its detection can be achieved using geometrical tools as well. We discuss this example further in section 7.5.

We give another example of detectability of changes in state-space models in section 7.5 when discussing the links between the statistical and geometrical detectability definitions.

Appendix: Signature of the Change on the Innovation

Assuming the steady-state behavior of the Kalman filter, let us now derive closed-form expressions of the signatures α, β, ρ defined in (7.2.109) and using both time domain and transfer function representations.

The computation of α is straightforward :

$$\alpha(k, t_0) = \sum_{i=1}^{k-t_0} F^{i-1}\Gamma\ \Upsilon_x(k-i, t_0) \tag{7.2.226}$$

For computing β, we make use of the first recursion concerning β in (7.2.110), which we rewrite as

$$\beta(k, t_0) = \tilde{F}_k \beta(k-1, t_0) + K_k \psi_y(k, t_0) \tag{7.2.227}$$

where

$$\begin{aligned} \tilde{F}_k &= (I_n - K_k H)F \\ \psi_y(k, t_0) &= H\alpha(k, t_0) + \Xi\Upsilon_y(k, t_0) \end{aligned} \tag{7.2.228}$$

The recursion (7.2.227) has the same form as the recursion for α in (7.2.110), except that the matrix coefficients are nonconstant. The solution for β is thus a little more complex :

$$\beta(k, t_0) = \sum_{i=0}^{k-t_0} \left(\prod_{j=0}^{i-1} \tilde{F}_{k-j} \right) K_{k-i} \psi_y(k-i, t_0) \tag{7.2.229}$$

where $\prod_{j=0}^{-1} = 1$. Using the solution of the recursion for α and assuming that the steady-state behavior of the Kalman filter is reached, we get

$$\beta(k, t_0) = \sum_{i=0}^{k-t_0} \tilde{F}^i K \left[H \sum_{j=1}^{k-i-t_0} F^{j-1}\Gamma\Upsilon_x(k-i-j, t_0) + \Xi\Upsilon_y(k-i, t_0) \right] \tag{7.2.230}$$

where

$$\tilde{F} = (I_n - KH)F \tag{7.2.231}$$

The computation of ρ proceeds in two steps. First, we compute

$$\gamma(k, t_0) = \alpha(k, t_0) - F\beta(k-1, t_0) \tag{7.2.232}$$

using (7.2.226) and (7.2.230), thus assuming the steady-state behavior. Let

$$\bar{F} = F(I_n - KH) \tag{7.2.233}$$

Reasoning by induction, it is easy to prove that

$$\bar{F}^n F - F\tilde{F}^n KH = \bar{F}^{n+1} \tag{7.2.234}$$

Straightforward but long computations then give

$$\gamma(k,t_0) = \sum_{i=0}^{k-t_0-1} \left[\bar{F}^i \Gamma \Upsilon_x(k-i-1,t_0) - F\tilde{F}^i K \Xi \Upsilon_y(k-i-1,t_0)\right] \tag{7.2.235}$$

Now we have

$$\begin{aligned} \tilde{F}^n &= (I_n - KH)\bar{F}^{n-1}F \\ \bar{F}^n &= F\tilde{F}^{n-1}(I_n - KH) \\ \bar{F}^n F &= F\tilde{F}^n \end{aligned} \tag{7.2.236}$$

and thus (7.2.235) can be simplified using only $\bar{F}$. This will be used later.

We now use

$$\rho(k,t_0) = H\gamma(k,t_0) + \Xi\Upsilon_y(k,t_0) \tag{7.2.237}$$

together with (7.2.235) to obtain

$$\begin{array}{rcll} \rho(k,t_0) & = & +\sum_{i=0}^{k-t_0-1} & H\bar{F}^i\Gamma\Upsilon_x(k-i-1,t_0) \\ & & -\sum_{i=0}^{k-t_0-1} & H\bar{F}^i F K \Xi \Upsilon_y(k-i-1,t_0) \\ & & + & \Xi\Upsilon_y(k,t_0) \end{array} \tag{7.2.238}$$

Using transfer function notation, this can be re-written as

$$\begin{aligned} \rho(k,t_0) &= \mathcal{K}_x(z)\Upsilon_x(k,t_0) + \mathcal{K}_y(z)\Upsilon_y(k,t_0) \\ \text{where } \mathcal{K}_x(z) &= \sum_{i=0}^{k-t_0-1} H\bar{F}^i\Gamma z^{-i-1} \\ \mathcal{K}_y(z) &= -\sum_{i=0}^{k-t_0-1} H\bar{F}^i F K \Xi z^{-i-1} + \Xi \end{aligned} \tag{7.2.239}$$

Straightforward computations lead to

$$\begin{aligned} \mathcal{K}_x(z) &= H(zI_n - \bar{F})^{-1}(I_n - \bar{F}^{k-t_0}z^{-k+t_0})\Gamma \\ \mathcal{K}_y(z) &= -H(zI_n - \bar{F})^{-1}(I_n - \bar{F}^{k-t_0}z^{-k+t_0})FK\Xi + \Xi \end{aligned} \tag{7.2.240}$$

7.3 Properties of the Statistical Algorithms

In this section, we investigate the properties of some of the CUSUM-type and GLR algorithms that we described in subsection 7.2.1 for the basic problem. It should be clear that the properties of the algorithms for on-line additive change detection in regression, ARMA, and state-space models can be deduced from these in a straightforward manner.

Referring to the eight cases that we introduced in subsection 7.2.1, let us first outline the new components of the derivation of these properties in the present situation of additive changes in a *multidimensional* parameter, with respect to the scalar case investigated in part I. In cases 1 and 2 of simple hypotheses, all the results that are derived for the optimality and the ARL function in chapter 5 can be used with small modifications, as we show later. This results from the fact that, in these two cases, the increments of the decision function are *independent* and moreover Gaussian. For case 3 of a composite hypothesis after change, namely with known magnitude but unknown direction of change, we derive a *new* result concerning the first-order optimality of the χ^2-CUSUM algorithm. In all other cases 4 to 8, to our knowledge the properties of the corresponding algorithms are unknown. Therefore, from now on in this section, we concentrate on the linear CUSUM and χ^2-CUSUM algorithms.

7.3.1 Linear CUSUM Algorithm

As in subsection 7.2.1, we distinguish between the first two cases concerning the hypotheses about the parameters before and after change. In both these cases, the increment of the cumulative sum is an independent Gaussian sequence. Therefore, it is possible to use all the theoretical results of chapter 5 about the optimality and the computation of the ARL function for change detection algorithms. To use these results, the only thing that has to be done is to compute the mean and variance of the increment of the cumulative sum.

7.3.1.1 Known θ_0 and θ_1

We thus investigate first the simplest case of known parameters before and after change. As is obvious from (7.2.3), the decision function is a linear combination of the observations, and thus has a Gaussian distribution with mean value

$$\mu = \mathbf{E}_\theta(s_k) = (\theta_1 - \theta_0)^T \Sigma^{-1} (\theta - \theta^*) \tag{7.3.1}$$

where

$$\theta^* = \frac{\theta_0 + \theta_1}{2} \tag{7.3.2}$$

and variance

$$\sigma^2 = \mathrm{var}\,(s_k) = (\theta_1 - \theta_0)^T \Sigma^{-1} (\theta_1 - \theta_0) \tag{7.3.3}$$

The results given in example 5.2.1 and in (5.5.7)-(5.5.9) can then be used together with these values.

7.3.1.2 Θ_0 and Θ_1 Separated by a Hyperplane

In section 4.4, with the aid of figure 4.6, we explained the connection between the shape of the ARL function and the expectation of the increment of the CUSUM decision function. From this we deduce that, for investigating the properties of the linear CUSUM algorithm, it is sufficient to investigate this expectation.

As we explained for case 2 of subsection 7.2.1, the increment (7.2.4) of the relevant decision function in this case is again a linear combination of the observations, and thus has a Gaussian distribution with mean

$$\mu = \mathbf{E}_\theta(s_k) = \Upsilon^T \Sigma^{-1} (\theta - \theta^*) \tag{7.3.4}$$

and variance

$$\sigma^2 = \mathrm{var}\,(s_k) = \Upsilon^T \Sigma^{-1} \Upsilon \tag{7.3.5}$$

where Υ is the unit vector of the *assumed* change direction. Let us analyze this case. For simplicity, we define $\tilde{\Upsilon}$ by

$$\theta - \theta^* = \nu \tilde{\Upsilon} \tag{7.3.6}$$

$\tilde{\Upsilon}$ is the unit vector of the *actual* change direction. Thus

$$\mu = \nu \Upsilon^T \Sigma^{-1} \tilde{\Upsilon} \tag{7.3.7}$$

Let us now insert (5.5.8) into (5.5.7) in order to rewrite Wald's approximation of the ARL function as follows :

$$\hat{L}_0(b) = \frac{2b\hat{L}_0^{\frac{1}{2}}(0) + e^{-2b\hat{L}_0^{\frac{1}{2}}(0)} - 1}{2b^2} \tag{7.3.8}$$

where

$$b = \nu \frac{\Upsilon^T \Sigma^{-1} \tilde{\Upsilon}}{(\Upsilon^T \Sigma^{-1} \Upsilon)^{\frac{1}{2}}} \neq 0 \tag{7.3.9}$$

This formula is useful because it gives the ARL function in terms of its single value at the origin $\hat{L}_0(0)$ and the ratio b, without using the threshold. Thus, for example, it allows us to assign a value $\hat{L}_0(0)$ (considering the middle point θ^*), and then compute the corresponding delay for different values of $b > 0$. But it also allows us to compute other mean times between false alarms for other values of $b < 0$.

On the other hand, Wald's approximation (5.5.8) gives the expression of the ARL function at θ^* in terms of the threshold h :

$$\hat{L}_0(0) = \frac{h^2}{\Upsilon^T \Sigma^{-1} \Upsilon} \tag{7.3.10}$$

The behavior of the function $\hat{L}_0(b)$ implies that, for fixed $\hat{L}_0(0)$, the performance of the CUSUM algorithm is improved when b^2 increases, namely the mean delay ($b > 0$) decreases and the mean time between false alarms ($b < 0$) increases.

Now, as in section 5.2, let us write Siegmund's approximation for the ARL function of the linear CUSUM :

$$\tilde{L}_0(b) = \frac{1}{b^2}\left[e^{-2\left(\frac{bh}{\sigma}+1.166b\right)} + 2\left(\frac{bh}{\sigma} + 1.166b\right) - 1\right] \tag{7.3.11}$$

for $b \neq 0$, and

$$\tilde{L}_0(0) = \left(\frac{h}{\sigma} + 1.166\right)^2 \tag{7.3.12}$$

Let us thus investigate the ratio b (7.3.9). We distinguish several cases.

1. We consider the hypotheses $\theta_0 = \theta^* - \nu\Upsilon$ and $\theta_1 = \theta^* + \nu\Upsilon$. In other words, in this case, the assumed and actual values of the change direction are the same : $\Upsilon = \tilde{\Upsilon}$. Then

$$b^2 = \nu^2 \Upsilon^T \Sigma^{-1} \Upsilon \tag{7.3.13}$$

or equivalently

$$b^2 = 2\mathbf{K}(\theta_1, \theta_0) = \mathbf{J}(\theta_0, \theta_1) \tag{7.3.14}$$

We thus recover the fact that the properties of the CUSUM algorithm are in terms of the Kullback information (or divergence in the present Gaussian case).

2. We assume again that $\Upsilon = \tilde{\Upsilon}$ and that now ν is a known constant. In this case, we can estimate the range of b^2 in (7.3.13) as follows. Because Σ is a positive definite matrix, we can write

$$\nu^2 \lambda_r \leq b^2 = \nu^2 \Upsilon^T \Sigma^{-1} \Upsilon \leq \nu^2 \lambda_1 \tag{7.3.15}$$

where $\lambda_r \leq \ldots \leq \lambda_1$ are the eigenvalues of Σ^{-1}.

3. We assume that $\tilde{\Upsilon}$ is any unit vector. Then the properties and the ARL function of the CUSUM algorithm depend upon the ratio in (7.3.9) :

$$f(\Upsilon, \tilde{\Upsilon}) = \frac{\Upsilon^T \Sigma^{-1} \tilde{\Upsilon}}{(\Upsilon^T \Sigma^{-1} \Upsilon)^{\frac{1}{2}}} \tag{7.3.16}$$

Let us consider the two following cases :

- The actual change direction $\tilde{\Upsilon}$ is fixed : then it is easy to prove that

$$\arg\min_{\Upsilon} f(\Upsilon, \tilde{\Upsilon}) = -\tilde{\Upsilon} \tag{7.3.17}$$

$$\arg\max_{\Upsilon} f(\Upsilon, \tilde{\Upsilon}) = +\tilde{\Upsilon} \tag{7.3.18}$$

  From this, it results that the best performances are reached when the actual and assumed change direction do coincide, which is intuitively obvious.
- The assumed change direction Υ is fixed : then (7.3.16) can be obviously rewritten as

$$f(\Upsilon, \tilde{\Upsilon}) = (\Upsilon^T \Sigma^{-1} \Upsilon)^{\frac{1}{2}} \ \frac{\cos(\alpha, \tilde{\Upsilon})}{\cos(\alpha, \Upsilon)} \tag{7.3.19}$$

  where $\alpha = \Sigma^{-1} \ \Upsilon$ and where $\cos(\alpha, \beta)$ is the cosine of the angle between the two vectors α and β. Therefore, f, and thus b^2, depends only upon

$$c = \frac{\cos(\alpha, \tilde{\Upsilon})}{\cos(\alpha, \Upsilon)} \tag{7.3.20}$$

  Now consider in the parameter space the cone with axis α and generatrix Υ, as depicted in figure 7.16. It is obvious that, when the actual change direction $\tilde{\Upsilon}$ is *inside* the cone, the ratio (7.3.20) is $c > 1$ and thus the CUSUM algorithm performs even better than predicted. Similarly, when $\tilde{\Upsilon}$ is *outside* the cone, the ratio is $c < 1$ and thus the CUSUM algorithm performs worse than predicted.
  The only case in which any deviation between the assumed and actual change directions certainly results in a loss of performance is when the matrix Σ has all its eigenvalues equal and thus the cone is reduced to the straight line defined by Υ.

7.3.2 χ^2-CUSUM Algorithm

In chapter 5, we discussed the properties of weighted CUSUM algorithms starting from [Pollak and Siegmund, 1975]. Recall that these properties are derived under the assumption that the weighting function is continuously differentiable in the neighborhood of θ_1. It turns out that this condition is not fulfilled for the χ^2-CUSUM algorithm, for a scalar parameter, or for a multidimensional one. Actually, as we explained in subsection 2.4.2, the χ^2-CUSUM for a scalar parameter is based upon a *degenerate* weighting function reduced to two Dirac masses on $\theta_0 \pm \nu$, and thus the previous result cannot be used in this case. Similarly, as explained in subsection 7.2.1, the χ^2-CUSUM solution to case 3 is based upon a weighting function concentrated on the surface of a sphere (7.2.11). In this subsection we investigate the asymptotic properties of the χ^2-CUSUM algorithm for both cases of scalar and multidimensional parameters.

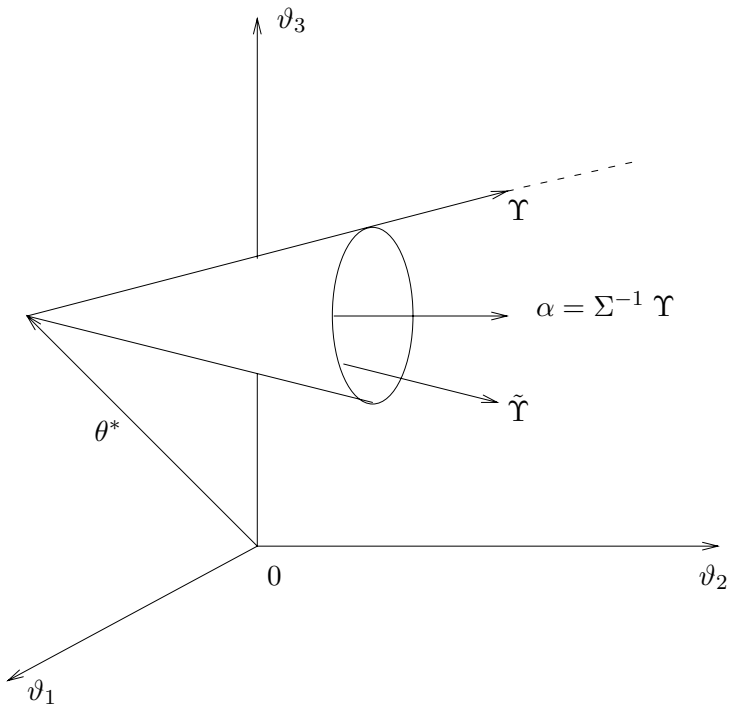

Figure 7.16 Assumed and actual change directions.

7.3.2.1 Main Idea

We first outline the main line of the proof of the first-order optimality property in both cases, and then we describe separately the proof for the two cases.

We explained in chapter 5 that, according to [Lorden, 1971], in the case of simple hypotheses before and after the change, the delay for detection $\bar{\tau}^*$ and the mean time between false alarms $\bar{T}$ for optimal change detection algorithms are related as follows :

$$\bar{\tau}^* \sim \frac{\ln \bar{T}}{\mathbf{K}(\theta_1, \theta_0)} \quad \text{when} \quad \bar{T} \to \infty \tag{7.3.21}$$

Furthermore, this result holds for a multidimensional parameter and known values of θ_0 and θ_1.

We prove that this result holds for the χ^2-CUSUM algorithm, namely in the case of known parameter θ_0 and change magnitude but *unknown* change direction :

$$\theta(k) = \begin{cases} \theta_0 & \text{when} \quad k < t_0 \\ \theta_1 \ : (\theta_1 - \theta_0)^T \Sigma^{-1} (\theta_1 - \theta_0) = b^2 & \text{when} \quad k > t_0 \end{cases} \tag{7.3.22}$$

where $b > 0$. Here we use the notation θ_1 for the parameter values after change; this makes the following computations more obvious. Note that in this situation, the Kullback information $\mathbf{K}(\theta_1, \theta_0) = \frac{b^2}{2}$ is *known*.

The two main bases for our result are the following. The mean time between false alarms is derived using Lorden's theorem and Wald's inequality, in a manner that is the same for both scalar and multidimensional parameters. On the other hand, the mean delay for detection is derived using a theorem in [Berk, 1973] concerning the ASN of the SPRT. Note that this derivation is different in the two scalar and multidimensional cases.

7.3.2.2 Mean Time Between False Alarms

The computation of the mean time between false alarms in our case proceeds in exactly the same way as the computation that was done for this quantity in subsection 5.2.3. For this purpose, it is necessary to introduce the open-ended tests corresponding to the χ^2-CUSUM algorithm :

$$T_k = \begin{cases} \min\{n \geq 1 : \tilde{\Lambda}_k^n \geq e^h\} \\ \infty \text{ if no such } n \text{ exists} \end{cases} \tag{7.3.23}$$

where

$$\ln \tilde{\Lambda}_k^n = -(n-k+1)\frac{b^2}{2} + \ln G\left[\frac{r}{2}, \frac{b^2(n-k+1)^2(\chi_k^n)^2}{4}\right] \tag{7.3.24}$$

Because $\tilde{\Lambda}_k^n$ is the likelihood ratio of χ^2 distributions (see subsection 4.3.4), it follows from Wald's inequality that

$$\mathbf{P}_{\theta_0}(T_k < \infty) \leq e^{-h} \tag{7.3.25}$$

From Lorden's theorem, we deduce that

$$\bar{T} = \mathbf{E}_{\theta_0}(t_a) \geq e^h \tag{7.3.26}$$

because the stopping time t_a is the extended stopping time associated with T_k.

7.3.2.3 Delay for Detection

It results from Berk's theorem that the delay for detection satisfies

$$\lim_{h\to\infty} \frac{\mathbf{E}_{\theta_1}(T)}{h} = \frac{1}{\varrho} \tag{7.3.27}$$

provided that

$$\frac{\ln \tilde{\Lambda}_1^n}{n} \rightarrow_{w.p.1} \varrho \in (0, +\infty) \tag{7.3.28}$$

and that the large deviation probability $p_n = \mathbf{P}_{\theta_1}\left(\frac{\ln \tilde{\Lambda}_1^n}{n} < \tilde{\varrho}\right)$, where $\tilde{\varrho} \in (0, \varrho)$ satisfies the two following conditions

$$\lim_{n\to\infty} np_n = 0 \tag{7.3.29}$$

$$\sum_{n=1}^{\infty} p_n < \infty \tag{7.3.30}$$

Now let us prove that (7.3.28) holds for the χ^2-CUSUM algorithm in both scalar and multidimensional cases. Without loss of generality, *we assume that the covariance matrix is identity*. By definition of the χ^2-CUSUM algorithm, we have

$$\frac{\ln \tilde{\Lambda}_1^n}{n} = \frac{\tilde{S}_1^n}{n} = -\frac{b^2}{2} + \frac{1}{n}\ln G\left[\frac{r}{2}, \frac{b^2 n^2 (\chi_1^n)^2}{4}\right] \tag{7.3.31}$$

where

$$(\chi_1^n)^2 = \|\bar{Y}_1^n - \theta_0\|^2 \tag{7.3.32}$$

It results from the strong law of large numbers [Loeve, 1964] that, under the distribution $\mathbf{P}_{\theta_1}$, $\bar{Y}_1^n \rightarrow_{w.p.1} \theta_1$. Then, by the continuity theorem [Borovkov, 1984] and under $\mathbf{P}_{\theta_1}$, we get

$$(\chi_1^n)^2 \rightarrow_{w.p.1} b^2 \tag{7.3.33}$$

because $\|\theta_1 - \theta_0\| = b$. Now let us define the following function :

$$f_n(x) = \frac{\tilde{S}_1^n(x)}{n} \tag{7.3.34}$$

considered as a function of $(\chi_1^n)^2$. It can be proven that $f_n(x)$ converges, when n goes to infinity and uniformly in x, toward the limit :

$$f(x) = -\frac{b^2}{2} + b\sqrt{x} \tag{7.3.35}$$

For $r = 1$, this result is a straightforward consequence of (2.4.9). For $r > 1$, this convergence results from the approximation (7.2.27) of the hypergeometric function G. It can be proven that, from (7.3.33), (7.3.35), and the uniform continuity of $f(x)$,

$$f_n((\chi_1^n)^2) \rightarrow_{w.p.1} f(b^2) \tag{7.3.36}$$

Finally, we get

$$\frac{\ln \tilde{\Lambda}_1^n}{n} \rightarrow_{w.p.1} \varrho = \frac{b^2}{2} = \mathbf{K}(\theta_1, \theta_0) \tag{7.3.37}$$

We now prove (7.3.29) and (7.3.30) for the one-dimensional and multidimensional cases separately.

One dimensional case We first prove (7.3.29). As we discussed in chapter 2, when $\sigma = 1$ and thus $b = \nu$, the decision function (2.4.9) is

$$\frac{\ln \tilde{\Lambda}_1^n}{n} = \frac{\ln \cosh[bn(\bar{y}_1^n - \mu_0)]}{n} - \frac{b^2}{2} \tag{7.3.38}$$

where $\theta_0 = \mu_0$. Let us estimate the large deviation probability, and first find an upper bound :

$$\begin{aligned} p_n = \mathbf{P}_{\theta_1}\left(\frac{\ln \tilde{\Lambda}_1^n}{n} < \tilde{\varrho}\right) &< \mathbf{P}_{\theta_1}\left[|b(\bar{y}_1^n - \mu_0)| - \frac{b^2}{2} - \frac{\ln 2}{n} < \tilde{\varrho}\right] \\ &= \mathbf{P}_{\theta_1}\left[|\bar{y}_1^n - \mu_0| < \frac{1}{2}(1+\alpha)b + \frac{\ln 2}{nb}\right] \\ &< \phi\left[-\sqrt{n}(b-c) + \frac{\ln 2}{b\sqrt{n}}\right] \end{aligned} \tag{7.3.39}$$

where $\alpha \in (0,1)$, $c = \frac{1}{2}(1+\alpha)b$ and ϕ is the cdf of the standard Gaussian distribution. From this and the asymptotic formula

$$\phi(-x) \sim \frac{1}{x\sqrt{2\pi}} e^{-\frac{x^2}{2}} \left(1 - \frac{1}{x^2} + \frac{3}{x^4} + \ldots\right) \text{ when } x \rightarrow +\infty \tag{7.3.40}$$

we deduce

$$\lim_{n\rightarrow\infty} np_n = 0 \tag{7.3.41}$$

Let us now prove (7.3.30). Using again the above-mentioned upper bound for p_n, we deduce that

$$\sum_{i=1}^{n} p_n < \sum_{i=1}^{n} \tilde{p}_n \tag{7.3.42}$$

where

$$\tilde{p}_n = \frac{1}{\sqrt{2\pi}[(b-c)\sqrt{n} - \frac{\ln 2}{b\sqrt{n}}]} \ \exp\left\{-\frac{1}{2}\left[(b-c)\sqrt{n} - \frac{\ln 2}{b\sqrt{n}}\right]^2\right\} \tag{7.3.43}$$

Using the D'Alembert's criterion for convergence of series, we now have to prove that

$$\lim_{n\to\infty} \frac{\tilde{p}_{n+1}}{\tilde{p}_n} < 1 \tag{7.3.44}$$

Straightforward computations show that this last inequality actually holds.

This completes the proof of (7.3.27) in the case of a scalar parameter. Note that we proved that

$$\mathbf{E}_{\theta_1}(T) \sim \frac{h}{\varrho} \ \text{ as } \ h \to \infty \tag{7.3.45}$$

where ϱ is defined in (7.3.37). When compared to Lorden's theorem, this result shows us that the worst mean delay satisfies

$$\bar{\tau}^* \leq \mathbf{E}_{\theta_1}(T) \sim \frac{h}{\mathbf{K}(\theta_1, \theta_0)} \ \text{ as } \ h \to \infty \tag{7.3.46}$$

In subsection 5.2.4 we got a slightly different relation (5.2.106) for $\bar{\tau}^*$, but the difference is negligible from the point of view of order of magnitude.

Multidimensional case We now prove (7.3.29) and (7.3.30) in the case of a multidimensional parameter. We estimate an upper bound for the large deviation probability $p_n = \mathbf{P}_{\theta_1}\left(\frac{\ln \tilde{\Lambda}_1^n}{n} < \tilde{\varrho}\right)$ using the following reasoning. From (7.3.31) and (7.3.34), and because f_n is an increasing function (logarithm of hypergeometric function G), we get

$$p_n = \mathbf{P}_{\theta_1}[(\chi_1^n)^2 < f_n^{-1}(\tilde{\varrho})] \tag{7.3.47}$$

We thus need only a lower bound $\tilde{f}_n$ for the function f_n, which gives the following upper bound for p_n :

$$p_n \leq \mathbf{P}_{\theta_1}[(\chi_1^n)^2 < \tilde{f}_n^{-1}(\tilde{\varrho})] \tag{7.3.48}$$

We now use the expression (7.2.27) of the function G with the aid of the hypergeometric function M and its expansion. We get

$$\begin{aligned} f_n(x) &> -\frac{b^2}{2} + b\sqrt{x} - \frac{r-1}{2n}\ln\left(2nb\sqrt{x}\right) + \frac{1}{n}\ln\frac{\Gamma(r-1)}{\Gamma\left(\frac{r-1}{2}\right)} \\ &> -\frac{b^2}{2} + b\sqrt{x} - \frac{r-1}{2n}\left(\sqrt{x} + \ln\left\{2nb + \left[\frac{\Gamma\left(\frac{r-1}{2}\right)}{\Gamma(r-1)}\right]^{\frac{2}{r-1}}\right\}\right) \qquad (7.3.49) \\ &= \tilde{f}_n(x) \qquad (7.3.50) \end{aligned}$$

Finally, it results from these computations that

$$p_n = \mathbf{P}_{\theta_1}[(\chi_1^n)^2 < f_n^{-1}(\tilde{\varrho})] \leq \tilde{p}_n = \mathbf{P}_{\theta_1}\left[(\chi_1^n)^2 < \tilde{f}_n^{-1}(\tilde{\varrho})\right] \tag{7.3.51}$$

Let us consider the random value $\xi = n(\chi_1^n)^2$, which is distributed as $\chi^2(r, \lambda_n)$, where $\lambda_n = b^2 n$ is the noncentrality parameter. When n goes to infinity, the noncentrality parameter also goes to infinity. As we showed in subsection 3.1.1, in this case the χ^2 distribution converges to the normal distribution :

$$\mathcal{L}\left[\frac{\xi - \mathbf{E}_n(\xi)}{\sigma_n}\right] \rightsquigarrow \mathcal{N}(0,1) \text{ when } n \to \infty \tag{7.3.52}$$

where $\mathbf{E}_n(\xi) = b^2 n + r$ is the expectation and $\sigma_n = \sqrt{\text{var}(\xi)}$, $\text{var}(\xi) = 4b^2 n + 2r$, is the standard deviation of ξ. Let us estimate the large deviation probability $\tilde{p}_n$,

$$\tilde{p}_n = \mathbf{P}_{\theta_1}\left(\frac{\xi}{n} < \tilde{x}_n\right) \tag{7.3.53}$$

where $\tilde{x}_n$ is computed from (7.3.49) :

$$\tilde{x}_n = \left(\frac{\tilde{\varrho} + \frac{b^2}{2}}{b - \frac{r-1}{2n}} - \frac{1}{2n} \ln\left\{2nb + \left[\frac{\Gamma\left(\frac{r-1}{2}\right)}{\Gamma(r-1)}\right]^{\frac{2}{r-1}}\right\} \frac{r-1}{b - \frac{r-1}{2n}}\right)^2 \tag{7.3.54}$$

Using the asymptotic normality of ξ, we have

$$\begin{aligned} \tilde{p}_n = \mathbf{P}_{\theta_1}\left(\frac{\xi}{n} < \tilde{x}_n\right) &= \phi\left[-\frac{n(b^2 + \frac{r}{n} - \tilde{x}_n)}{\sigma_n}\right] \\ &= \phi\left\{-\sqrt{n}(b^2 - c^2)[1 + O(n^{-1})] + O(n^{-1}\ln n)\right\} \end{aligned} \tag{7.3.55}$$

where c is as in (7.3.39). Thus, we are now in the same situation as in (7.3.39), and we can conclude the proof of (7.3.29), (7.3.30), and (7.3.46) as in the case of a one-dimensional parameter.

Finally, we have shown that the χ^2-CUSUM algorithm keeps the first-order optimal property of the CUSUM algorithm :

$$\bar{\tau}^* \sim \frac{\ln \bar{T}}{\mathbf{K}(\theta_1, \theta_0)} \text{ when } \bar{T} \to \infty \tag{7.3.56}$$

7.3.3 GLR Algorithm

We now discuss briefly the properties of the GLR algorithm (7.2.21). Let us recall that the GLR algorithm corresponds here to the same hypotheses (7.3.22) as the χ^2-CUSUM algorithm investigated in the previous subsection. For deriving the properties of the GLR algorithm, we basically use the main idea of the previous subsection. It is obvious that the asymptotic formula

$$\mathbf{E}_{\theta_1}(T) \sim \frac{h}{\varrho} \text{ as } h \to \infty \tag{7.3.57}$$

for the mean delay holds true if we replace $\tilde{S}_1^n/n$ by $\hat{S}_1^n/n$, where

$$\frac{\hat{S}_1^n}{n} = \frac{1}{n} \ln \frac{\sup_{\|\theta_1 - \theta_0\| = b} \prod_{i=1}^n p_{\theta_1}(Y_i)}{\prod_{i=1}^n p_{\theta_0}(Y_i)} = -\frac{b^2}{2} + b\chi_1^n = f[(\chi_1^n)^2] \tag{7.3.58}$$

For proving this, the same arguments as in the previous subsection can be used.

Unfortunately, for $r > 1$, it is not obvious how to compute a lower bound for the mean time between false alarms $\bar{T}$. It results from (7.2.30) that, for a given threshold h, the generatrix $\hat{c}_n$ of the stopping surface

Table 7.1 Comparison between the "exact" value of $\bar{\tau}^*$ for the two-sided CUSUM algorithm for $r = 1$, the asymptotic formula (7.3.46) for the χ^2-CUSUM algorithm, and simulation of this algorithm for $b = 1, r = 1, 2, 10$, and 40.

h	Fredholm	(7.3.46)	Simulation			
	$r = 1$	$r \geq 1$	$r = 1$	$r = 2$	$r = 10$	$r = 40$
5	11.8	10	11.9 ± 0.3	13.5 ± 0.3	21.1 ± 0.4	24.8 ± 0.2
10	21.8	20	22.2 ± 0.4	23.8 ± 0.4	34.7 ± 0.5	41.5 ± 0.7
15	31.7	30	32.0 ± 0.5	34.2 ± 0.5	46.4 ± 0.6	61.0 ± 0.9
20	41.7	40	42.3 ± 0.6	44.8 ± 0.6	58.9 ± 0.7	78.1 ± 0.9
25	51.7	50	51.9 ± 0.7	55.2 ± 0.6	70.6 ± 0.8	94.0 ± 1.0
30	61.7	60	62.2 ± 0.7	65.3 ± 0.7	82.2 ± 0.8	109.8 ± 1.1
35	71.7	70	72.3 ± 0.8	75.3 ± 0.8	93.4 ± 0.9	123.4 ± 1.1
40	81.7	80	81.5 ± 0.8	85.5 ± 0.8	104.8 ± 0.9	137.0 ± 1.2
45	91.7	90	91.5 ± 0.9	96.2 ± 0.9	115.6 ± 1.0	152.8 ± 1.2
50	101.7	100	101.3 ± 0.9	106.4 ± 0.9	126.8 ± 1.0	164.8 ± 1.2

of the GLR is less than the generatrix $\tilde{c}_n$ of the χ^2-CUSUM. For this reason, the following inequality takes place :

$$\mathbf{P}_{\theta_0}(\min\{n \geq 1 \ : \tilde{S}_1^n \geq h\} < \infty) \leq \mathbf{P}_{\theta_0}(\min\{n \geq 1 \ : \hat{S}_1^n \geq h\} < \infty) \tag{7.3.59}$$

Therefore we have :

$$\bar{T}_{\text{GLR}} \leq \bar{T}_{\text{CUSUM}} \tag{7.3.60}$$

7.3.4 Simulation Results

The purpose of this subsection is to compare, for the χ^2-CUSUM algorithm, the asymptotic formulas for the mean delay $\bar{\tau}^*(h)$ (7.3.46), the mean time between false alarms $\bar{T}(h)$ (7.3.26), and the first-order optimality $\bar{\tau}^*(\bar{T})$ (7.3.56), which we derived before, with results of simulation of this change detection algorithm. The main aim here is an investigation of the multidimensional case. But we also add results for the one-dimensional case $r = 1$, because in this case we can compare asymptotic formulas with "exact" results of Fredholm integral equations, and this can help in guessing which precision can be expected in the multidimensional case.

7.3.4.1 Experiment 1 - Mean Delay for Detection

Because we are estimating the *worst* mean delay, we assume that the change occurred at the beginning of the sample (see the discussion in section 4.4). Therefore, the simulations consist of generating, for each value of the threshold h, 500 i.i.d. sequences of pseudo-random Gaussian vectors with unit covariance matrix and with mean $\theta_1 = \frac{1}{\sqrt{r}} \mathbb{1}_r$. In other words, we use the *same* Kullback information (or equivalently signal-to-noise ratio $b = 1$) for different dimensions r. Then we estimate the empirical mean and standard deviation of the resulting mean delays for detection, which are shown in columns 4-7 of table 7.1.

Scalar Case First we discuss the scalar case, and compare the "exact," asymptotic, and empirical delays. The results of these comparisons are summarized in columns 2-4 of table 7.1 for values of the threshold h between 5 and 50 shown in the first column.

As we explained in subsections 2.4.2 and 5.2.4, in the scalar case, the stopping boundaries of the χ^2-CUSUM and two-sided CUSUM algorithms are asymptotically equivalent. Therefore, for the "exact" value

Table 7.2 Comparison between the "exact" value of $\bar{T}$ for the two-sided CUSUM algorithm for $r = 1$, the asymptotic formula (7.3.26) for the χ^2-CUSUM algorithm, and simulation of this algorithm for $b = 1$, $r = 1, 210$, and 40.

h	Fredholm	(7.3.26)	Simulation			
	$r = 1$	$r \geq 1$	$r = 1$	$r = 2$	$r = 10$	$r = 40$
1	13.3	2.72	12.6 ± 1.1	12.5 ± 1.0	18.5 ± 1.3	22.5 ± 1.3
2	42.0	7.39	39.3 ± 4.1	47 ± 4.7	55.0 ± 5.2	70.1 ± 4.8
3	121.5	20.10	109 ± 10	124.4 ± 11.6	155.6 ± 12.8	177.5 ± 13.0
4	340.0	54.60	315 ± 31	324.5 ± 29.5	369.5 ± 32.2	416.2 ± 38.6

of the delay in this case, we use the result of the numerical solution of the Fredholm integral equation for the two-sided CUSUM algorithm discussed in the first example of subsection 5.2.2. We show this in the second column.

The table proves again that, in the scalar case, the χ^2-CUSUM and two-sided CUSUM algorithms are equivalent, as is obvious from a comparison of columns 2 and 4. Moreover, a comparison of column 3 and columns 2 and 4 shows that the asymptotic formula (7.3.46) for $\bar{\tau}^*$ underestimates the true value of the mean delay because it ignores the excess over the threshold h. However, this precision is sufficient in practice.

Multidimensional Case A comparison of column 3 and columns 5-7 shows again that the asymptotic formula (7.3.46) is convenient in practice, especially for high values of the threshold, but that it still underestimates the mean delay.

7.3.4.2 Experiment 2 - Mean Time Between False Alarms

Since we are estimating the mean time between false alarms, we assume that there is no change. Therefore, the simulations consist of generating, for each value of the threshold h, 100 i.i.d. sequences of pseudo-random Gaussian vectors with unit covariance matrix and with mean $\theta_0 = 0$. Then we estimate the empirical mean and standard deviation of the resulting mean times between false alarms, which are shown in columns 4-7 of table 7.2.

As before, it results from the comparison between the columns 2 and 4 that the χ^2-CUSUM and two-sided CUSUM algorithms are equivalent even for small values of the threshold. Columns 4-7 of table 7.2 also show that the asymptotic formula (7.3.26) underestimates the true value of the mean time between false alarms for $r \geq 1$, again because the excess over the threshold is ignored as in the previous case. Note that this is predictable from the inequality in (7.3.26) itself!

7.3.4.3 Experiment 3 - First-Order Optimality

The aim of this experiment is to compare the main result (7.3.56) of the previous subsection, which is the asymptotic relation between $\bar{\tau}^*$ and $\bar{T}$. The results of this comparison are presented in table 7.3 for $r = 2$, using values obtained in the two previous experiments (with lower values of the threshold). This table shows that the precision of the formula (7.3.56) is sufficient in practice, even for small values of $\bar{T}$.

7.4 Geometrical Approach

In this section, we describe several geometrical tools for failure detection, which are known in the control literature as *analytical redundancy* and *parity spaces*. Following the main lines of this book, the goal of this

Table 7.3 Comparison between the asymptotic formula (7.3.56) for $\bar{\tau}^*$ and simulation for the χ^2-CUSUM algorithm for $b = 1, r = 2$.

$\bar{T}$	(7.3.56)	Simulation
12.5	5.1	4.25 ± 0.12
47.0	7.7	6.68 ± 0.17
124.4	9.7	8.68 ± 0.20
324.5	11.6	11.06 ± 0.24

section is to introduce some basic tools upon which the geometrical approach relies, in order to emphasize the *basic links* between the statistical and geometrical approaches. We do not pretend to give an exhaustive discussion of the geometrical issues; for this we refer the reader to the surveys [Mironovski, 1980, Frank and Wünnenberg, 1989, Frank, 1990, Gertler, 1991, Patton and Chen, 1991] and the books [Viswanadham *et al*., 1987b, Patton *et al*., 1989]. We simply introduce several key ideas for the purpose of establishing some links.

The additive failure models that are to be assumed in this section are basically the regression and state-space models in (7.1.4) and (7.1.6), and, to a lesser extent, the input-output transfer function $\mathcal{T}_U$ in (7.2.101). The key tools upon which the analytical redundancy techniques rely are bases for the orthogonal complement of the observability matrix, stable factorizations of the transfer function $\mathcal{T}_U$, and observers design.

In some sense, the methods that we describe here follow basically a one-model approach, as discussed in section 1.4 about prior information. We mean that, for the *detection* problem, these methods do not use prior information and models about the type of changes that are to occur. We distinguish between two classes of methods. In the first class, we consider methods that work basically with the unfailed model (7.1.3) and that deal with a one-model approach for solving detection problems only. This is the case of the methods described in this section, namely the direct and temporal redundancy and the generalized parity space approaches. In the second class, we consider methods that follow in some sense a two-model approach and directly incorporate the failure models (7.1.6), and that most often solve the detection and the isolation problems simultaneously. This is the case of the unknown input observers and detection filters approaches, which we do not describe here; we refer the reader to the above-mentioned survey papers and other references given at the end of this chapter.

We discuss the geometrical failure detectability issue in subsection 7.4.4.

The connections between these geometrical tools and the statistical tools introduced in section 7.2 are investigated in section 7.5. The result is that, in some sense, they monitor two different functions of the same sufficient statistic. The advantage of the statistical approach based upon the likelihood ratio is that it automatically *takes into account sensor noise and calibration problems*. These links are discussed for the residual generation problem in regression and state-space models, and also for the diagnosis problem and the detectability issue.

7.4.1 Direct Redundancy

We now introduce the first geometrical failure detection technique, which is known as indirect or analytical redundancy and which is an extension of the direct or physical redundancy that we first introduce in this subsection. We mainly follow [Viswanadham *et al*., 1987b, Ray and Luck, 1991].

Direct redundancy refers to a situation where several sensors measure the same physical quantity. Let us thus consider the following unfailed model :

$$Y_k \quad = \quad HX_k \quad + \quad V_k \tag{7.4.1}$$

where the state X and the observation Y have dimensions n and $r > n$, respectively, H is of rank n, and $(V_k)_k$ is a white noise sequence with covariance matrix R. In the present situation of multiple identical sensors measuring a same quantity, we can assume that

$$R = \sigma^2 I_r \tag{7.4.2}$$

and this assumption is of nonnegligible importance, as we discuss later. A typical example of (7.4.1) is a four-dimensional measurement vector Y $(r = 4)$, which is the output of a quadruplicate set of one-dimensional sensors estimating a three-dimensional unknown physical parameter $(n = 3)$, for example, accelerometers in the three dimensions of the real space, as encountered in inertial navigation systems.

We consider here additive changes as modeled in (7.1.4) and subsection 7.2.2. The measurements Y can be combined into a set of $(r - n)$ linearly independent equations by projection onto the left null space of H.

Definition 7.4.1 (Parity vector and parity space). *Let C be a $(r-n) \times r$ matrix such that its $(r-n)$ rows are an orthonormal basis of the left null space of H, i.e., such that*

$$\begin{aligned} CH &= 0 \\ CC^T &= I_{r-n} \end{aligned} \tag{7.4.3}$$

The vector ζ_k defined by

$$\zeta_k = CY_k \tag{7.4.4}$$

is called the parity vector *[Potter and Suman, 1977] and is a measure of the relative consistencies between the redundant measurements Y_k. The column space of C is called the* parity space *$\mathcal{S}$ of H.*

Note that

$$\zeta_k = CV_k \tag{7.4.5}$$

i.e., ζ_k is independent of the true value of X, is zero in the noise free case, and reflects only measurements errors including failures. The columns of C define r distinct failure directions associated with the r measurements. This is because the failure of the ith measurement implies the growth of the parity vector ζ in the direction of the ith column of C. This can be used for the purpose of diagnosis as we explain after.

Consider now the *residual vector* resulting from the least-squares estimation of X, namely

$$e_k = Y_k - H\hat{X}_k \tag{7.4.6}$$

where

$$\hat{X}_k = \arg\min_X \|Y - HX\|^2 \tag{7.4.7}$$

We already used the fact that

$$\min_X \|Y - HX\|^2 = \|P_H^* Y\|^2 \tag{7.4.8}$$

where P_H^* is the projection matrix defined by

$$P_H^* = I_r - H(H^T H)^{-1} H^T \tag{7.4.9}$$

namely, the orthogonal projection onto the orthogonal complement of the range of H - or equivalently onto the left null space of H. Therefore,

$$e_k = P_H^* Y_k \tag{7.4.10}$$

But, because of the two conditions in (7.4.3), C is necessarily such that

$$\begin{aligned} C^T C &= I_r - H(H^T H)^{-1} H^T \\ &= P_H^* \end{aligned} \tag{7.4.11}$$

Thus,

$$\begin{aligned} e_k &= C^T C Y_k \\ &= C^T \zeta_k \end{aligned} \tag{7.4.12}$$

and because of (7.4.3) again, we get

$$\|e_k\|^2 = \|\zeta_k\|^2 \tag{7.4.13}$$

In other words, the magnitude of the parity vector ζ_k is the same as the magnitude of the residual e_k. It is worth emphasizing that this is true only under the hypothesis (7.4.2) of equal noise variances, and raises the problems of calibration when variances are different and of sensor correlations when R is not even diagonal. This question is addressed in subsection 7.5.1.

The magnitude of the residual e_k grows in time when a failure occurs. In the direct parity approach, the detection is thus based upon the squared norm of the residual e_k defined in (7.4.10). In this sense, direct redundancy is a one-model approach.

Furthermore, a simple isolation or diagnosis decision scheme can be simply deduced from the properties of the parity space stated before. Actually, as we already mentioned, the columns c_i of C define r distinct failure directions associated with the r measurements. Thus, a simple diagnosis rule consists of looking for the vector c_i that is the closest to the parity check ζ, namely the c_i for which the *correlation* $c_i^T \zeta$ between the precomputed change signature c_i and the parity check ζ is maximum, or equivalently for which the angle is minimum [Ray and Desai, 1984, Viswanadham *et al*., 1987b, Patton and Chen, 1991]. Note that this technique is very close to a statistical approach. The only thing that is *not* achieved in a statistical framework is the design of the precomputed signatures.

Finally, it is of interest to note that the technique of *least-squares residual vector in a regression model* is also useful for investigating the concept of analytical redundancy in state-space models, as we describe now.

7.4.2 Analytical Redundancy

Here we follow [E.Chow and Willsky, 1984, Lou *et al*., 1986, Viswanadham *et al*., 1987b]. Analytical redundancy can be defined as the set of all nontrivial instantaneous or temporal relationships existing between the inputs and outputs of the dynamic system (7.1.3), and which are ideally zero when no fault occurs. We explain how such relationships can be derived using either a time domain or a transfer function representation of the system.

A trivial example of a direct redundancy relation is given by algebraic invariants, which are constant algebraic relations among the outputs Y, the value of which does not depend upon the inputs U. Such invariants often come from physical laws of conservation of energy, charge, heat, or static balance equations or plausibility checks, and are related to the symmetry groups of the differential equations describing the system.

The definitions of parity vector and space given in subsection 7.4.1 can be extended to the dynamic model (7.1.3) in the following manner. To combine measured and estimated outputs as before, we first make use of this unfailed model, which we rewrite here :

$$\left\{ \begin{array}{rcl} X_{k+1} & = & FX_k + GU_k + W_k \\ Y_k & = & HX_k + JU_k + V_k \end{array} \right. \tag{7.4.14}$$

where W and V have covariance matrices Q and R, respectively. Basically, instead of considering the left null space of the *observation matrix* H, we consider the left null space of the *observability matrix* $\mathcal{O}_n(H,F)$ defined by

$$\mathcal{O}_n(H,F) = \left(\begin{array}{c} H \\ HF \\ \vdots \\ HF^{n-1} \end{array} \right) \tag{7.4.15}$$

Definition 7.4.2 (Parity space of order l). *The* parity space of order l $(1 \leq l \leq n)$ *is the left null space of the observability matrix, namely the set*

$$\mathcal{S}_l = \textit{span}\ \{v | v^T \mathcal{O}_l(H,F) = 0\} \tag{7.4.16}$$

where span denotes the linear space spanned by the considered vectors.

Note [Lou *et al*., 1986] that this parity space is different from the $(l-1)$-step unobservable subspace that corresponds to the right null space of $\mathcal{O}_l(H,F) = \mathcal{O}_l$. Following the computations made in subsections 3.2.2 and 7.2.5, let us use repeatedly the equations (7.4.14) to get

$$\mathcal{Y}_{k-l+1}^k = \mathcal{O}_l X_{k-l+1} + \mathcal{J}_l(G,J)\mathcal{U}_{k-l+1}^k + \mathcal{J}_l(I_n,0)\mathcal{W}_{k-l+1}^k + \mathcal{V}_{k-l+1}^k \tag{7.4.17}$$

where $\mathcal{W}_{k-l+1}^l$ and $\mathcal{V}_{k-l+1}^l$ have covariance matrices $I_l \otimes Q$ and $I_l \otimes R$, respectively. The lower triangular matrix $\mathcal{J}_l(G,J)$ is defined in (7.2.146). Let us rewrite this *regression model* as follows :

$$\tilde{\mathcal{Y}}_{k-l+1}^k = \mathcal{O}_l\ X_{k-l+1} + \tilde{\mathcal{V}}_{k-l+1}^k \tag{7.4.18}$$

where the input-adjusted observation is

$$\tilde{\mathcal{Y}}_{k-l+1}^k = \mathcal{Y}_{k-l+1}^k - \mathcal{J}_l(G,J)\ \mathcal{U}_{k-l+1}^k \tag{7.4.19}$$

and the noise input

$$\tilde{\mathcal{V}}_{k-l+1}^k = \mathcal{J}_l(I_n,0)\ \mathcal{W}_{k-l+1}^k + \mathcal{V}_{k-l+1}^k \tag{7.4.20}$$

has a covariance matrix $\mathcal{R}_l$, which is given in (7.2.151). Now, with each vector v in (7.4.16) we can associate a scalar *parity check* ζ defined by

$$\zeta_k = v^T \left[\mathcal{Y}_{k-l+1}^k - \mathcal{J}_l(G,J)\ \mathcal{U}_{k-l+1}^k \right] = v^T\ \tilde{\mathcal{Y}}_{k-l+1}^k \tag{7.4.21}$$

Because of (7.4.18), and using (7.4.16), we have

$$\begin{array}{rcl} \zeta_k & = & v^T\ \mathcal{O}_l(H,F)\ X_{k-l+1} + v^T\ \tilde{\mathcal{V}}_{k-l+1}^k \\ & = & v^T\ \tilde{\mathcal{V}}_{k-l+1}^k \end{array} \tag{7.4.22}$$

Equation (7.4.21) should be compared with (7.4.4). It is obvious from (7.4.22) that a parity check or temporal redundancy relation does not depend upon the nonmeasured state X as in the direct case, but (7.4.21) shows

that it is a linear combination of present and past inputs and outputs to the system (7.4.14), as opposed to a direct redundancy relation which is a linear combination of present outputs only. Because of (7.4.16), the geometrical interpretation of the parity space is the same as before for the direct redundancy case : it is the orthogonal complement of the range of $\mathcal{O}_l(H,F)$ instead of H.

Such a parity equation can be used to compute recursively an estimate $\hat{y}_k$ of some of the observations, using the relationship $\zeta_k = y_k - \hat{y}_k$. Note that the resulting recursive equation can have, in the presence of noise, different behavior from the equation for ζ, in terms of statistical properties, signature of a given change, and instabilities [Patton and Chen, 1991].

There are several ways to compute parity checks as defined in (7.4.21). One way consists of the orthogonal projection of the input-adjusted observations $\left(\mathcal{Y}_{k-l+1}^{k} - \mathcal{J}_l(G,J)\,\mathcal{U}_{k-l+1}^{k}\right)$ onto $\mathcal{S}_l$, which is the orthogonal complement of the range of $\mathcal{O}_l = \mathcal{O}_l(H,F)$. Let P_l^* be the matrix associated with this projection. We have

$$P_l^* = I_{lr} - \mathcal{O}_l \left(\mathcal{O}_l^T \mathcal{O}_l\right)^{-1} \mathcal{O}_l^T \tag{7.4.23}$$

As usual, we need to ensure the invertibility of the matrix inside the parentheses in (7.4.23), which depends upon the choice of l. If the system is observable, namely if rank $\mathcal{O}_n = n$, an obvious relevant choice is $l = n$. If the system is not observable, then the order l of the parity space has to be chosen in terms of the observability index, as discussed in [E.Chow and Willsky, 1984, Lou *et al.*, 1986].

To pursue the parallelism with the direct redundancy, let us now consider again equation (7.4.18), and discuss the least-squares estimation of the unknown value of state X. The key issue here is that X_{k-l+1} is independent of $\tilde{\mathcal{V}}_{k-l+1}^{k}$, as we explain in subsection 7.2.5.

Let us assume for the moment that

$$\mathcal{R}_l = \sigma^2 I_{lr} \tag{7.4.24}$$

exactly as we do in (7.4.2) for direct redundancy. Therefore, the standard solution (7.4.8) to least-squares estimation in a regression model can be applied, namely

$$\hat{X}_{k-l+1} = \left(\mathcal{O}_l^T \mathcal{O}_l\right)^{-1} \mathcal{O}_l^T \tilde{\mathcal{Y}}_{k-l+1}^{k} \tag{7.4.25}$$

Now, let us define the corresponding *residual vector* by

$$\begin{aligned} e_k &= \tilde{\mathcal{Y}}_{k-l+1}^{k} - \mathcal{O}_l\, \hat{X}_{k-l+1} \\ &= P_l^*\, \tilde{\mathcal{Y}}_{k-l+1}^{k} \end{aligned} \tag{7.4.26}$$

where P_l^* is exactly as in (7.4.23). Note that e_k in (7.4.26) is defined in a similar way as in (7.4.10), and is the residual of the least-squares smoothing of state X. The residual e_k can be thought of as being made of a collection of parity checks ζ_k corresponding to *independent* vectors v as in (7.4.21), the rows of P_l^* being made of these v. The decision function associated with analytical redundancy is thus to monitor the norm of the residual e_k as before, and results basically in a one-model approach.

Let us emphasize that the assumption (7.4.24) is never true when the system is dynamic and not static, because in this case $\check{\mathcal{J}} \neq I$, as the expression (7.2.151) of $\mathcal{R}$ shows. In this case, the analytical redundancy approach consists again of monitoring the norm of the vector of parity checks ζ_k, which is now different from the residual e_k. The link between this approach and the GLR algorithm is investigated in subsection 7.5.

7.4.3 Generalized Parity Checks

In this subsection, we describe another geometrical technique for residual generation, which is based upon the input-output *transfer function* representation (7.2.95). The analytical redundancy relations and parity

spaces have been defined basically using a *time domain* representation of the dynamic system and *polynomial* relations. Similar derivations are also possible for *rational* redundancy relations when starting from an input-output transfer function representation. This can be achieved with the aid of stable factorizations, as shown in [Viswanadham *et al*., 1987a]. Again, this is mainly a one-model approach as far as detection is concerned, because it makes use of the only deterministic input-output transfer function associated with the unfailed model (7.4.14), namely

$$Y_k = \mathcal{T}_U(z)U_k \tag{7.4.27}$$

First, note [Lou *et al*., 1986, Gertler, 1991, Patton and Chen, 1991] that multiplying the quantity $Y_k - \mathcal{T}_U(z)U_k$ by any polynomial or rational transfer function gives rise to various parity checks, achieving desired properties. More precisely, because of (7.4.21), any parity check is of the form

$$\zeta_k = A(z)Y_k + B(z)U_k \tag{7.4.28}$$

where the transfer function matrices A and B are stable and satisfy the following condition :

$$A(z)\mathcal{T}_U(z) + B(z) = 0 \tag{7.4.29}$$

Imposing different structures for A and B results in different residual generations. Let us now explain one very powerful example of such design.

Assume that the rational transfer function (7.4.27) is proper and stable. We can define the stable right coprime factorization (rcf) of $\mathcal{T}_U$ as [Kailath, 1980, Vidyasagar, 1985]

$$\mathcal{T}_U(z) = N(z)D^{-1}(z) \tag{7.4.30}$$

where the *rational* (and not necessarily polynomial, as before) functions N and D have no common zeroes in the right half complex plane. The corresponding Bezout identity is

$$A(z)N(z) + B(z)D(z) = I_m \tag{7.4.31}$$

and, similarly, the stable left coprime factorization (lcf)

$$\mathcal{T}_U(z) = \tilde{D}^{-1}(z)\tilde{N}(z) \tag{7.4.32}$$

with the corresponding Bezout identity

$$\tilde{N}(z)\tilde{A}(z) + \tilde{D}(z)\tilde{B}(z) = I_r \tag{7.4.33}$$

where A, B, $\tilde{A}$, and $\tilde{B}$ are stable. Because of the equality between (7.4.30) and (7.4.32), and because of (7.4.31) and (7.4.33), the following generalized Bezout identity holds :

$$\begin{pmatrix} B(z) & A(z) \\ -\tilde{N}(z) & \tilde{D}(z) \end{pmatrix} \begin{pmatrix} D(z) & -\tilde{A}(z) \\ N(z) & \tilde{B}(z) \end{pmatrix} = \begin{pmatrix} I_m & 0 \\ 0 & I_r \end{pmatrix} \tag{7.4.34}$$

Now these rcf and lcf of $\mathcal{T}_U(z)$ can be obtained from a state-space description of $\mathcal{T}_U(z)$. Let us assume that (7.4.14) is a *stabilizable* and *detectable* system. Then there exist two gain matrices K and L such that $\bar{F} = F - FKH$ and $F - GL$ are stable. Note that the choice of the Kalman filter gain as a matrix K is admissible. The input-output transfer function associated with this deterministic state-space model is defined by

$$\mathcal{T}_U(z) = H(zI_n - F)^{-1}G + J \tag{7.4.35}$$

and its rcf and lcf are given by

$$\begin{pmatrix} B(z) & A(z) \\ -\tilde{N}(z) & \tilde{D}(z) \end{pmatrix} = \begin{pmatrix} I_m & 0 \\ -J & I_r \end{pmatrix} + \begin{pmatrix} L \\ -H \end{pmatrix} (zI_n - \bar{F})^{-1} \begin{pmatrix} (G - FKJ) & FK \end{pmatrix} \tag{7.4.36}$$

and

$$\begin{pmatrix} D(z) & -\tilde{A}(z) \\ -N(z) & \tilde{B}(z) \end{pmatrix} = \begin{pmatrix} I_m & 0 \\ J & I_r \end{pmatrix} - \begin{pmatrix} L \\ -(H - JL) \end{pmatrix} (zI_n - F + GL)^{-1} \begin{pmatrix} G & FK \end{pmatrix} \tag{7.4.37}$$

which can be rewritten as

$$\begin{array}{rcccl} N(z) & = & -J & - & (H - JL)(zI_n - F + GL)^{-1}G \\ D(z) & = & I_m & - & L(zI_n - F + GL)^{-1}G \end{array} \tag{7.4.38}$$

and

$$\begin{array}{rcccl} \tilde{N}(z) & = & J & + & H(zI_n - \bar{F})^{-1}(G - FKJ) \\ \tilde{D}(z) & = & I_r & - & H(zI_n - \bar{F})^{-1}FK \end{array} \tag{7.4.39}$$

These results are used for the design of generalized parity relations in the following manner [Viswanadham *et al*., 1987a]. Let us define a *generalized parity vector*,

$$\zeta_k = \tilde{D}(z)\,[Y_k - \mathcal{T}_U(z)U_k] \tag{7.4.40}$$

$$= \tilde{D}(z)Y_k - \tilde{N}(z)U_k \tag{7.4.41}$$

where the second equality holds because of the lcf (7.4.32). This is an extension of the parity vectors defined in (7.4.4) for the direct redundancy and in (7.4.21) for the temporal redundancy. Furthermore, this generalized parity vector is also equivalent to the detection filter approach to be presented next. Note that it is possible to design specific parity vectors for monitoring (subsets of) actuators or sensors using a selection of appropriate rows of $\tilde{N}$ or $\tilde{D}$.

The two-model approach corresponding to this factorization technique for generating parity checks consists of computing the signature of the changes as modeled in (7.2.97) on the parity checks ζ defined in (7.4.40). Following (7.2.109), we denote this signature by ϱ, namely

$$\zeta_k = \zeta_k^0 + \varrho(k, t_0) \tag{7.4.42}$$

Using the transfer function expression of the additive change model (7.2.101), the factorization (7.4.32), and the closed form expression of $\tilde{D}$ given in (7.4.39), it is straightforward to show that

$$\begin{array}{rcl} \varrho(k, t_0) & = & \mathcal{H}_x(z)\Upsilon_x(k, t_0) + \mathcal{H}_y(z)\Upsilon_y(k, t_0) \\ \text{where } \mathcal{H}_x(z) & = & \left[I_r - H(zI_n - \bar{F})^{-1}FK\right] H(zI_n - F)^{-1}\Gamma \\ \text{and } \mathcal{H}_y(z) & = & \left[I_r - H(zI_n - \bar{F})^{-1}FK\right] \Xi \end{array} \tag{7.4.43}$$

We use this result when discussing the basic links between the statistical and geometrical approaches for additive change detection, and also when investigating the geometrical detectability.

7.4.4 Geometrical Detectability

We discussed this detectability issue in the context of the statistical approach in subsection 7.2.6. We now investigate it in the framework of the geometrical approach, following a definition introduced in [Caglayan, 1980]. We show the link between the two points of view in section 7.5.

Referring to the discussion we had in section 6.3 about intrinsic and detection-based definitions of failure detectability, we now consider *intrinsic* definitions of detectability, but using only the deterministic part of the system. These definitions are in terms of the observability and detectability of the augmented system which combines the state X and the failures of (7.2.97).

We assume that the change profiles are constant, namely that the two changes, on the state and observation equations in (7.2.97), are simply *steps*. Thus, we use the model

$$\left\{ \begin{array}{rcl} X_{k+1} &=& FX_k + GU_k + W_k + \Gamma\,\Upsilon_x \mathbf{1}_{\{k\geq t_0\}} \\ Y_k &=& HX_k + JU_k + V_k + \Xi\,\Upsilon_y \mathbf{1}_{\{k\geq t_0\}} \end{array} \right. \tag{7.4.44}$$

which we rewrite using an extended state $\mathcal{X} = \begin{pmatrix} X \\ \Upsilon_x \mathbf{1}_{\{k\geq t_0\}} \\ \Upsilon_y \mathbf{1}_{\{k\geq t_0\}} \end{pmatrix}$ in the following manner :

$$\left\{ \begin{array}{rcl} \mathcal{X}_{k+1} &=& F_\Upsilon \mathcal{X}_k + \begin{pmatrix} GU_k \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} W_k \\ 0 \\ 0 \end{pmatrix} \\ Y_k &=& H_\Upsilon \mathcal{X}_k + JU_k + V_k \end{array} \right. \tag{7.4.45}$$

where

$$F_\Upsilon = \begin{pmatrix} F & \Gamma & 0 \\ 0 & I_{\tilde{n}} & 0 \\ 0 & 0 & I_{\tilde{r}} \end{pmatrix} \tag{7.4.46}$$

and

$$H_\Upsilon = \begin{pmatrix} H & 0 & \Xi \end{pmatrix} \tag{7.4.47}$$

Two possible definitions of the geometrical detectability can be given.

Definition 7.4.3 (Strong geometrical detectability). *The input jumps modeled in (7.4.44) are* strongly detectable *if the system defined by the pair (H_Υ, F_Υ) is observable.*

This definition can be thought of as being too strong, because, from an intuitive point of view, the only condition that is reasonably needed is that the transfer function from Υ toward Y, or toward ε or ζ, is full rank. But we show later that all these requirements result in the same conditions on the system matrices in (7.4.44).

Let us introduce a similar but weaker definition.

Definition 7.4.4 (Weak geometrical detectability). *The input jumps modeled in (7.4.44) are* weakly detectable *if the system defined by the pair (H_Υ, F_Υ) is detectable.*

Recall that a pair (H, F) is detectable if and only if every unstable mode of F is observable.

Now it can be proven [Caglayan, 1980] that the pair (H_Υ, F_Υ) is observable (respectively detectable) if and only if the pair (H, F) is observable (respectively detectable) and

$$\begin{array}{rcl} \text{rank } H(I_n - \bar{F})^{-1}\Gamma &=& \tilde{n} \\ \text{rank } \left[I_r - H(I_n - \bar{F})^{-1}FK\right]\Xi &=& \tilde{r} \end{array} \tag{7.4.48}$$

where K is any $n \times r$ matrix such that 1 is not an eigenvalue of $\bar{F} = F - FKH$. These two conditions imply that $r \geq \tilde{n}$ and $r \geq \tilde{r}$, in other words that there exist at least as many sensors as simultaneous failures on the state and on the observation equations.

Let us emphasize that the first consequence of this result is that both the strong and weak detectability definitions result in the same constraints between the change vectors Υ_x and Υ_y and change gains Γ and Ξ on one hand, and the system matrices on the other. Thus, the choice of the geometrical detectability definition depends only upon the assumption that is made about the unfailed system, namely on whether it is observable or simply detectable. These two conditions have to be compared with the "statistical condition" (7.2.218). This is done in section 7.5.

Another possible intrinsic definition of failure detectability is given in [Beard, 1971, White and Speyer, 1987] for the failure model given by (7.2.97) with the conditions (7.2.102) together with $J = 0, \nu_y = 0$. The *detectability* is then defined as the existence of a stable filter gain matrix K such that

$$\text{rank}\left\{H\left[\begin{array}{cccc}\Upsilon_x & (F-KH)\,\Upsilon_x & \ldots & (F-KH)^{n-1}\Upsilon_x\end{array}\right]\right\} = 1 \tag{7.4.49}$$

This condition ensures that the output error is constrained in a fixed direction when a failure occurs in the direction Υ_x, and is thus in fact a diagnosability condition.

7.5 Basic Geometrical/Statistical Links

In this section, we investigate the relationship between the statistical and geometrical approaches to additive change detection and diagnosis. As already mentioned in section 6.3, we investigate this link considering several issues : design of residuals for change detection, diagnosis, and detectability.

As far as the design of residuals and decision functions is concerned, several possibilities exist for comparing the statistical and geometrical approaches. The first consists of writing the log-likelihood ratio s_k in terms of the parity check ζ_k, showing that the statistical and geometrical tools *monitor two different functions of the same sufficient statistic*. This is what we do in subsection 7.5.1, first in the case of regression models where there is identity between the two decision functions, and then in the case of state-space models.

The second tool uses the following fact, which we prove in subsection 7.4.2. The key common feature to both the geometrical and statistical approaches and in both regression and state-space models is the *projection associated with least-squares estimation in regression models*, as is obvious from equation (7.4.18). This common tool can serve as a basis for establishing the link between both approaches. But it is important to note that, in the case of *dynamic* - and not static - systems, the projection $\tilde{P}_l$ associated with state smoothing is *not* identical to the projection P_l^* associated with analytical redundancy relations, as we show in subsection 7.5.1.

In subsection 7.5.2, considering the case of state-space models, we also focus on the generalized parity checks approach, and we use a third possibility, which consists of comparing the *signatures* of the additive changes on the Kalman filter innovation and on the generalized parity check. We also find that the transfer functions from the observations Y toward the innovation ε and toward the generalized parity check ζ are the same.

Then, in subsection 7.5.3, we discuss the *diagnosis* problem. We show that, when introducing a statistical decoupling criterion for achieving failure isolation, the first step of the resulting algorithm is nothing but a standard geometric decoupling procedure.

Finally, in subsection 7.5.4, we discuss the *detectability* issue, showing that the intuitive, geometrical, and statistical points of view basically lead to the same detectability definitions. In particular, we show that the statistical detectability definition results, on the deterministic part of the system, in conditions that are very close to the classical detectability used in system theory.

7.5.1 Analytical Redundancy and GLR

Here we compute the log-likelihood ratio s_k in terms of the parity checks ζ_k. We begin our explanation about the connection between the two approaches in the simplest case of regression models. We assume first that the measurement covariance matrix is $R = \sigma^2 I_r$, and we discuss the robustness with respect to this assumption. Then we investigate the case of state-space models.

7.5.1.1 Regression Models

In subsection 7.4.1, the direct redundancy approach was described as a one-model approach to change detection. Now let us investigate the corresponding two-model approach, namely the relevant decision function in this context when using also the information contained in the failed model associated with (7.4.1). We consider the alternative hypothesis $\mathbf{H}_1$ introduced in (7.1.4), where we assume that an additive change of magnitude ν occurs in the *constant* direction Υ, namely

$$Y_k \;=\; HX_k \;+\; V_k \;+\; \nu\,\Upsilon\,\mathbf{1}_{\{k\geq t_0\}} \tag{7.5.1}$$

where we assume that $\|\Upsilon\| = 1$.

Equal noise variances Under the no-change hypothesis $\mathbf{H}_0$ corresponding to the model (7.4.1), we have

$$e_k = P_H^* Y_k = P_H^* V_k \tag{7.5.2}$$

Assuming the alternative hypothesis $\mathbf{H}_1$, we have

$$e_k = P_H^* Y_k = P_H^* V_k + \nu\, P_H^*\;\Upsilon\;\mathbf{1}_{\{k\geq t_0\}} \tag{7.5.3}$$

The comparison between (7.5.2) and (7.5.3), remembering that the covariance matrix of V is diagonal, leads to monitor

$$\|P_H^* Y_k\|^2 - \|P_H^*(Y_k - \nu\,\Upsilon)\|^2 \tag{7.5.4}$$

that is, the squared norm of the residual vector defined in (7.4.10).

Let us now compare this approach with the GLR technique for regression models. The generalized likelihood ratio algorithm for hypotheses testing is introduced in subsection 4.2.7. As discussed in subsection 7.2.2, when applied to the regression model (7.5.1) with unknown input variable X, this approach consists of testing between composite hypotheses and thus in monitoring the following quantity :

$$S_k \;=\; \sum_{j=1}^{k} \ln \frac{\sup_{X_j} p_{\nu\Upsilon}(Y_j|X_j)}{\sup_{X_j} p_0(Y_j|X_j)} \tag{7.5.5}$$

$$\;=\; \frac{1}{2\sigma^2}\sum_{j=t_0}^{k}\left(\inf_{X_j}\|Y_j - HX_j\|^2 - \inf_{X_j}\|Y_j - HX_j - \nu\,\Upsilon\|^2\right) \tag{7.5.6}$$

From (7.4.8), we get

$$2\sigma^2 S_k = \|P_H^* Y_k\|^2 - \|P_H^*(Y_k - \nu\,\Upsilon)\|^2 \tag{7.5.7}$$

where P_H^* is defined in (7.4.9). This results in

$$\sigma^2 S_k = \nu\,\Upsilon^T \sum_{j=t_0}^{k} P_H^*\left(Y_k - \frac{\nu}{2}\,\Upsilon\right) \tag{7.5.8}$$

It is clear from (7.5.4) and (7.5.7) that in the particular case of additive changes in regression models with equal sensor noise variances, the direct redundancy parity approach and the likelihood approach involve exactly the same projection operation, and *monitor the same sufficient statistic*. Moreover, (7.5.8) can be rewritten as

$$2\sigma^2 S_k = 2\nu \sum_{j=t_0}^{k} \rho^T e_j - \nu^2 \sum_{j=t_0}^{k} \rho^T \rho \tag{7.5.9}$$

where $\rho = P_H^* \Upsilon$ and now it is obvious that the decision function is nothing but the correlation between the projection of the observations $P_H^* Y_k$, namely the residuals e_k, and the signature $P_H^* \Upsilon$ of the assumed failure on the residual.

Finally, the likelihood ratio statistic can be written as a function of the parity check ζ in the following manner :

$$2\sigma^2 S_k = 2\nu \sum_{j=t_0}^{k} \rho^T C^T \zeta_j - \nu^2 \sum_{j=t_0}^{k} \rho^T \rho \tag{7.5.10}$$

where C satisfies (7.4.11).

Arbitrary sensor noise covariance matrix When R - the covariance matrix of V - is neither proportional to identity nor diagonal, the squared norm of ζ is not the same as the norm of e, as we show now. Let us thus investigate the situation where the sensor noise covariance matrix R is an arbitrary positive definite matrix, factorized as in (7.2.57). We show in subsection 7.2.2 that, in this case, the relevant projection matrix is not P_H^* but P_H defined in (7.2.65) and that the transformation from the observations Y to the residuals e is

$$e_k = \tilde{P}_H Y_k \tag{7.5.11}$$

where $\tilde{P}_H$, defined in (7.2.63), is not a projection (except if R is diagonal). Let C be the factorization of P_H, and define the parity check ζ by

$$\zeta_k = C Y_k \tag{7.5.12}$$

Then the norms of the residual e and the parity check are *different* :

$$\|e_k\|^2 = Y_k^T \tilde{P}_H^T \tilde{P}_H Y_k \tag{7.5.13}$$

$$\|\zeta_k\|^2 = Y_k^T P_H Y_k \tag{7.5.14}$$

Furthermore, the expression, in terms of the parity check ζ, of the log-likelihood ratio s given in (7.2.68) is less simple and requires the expression of

$$e_k = A C^T C A^{-1} Y_k \tag{7.5.15}$$

in terms of ζ.

Another question of interest is the issue of robustness of the geometrical and statistical approaches with respect to the assumption about R, namely the behavior of both decision functions built under the assumption (7.4.2) when R actually is not so. This question needs further investigation.

7.5.1.2 State-Space models

Now let us extend this result to the more complex case of additive changes in state-space models, and compare similarly the analytical redundancy and the likelihood approaches introduced in subsections 7.4.2 and 7.2.4, respectively. Again the analytical redundancy method for state-space models is described as

a one-model approach for change detection in subsection 7.4.2. Let us now investigate the corresponding two-model approach, namely the behavior of this residual under the no-change (unfailed) and change (failed) hypotheses $\mathbf{H}_0$ and $\mathbf{H}_1$ corresponding to the noisy models (7.1.3) and (7.1.6), respectively. Under $\mathbf{H}_0$, the residual (7.4.26) is equal to

$$\begin{aligned} e_k &= P_l^* \, \tilde{\mathcal{Y}}_{k-l+1}^k \\ &= P_l^* \, \tilde{\mathcal{V}}_{k-l+1}^k \end{aligned} \tag{7.5.16}$$

Using repeatedly (7.1.6), which we rewrite here,

$$\left\{ \begin{array}{rcccccccc} X_{k+1} & = & FX_k & + & GU_k & + & W_k & + & \Gamma \, \Upsilon_x(k,t_0) \\ Y_k & = & HX_k & + & JU_k & + & V_k & + & \Xi \, \Upsilon_y(k,t_0) \end{array} \right. \tag{7.5.17}$$

we extend to the failure hypothesis $\mathbf{H}_1$ the relationship (7.4.18) and get

$$\tilde{\mathcal{Y}}_{k-l+1}^k = \mathcal{O}_l \; X_{k-l+1} + \tilde{\mathcal{V}}_{k-l+1}^k + \mathcal{J}_l(\Gamma, 0) \; (\Psi_x)_{k-l+1}^k + \mathcal{J}_l(0, \Xi)(\Psi_y)_{k-l+1}^k \tag{7.5.18}$$

where $\mathcal{O}_l$ and $\mathcal{J}_l(G, J)$ are as defined before, and

$$(\Psi_x)_{k-l+1}^k = \left(\begin{array}{c} \Upsilon_x(k-l+1,t_0) \\ \Upsilon_x(k-l+2,t_0) \\ \vdots \\ \Upsilon_x(k,t_0) \end{array} \right) \text{ and } (\Psi_y)_{k-l+1}^k = \left(\begin{array}{c} \Upsilon_y(k-l+1,t_0) \\ \Upsilon_y(k-l+2,t_0) \\ \vdots \\ \Upsilon_y(k,t_0) \end{array} \right) \tag{7.5.19}$$

Under $\mathbf{H}_1$, the residual is thus

$$\begin{aligned} e_k &= P_l^* \, \tilde{\mathcal{Y}}_{k-l+1}^k \\ &= P_l^* \, \tilde{\mathcal{V}}_{k-l+1}^k + P_l^* \, \mathcal{J}_l(\Gamma, 0) \; (\Psi_x)_{k-l+1}^k + P_l^* (I_l \otimes \Xi) \; (\Psi_y)_{k-l+1}^k \end{aligned} \tag{7.5.20}$$

because of (7.5.18). Equations (7.5.16) and (7.5.20) should be compared with (7.5.2) and (7.5.3), respectively. Thus, in the case of equal noise variances (7.4.24), the analytical redundancy approach, as summarized in (7.5.16) and (7.5.20), thus leads to the comparison between the squared norms of

$$P_l^* \, \tilde{\mathcal{Y}}_{k-l+1}^k \text{ and } P_l^* \left[\tilde{\mathcal{Y}}_{k-l+1}^k - \mathcal{J}_l(\Gamma, 0) \; (\Psi_x)_{k-l+1}^k - (I_l \otimes \Xi)(\Psi_y)_{k-l+1}^k \right] \tag{7.5.21}$$

and is identical to the GLR approach as in the case of regression models.

But, as noticed before, in the case of dynamic systems, the assumption (7.4.24) is never true. The extension to arbitrary noise covariance matrices is as follows. In this case, the relevant definition of the residual e is the following :

$$e_k = \tilde{P}_l \tilde{\mathcal{Y}}_{k-l+1}^k \tag{7.5.22}$$

where

$$\tilde{P}_l = \mathcal{A}_l P_l \mathcal{A}_l^{-1} \tag{7.5.23}$$

and P_l is defined according to the general case as in (7.2.65) :

$$P_l = I - \mathcal{A}_l^{-1} \mathcal{O}_l (\mathcal{O}_l^T \mathcal{R}_l^{-1} \mathcal{O}_l)^{-1} \mathcal{O}_l^T \mathcal{A}_l^{-T} \tag{7.5.24}$$

and where the covariance factorization is

$$\mathcal{R}_l = \mathcal{A}_l \mathcal{A}_l^T \tag{7.5.25}$$

Now, the covariance matrix of e is

$$\tilde{\mathcal{R}}_l = \tilde{P}_l \mathcal{R}_l \tilde{P}_l^T \tag{7.5.26}$$

and, according to (7.2.70), the log-likelihood ratio can be written as

$$s_k = \tilde{\rho}_k^T \tilde{\mathcal{R}}_l^{-1} e_k - \frac{1}{2} \tilde{\rho}_k^T \tilde{\mathcal{R}}_l^{-1} \tilde{\rho}_k \tag{7.5.27}$$

where

$$\tilde{\rho}_k = \tilde{P}_l \left[\mathcal{J}_l(\Gamma, 0) \; (\Psi_x)_{k-l+1}^k + (I \otimes \Xi)(\Psi_y)_{k-l+1}^k \right] \tag{7.5.28}$$

Recall that, in the case of dynamic systems, the residual e_k is different from a collection of parity checks ζ_k, because the parity checks are generated with the aid of the projection matrix P_l^* and not $\tilde{P}_l$.

From our discussion in this subsection, we deduce that the obvious advantage of the statistical approach over the geometrical approach is that it automatically takes into account sensor noise and calibration issues.

7.5.2 Innovations and Generalized Parity Checks

As we mention in the introduction to this section, in the case of additive changes in state-space models, we can also investigate the *signature* of a change on the innovation and on the generalized parity checks. Here we follow this second line.

Let us first recall the two results concerning these signatures that we obtained in subsections 7.2.4 and 7.4.3. Considering the general additive changes model (7.2.97), we have shown that the signature of the changes Υ on the innovation of a Kalman filter is

$$\rho(k, t_0) = \mathcal{K}_x(z) \Upsilon_x(k, t_0) + \mathcal{K}_y(z) \Upsilon_y(k, t_0) \tag{7.5.29}$$

where z is the forward shift operator and

$$\begin{aligned} \mathcal{K}_x(z) &= H(zI_n - \bar{F})^{-1}(I_n - \bar{F}^{k-t_0} z^{-k+t_0})\Gamma \\ \mathcal{K}_y(z) &= -H(zI_n - \bar{F})^{-1}(I_n - \bar{F}^{k-t_0} z^{-k+t_0})FK\Xi + \Xi \end{aligned} \tag{7.5.30}$$

which asymptotically simplify into

$$\begin{aligned} \mathcal{K}_x(z) &= H(zI_n - \bar{F})^{-1}\Gamma \\ \mathcal{K}_y(z) &= \left[I_n - H(zI_n - \bar{F})^{-1} FK \right] \Xi \end{aligned} \tag{7.5.31}$$

On the other hand, the signature $\varrho(k, t_0)$ of the same additive changes on the generalized parity check ζ_k designed with the aid of the stable factorization approach has been shown to be

$$\varrho(k, t_0) = \mathcal{H}_x(z) \Upsilon_x(k, t_0) + \mathcal{H}_y(z) \Upsilon_y(k, t_0) \tag{7.5.32}$$

where

$$\begin{aligned} \mathcal{H}_x(z) &= \left[I_r - H(zI_n - \bar{F})^{-1} FK \right] H(zI - F)^{-1}\Gamma \\ \mathcal{H}_y(z) &= \left[I_r - H(zI_n - \bar{F})^{-1} FK \right] \Xi \end{aligned}$$

It results from identity (3.2.28) and straightforward computations that these two signatures are the same, provided that the parity check is designed with the aid of the Kalman filter gain K, which is an admissible choice.

The first consequence of this result concerns the relationship between the innovation of the Kalman filter and the generalized parity check. Actually, with obvious transfer function notations, we can write

$$\begin{aligned} \mathcal{T}_{Y\to\varepsilon}\cdot\mathcal{T}_{\Upsilon\to Y} &= \mathcal{T}_{\Upsilon\to\varepsilon} \\ \mathcal{T}_{Y\to\zeta}\cdot\mathcal{T}_{\Upsilon\to Y} &= \mathcal{T}_{\Upsilon\to\zeta} \end{aligned} \tag{7.5.33}$$

for any change vector Υ. Considering a single change Υ_y associated to a gain $\Xi = I_r$, and using the equality between the two right-hand sides of these equations result in

$$\begin{aligned} \mathcal{T}_{Y\to\varepsilon}(z) &= \mathcal{T}_{Y\to\zeta}(z) \\ &= \left[I_r - H(zI_n - \bar{F})^{-1}FK\right] \\ &= \tilde{D}(z) \end{aligned} \tag{7.5.34}$$

The equality (7.5.34) concerning the innovation is a known result of Kalman filtering. Using identity (3.2.28), this transfer function can easily be shown to be exactly the inverse of the transfer function associated with the innovation model (3.2.22), which we gave in chapter 3. Moreover, here we recover the fact that the innovation of the Kalman filter and the generalized parity check designed with the aid of a stable factorization of the input-output transfer function, operate the same compression of the information contained in the observations Y, in terms of transfer functions.

The second consequence of this result could be the following. To design a statistical decision function based upon the generalized parity checks designed with the aid of spectral factorization techniques, a possible solution consists of applying the CUSUM or GLR algorithms discussed in subsection 7.2.1. Note that the resulting algorithm is necessarily *suboptimal* because the generalized parity checks ζ_k are *not* independent.

Another consequence concerns the degenerate example (7.2.225) introduced in the discussion about statistical detectability in subsection 7.2.6. A possible statistical solution to this detection problem consists of assuming, instead of (7.2.225), the following model :

$$X_{k+1} = FX_k + GU_k + \begin{pmatrix} I_{\bar{n}} & 0 \\ 0 & \alpha I_{\underline{n}} \end{pmatrix} W_k + \begin{pmatrix} 0 \\ I_{\underline{n}} \end{pmatrix} \Upsilon_x(k, t_0) \tag{7.5.35}$$

with α small, and applying the GLR algorithm of subsection 7.2.4. The link between the signatures on the Kalman filter innovation and the generalized parity check, together with the equivalence between the various geometrical techniques, shows that if there is actually no noise - namely, if $\alpha = 0$ - on the $\underline{n}$ remaining components of the state, then the resulting processing is nothing but an observer.

7.5.3 Diagnosis

Now, we emphasize the bridge between the statistical and geometrical points of view for the diagnosis problem in state-space models again. More precisely, we show that when applying a statistical decoupling criterion for achieving failure diagnosis, the first step of the resulting algorithm is nothing but a standard geometric decoupling procedure.

Actually, we showed in subsection 7.2.5 that a relevant statistical diagnosis rule consists of computing a χ^2 test associated with some conveniently *transformed* observations (7.2.187). We also notice that this transformation is not unique, and is defined up to a multiplication by an invertible matrix. It turns out that a transformation of observations, which uses the single deterministic part of the system, is an admissible choice for the transformation required by the statistical criterion. More precisely, any full row rank matrix satisfying (7.2.185) is convenient.

The consequences of this result are interesting. First, the benefit of the statistical change detection techniques, namely the fact that they automatically take into account sensor noises and calibration problems, can

be incorporated in *existing* failure diagnosis methods without complete redesign of the part of the algorithm that concerns the deterministic part of the system. The only thing that has to be done after geometrical decoupling is to apply to the resulting residuals a Kalman filter for whitening them and to compute one of the appropriate statistical change detection algorithms described in section 7.2.1. Here we recover a result similar to that of [E.Chow and Willsky, 1984]. Second, this result also means that, in some cases, statistical decoupling can be achieved in an efficient and simple manner using a standard geometrical decoupling technique for achieving the processing which concerns the deterministic part of the system. In other words, in some cases, the transformation (7.2.187) with (7.2.177) may be simpler to compute than the transformation (7.2.175).

7.5.4 Detectability

Let us now summarize the different detectability conditions that we derived in the case of additive changes in state-space models, and show the strong existing connections between them.

According to our previous discussions, the detectability of an additive change Υ in a state-space model can be defined, in the geometrical and statistical approaches, from three different points of view :

- *intuitive definitions :* the transfer functions
  - $\mathcal{T}_{\Upsilon \to Y}$ (7.2.101) from Υ toward the observation Y
  - $\mathcal{T}_{\Upsilon \to \varepsilon}$ (7.2.112) from Υ toward the innovation ε
  - $\mathcal{T}_{\Upsilon \to \zeta}$ (7.4.43) from Υ toward the parity vector ζ

  should be full rank, or equivalently left invertible; see the discussion in [Massoumnia and Vander Velde, 1988, Massoumnia *et al*., 1989];
- *statistical definition :* the increments (7.2.217) and (7.2.221) of the Kullback divergence in the transformed problem - on the innovation - should be strictly positive;
- *geometrical definition :* the extended state-space model - which incorporates the change vector in the state - should be detectable, which results in conditions (7.4.48).

Let us rewrite together all the corresponding necessary formulas to show the basic equivalence between these different points of view.

- *intuitive definitions :* the transfer functions
  - observation :
$$\mathcal{T}_{\Upsilon \to Y} = \mathcal{T}_{\Upsilon}(z) = \left[\begin{array}{c|c} H(zI_n - F)^{-1}\Gamma & \Xi \end{array} \right] \tag{7.5.36}$$
  - innovation :
$$\begin{array}{rcl} \mathcal{T}_{\Upsilon \to \varepsilon} & = & \left[\begin{array}{cc} \mathcal{K}_x & \mathcal{K}_y \end{array} \right] \\ \mathcal{K}_x(z) & = & H(zI_n - \bar{F})^{-1}\Gamma \\ \mathcal{K}_y(z) & = & \left[I_r - H(zI_n - \bar{F})^{-1}FK \right] \Xi \end{array} \tag{7.5.37}$$
  - parity vector :
$$\begin{array}{rcl} \mathcal{T}_{\Upsilon \to \zeta} & = & \left[\begin{array}{cc} \mathcal{H}_x & \mathcal{H}_y \end{array} \right] \\ \mathcal{H}_x(z) & = & \left[I_r - H(zI_n - \bar{F})^{-1}FK \right] H(zI_n - F)^{-1}\Gamma \\ \mathcal{H}_y(z) & = & \left[I_r - H(zI_n - \bar{F})^{-1}FK \right] \Xi \end{array} \tag{7.5.38}$$

should be full rank, or equivalently left invertible;

- *statistical conditions :*

$$\begin{aligned} \tilde{\mathbf{J}}_x &= \Upsilon_x^T \Gamma^T (I_n - \bar{F})^{-T} H^T \Sigma^{-1} H (I_n - \bar{F})^{-1} \Gamma \Upsilon_x > 0 \\ \tilde{\mathbf{J}}_y &= \Upsilon_y^T \Xi^T \left[I_r - H(I_n - \bar{F})^{-1} FK \right]^T \Sigma^{-1} \left[I_r - H(I_n - \bar{F})^{-1} FK \right] \Xi \Upsilon_y > 0 \end{aligned} \tag{7.5.39}$$

- *geometrical conditions :*

$$\begin{aligned} \text{rank } \; H(I_n - \bar{F})^{-1} \Gamma &= \tilde{n} \\ \text{rank } \; \left[I_r - H(I_n - \bar{F})^{-1} FK \right] \Xi &= \tilde{r} \end{aligned} \tag{7.5.40}$$

In other words, these two matrices should be full rank.

Let us now discuss the connections between all these conditions. First, we consider the three intuitive definitions based upon transfer functions. We have shown before that, assuming that the parity vector ζ is computed with the aid of the *Kalman gain* K, the signatures of the change on the innovation and on the generalized parity vector ζ are the same, namely that $\mathcal{K}_x = \mathcal{H}_x$ and $\mathcal{K}_y = \mathcal{H}_y$. The innovation-based and parity-based intuitive definitions are thus the same. Finally, the fact that the transfer functions $\mathcal{K}_x$ and $\mathcal{K}_y$ have full rank if and only if $\mathcal{T}_\Upsilon$ has full rank results from the two identities :

$$H(I_n - \bar{F})^{-1} = \left[I_r + HF(I_n - F)^{-1} K \right]^{-1} H(I_n - F)^{-1} \tag{7.5.41}$$

$$\left[I_r - H(I_n - \bar{F})^{-1} FK \right] H = H(I_n - \bar{F})^{-1} (I_n - F) \tag{7.5.42}$$

and from the following lemma :

Lemma 7.5.1 *For all z outside the set of eigenvalues of $\bar{F}$, we have*

$$\textit{rank } \; H(zI_n - \bar{F})^{-1} \Gamma = \textit{rank } \; H(I_n - \bar{F})^{-1} \Gamma \tag{7.5.43}$$

[Kailath, 1980, Vidyasagar, 1985].

Thus, the three intuitive definitions are equivalent, under the assumption concerning the choice of the gain matrix K in the design of the parity check.

Second, let us discuss the link between this full rank transfer function condition (7.5.37) and the statistical condition (7.5.39) of positivity of the Kullback divergence increments. We use the same lemma as before, and from this lemma, we deduce that conditions (7.5.37) are true if and only if conditions (7.5.39) are satisfied for any nonzero change vectors Υ_x and Υ_y.

Third, the same lemma implies the equivalence between the full rank transfer function condition (7.5.37) and the geometrical condition (7.5.40).

We thus reach the interesting conclusion that the intuitive, statistical, and geometrical definitions of the detectability of a change are basically equivalent.

7.6 Notes and References

Section 7.2

Basic problem Cases 1, 2 and 3 of the basic problem are discussed in [Nikiforov, 1978, Nikiforov, 1980, Nikiforov, 1983]. Other cases are discussed in [Basseville and Nikiforov, 1991].

Regression models Additive changes in regression models are investigated in [Kireichikov *et al.*, 1990, Nikiforov *et al.*, 1991].

ARMA models Cases 1, 2 and 3 of additive changes in ARMA models are discussed in [Nikiforov, 1978, Nikiforov, 1980, Nikiforov, 1983].

State-space models Cases 7 and 8 are discussed in [Willsky, 1976, Willsky and Jones, 1976, Basseville and Benveniste, 1983a, Willsky, 1986]. Other more elementary statistical techniques have been used in the literature, as discussed in [Willsky, 1976, Willsky and Jones, 1976, Basseville, 1982].

Statistical decoupling The problem of decoupling or isolation has been stated as a statistical hypotheses testing problem with nuisance parameters in [Wahnon *et al.*, 1991a], following the ideas developed in [Rougée *et al.*, 1987].

Statistical detectability The statistical detectability of a change in terms of the Kullback divergence between the two models before and after change was introduced in [Tanaka, 1989] for additive changes with step profiles in state-space models.

Section 7.3

The properties of the linear CUSUM algorithm are discussed in [Nikiforov, 1980, Nikiforov, 1983]. The χ^2-CUSUM is discussed in [Nikiforov *et al.*, 1993].

Section 7.4

The use of analytical redundancy for detecting faults or changes in measurement systems was introduced in the early 1970s independently in the United States [Potter and Suman, 1977] and in Soviet Union [Britov and Mironovski, 1972]. Pioneering works include [Beard, 1971, H.Jones, 1973, Willsky *et al.*, 1975, Deckert *et al.*, 1977, Satin and Gates, 1978, Mironovski, 1979]. Developments during the 1980s are reported in [E.Chow, 1980, Ray *et al.*, 1983, Ray and Desai, 1984, E.Chow and Willsky, 1984, Lou *et al.*, 1986, Massoumnia, 1986, White and Speyer, 1987, Viswanadham *et al.*, 1987a, Viswanadham and Srichander, 1987, Viswanadham and Minto, 1988, Massoumnia *et al.*, 1989, Wünnenberg, 1990]. The research in this area is described in the survey papers [Willsky, 1976, Mironovski, 1980, Gertler, 1988, Frank, 1990, Frank, 1991, Gertler, 1991, Patton and Chen, 1991, Ray and Luck, 1991] and in the books [Viswanadham *et al.*, 1987b, Patton *et al.*, 1989].

The definition of the geometrical detectability in terms of the detectability (observability) of the extended state-space model was introduced in [Caglayan, 1980]. Another system oriented detectability definition was introduced in [Emami-Naeini *et al.*, 1988], where a failure is said to be detectable if the H_2 norm of its signature on the innovation is greater than the norm of both the noise and model uncertainty effect.

Section 7.5

The link between the statistical and geometrical approaches has been investigated for additive change detection in regression models in [Nikiforov, 1991, Nikiforov *et al.*, 1991]. In state-space models, the relationship between the projections associated with Kalman filtering on one hand and with parity checks on the other hand does not seem to have been investigated yet. The link we give between the signatures of the change on the innovation of the Kalman filter and on the generalized parity check obtained through factorization of

the input-output transfer function seems to be new. The connections between the detectability conditions resulting from the intuitive, geometrical, and statistical points of view also seem to be derived here for the first time.

7.7 Summary

Statistical Approach

Basic problem

Known θ_0 and θ_1

$$\begin{aligned} t_a &= \min\{k : g_k \geq h\} \\ g_k &= (g_{k-1} + s_k)^+ \\ s_k &= (\theta_1 - \theta_0)^T \Sigma^{-1}(Y_k - \theta_0) - \frac{1}{2}(\theta_1 - \theta_0)^T \Sigma^{-1}(\theta_1 - \theta_0) \end{aligned}$$

Known θ_0 but unknown θ_1

$$\begin{aligned} t_a &= \min\{k : g_k \geq h\} \\ g_k &= \max_{1 \leq j \leq k} \frac{k-j+1}{2}(\bar{Y}_j^k - \theta_0)^T \Sigma^{-1}(\bar{Y}_j^k - \theta_0) \end{aligned}$$

Regression models - Known Υ

$$Y_k = HX_k + V_k + \Upsilon \mathbf{1}_{\{k \geq t_0\}}$$

The decision rule is as in the corresponding case of the basic problem, with

$$s_k = \rho^T R^{-1} e_k - \frac{1}{2}\rho^T R^{-1} \rho$$

where $e_k = P_{\tilde{H}} Y_k$ and $\rho = P_{\tilde{H}} \Upsilon$.

ARMA models

$$Y_k = \sum_{i=1}^{p} A_i Y_{k-i} + \sum_{j=0}^{q} B_j V_{k-j} + \Upsilon(k, t_0)$$

Algorithms Use formulas for basic problem, replacing $Y_k - \theta_0$ by the innovation ε_k and Σ by R, and keeping $\theta_1 - \theta_0$ or Υ as they are.

State-space models

$$\left\{ \begin{array}{rcccccccc} X_{k+1} &=& FX_k &+& GU_k &+& W_k &+& \Gamma\, \Upsilon_x(k, t_0) \\ Y_k &=& HX_k &+& JU_k &+& V_k &+& \Xi\, \Upsilon_y(k, t_0) \end{array} \right.$$

where $\dim X = n$, $\dim U = m$, $\dim Y = r$, $\mathrm{cov}(W_k) = Q$, and $\mathrm{cov}(V_k) = R$, and where $\dim \Upsilon_x(k, t_0) = \tilde{n} \leq n$, $\dim \Upsilon_y(k, t_0) = \tilde{r} \leq r$, $\dim \Gamma = n \times \tilde{n}$, and $\dim \Xi = r \times \tilde{r}$.

Associated transfer functions

$$\begin{aligned} \mathcal{T}_U(z) &= H(zI_n - F)^{-1}G + J \\ \mathcal{T}_{\Upsilon \to Y}(z) = \mathcal{T}_\Upsilon(z) &= \left[\begin{array}{c:c} H(zI_n - F)^{-1}\Gamma & \Xi \end{array} \right] \end{aligned}$$

Signature of the changes on the innovation of the Kalman filter

$$\begin{aligned} \rho(k, t_0) &= \mathcal{K}_x(z)\Upsilon_x(k, t_0) + \mathcal{K}_y(z)\Upsilon_y(k, t_0) \\ \text{where } \mathcal{K}_x(z) &= H(zI_n - \bar{F})^{-1}(I_n - \bar{F}^{k-t_0}z^{-k+t_0})\Gamma \\ \mathcal{K}_y(z) &= -H(zI_n - \bar{F})^{-1}(I_n - \bar{F}^{k-t_0}z^{-k+t_0})FK\Xi + \Xi \end{aligned}$$

which asymptotically for k large simplifies into

$$\begin{aligned} \mathcal{K}_x(z) &= H(zI_n - \bar{F})^{-1}\Gamma \\ \mathcal{K}_y(z) &= \left[I_r - H(zI_n - \bar{F})^{-1}FK\right]\Xi \end{aligned}$$

Algorithms Use formulas for basic problem, replacing $Y_k - \theta_0$ by the innovation ε_k, $\theta_1 - \theta_0$ or Υ by the signature ρ, and Σ by Σ_k.

Statistical decoupling

GLR approach *Off-line problem statement :*

$$\mathbf{H}_0 : \left\{ \begin{array}{rcl} X_{k+1} &=& FX_k + GU_k + W_k + \Gamma_2\,\tilde{\Upsilon}_2(k) \\ Y_k &=& HX_k + JU_k + V_k \end{array} \right.$$

$$\mathbf{H}_1 : \left\{ \begin{array}{rcl} X_{k+1} &=& FX_k + GU_k + W_k + \Gamma_1\,\Upsilon_1(k) + \Gamma_2\,\Upsilon_2(k) \\ Y_k &=& HX_k + JU_k + V_k \end{array} \right.$$

Static problem statement :

$$\begin{aligned} Y &\sim \mathcal{N}(\bar{\mu}, \Sigma) \\ \mathbf{H}_0 : \bar{\mu} &= M_2\tilde{\mu}_2 \\ \mathbf{H}_1 : \bar{\mu} &= M\mu = M_1\mu_1 + M_2\mu_2 \end{aligned}$$

GLR solution :

$$\begin{aligned} S_N &= Y^T\left[\bar{P}_2 M_1(M_1^T\bar{P}_2M_1)^{-1}M_1^T\bar{P}_2\right]Y \\ \bar{P}_2 &= \Sigma^{-1}\left[\Sigma - M_2(M_2^T\Sigma^{-1}M_2)^{-1}M_2^T\right]\Sigma^{-1} \end{aligned}$$

Statistical detectability Criterion : strict positivity of the Kullback divergence.

Basic problem and ARMA models

$$\mathbf{J}(\theta_0, \theta_1) = (\theta_1 - \theta_0)^T\Sigma^{-1}(\theta_1 - \theta_0) > 0$$

Regression model

$$\mathbf{J}(0,\Upsilon) = \Upsilon^T A^{-T} P_H A^{-1} \Upsilon > 0$$

State-space model Positivity of the time increment of the Kullback divergence (because of the dynamic profile of the change on the innovation).

For a change in the state transition equation :

$$\tilde{\mathbf{J}}_x = \Upsilon_x^T \Gamma^T (I_n - \bar{F})^{-T} H^T \Sigma^{-1} H (I_n - \bar{F})^{-1} \Gamma \Upsilon_x > 0$$

For a change in the observation equation :

$$\tilde{\mathbf{J}}_y = \Upsilon_y^T \Xi^T \left[I_r - H(I_n - \bar{F})^{-1} FK \right]^T \Sigma^{-1} \left[I_r - H(I_n - \bar{F})^{-1} FK \right] \Xi \Upsilon_y > 0$$

Properties of the Statistical Algorithms

Linear CUSUM algorithm

Wald's approximation of the ARL function

$$\hat{L}_0(b) \quad = \quad \frac{2b\hat{L}_0^{\frac{1}{2}}(0) + e^{-2b\hat{L}_0^{\frac{1}{2}}(0)} - 1}{2b^2}$$

where

$$b = \nu \frac{\Upsilon^T \Sigma^{-1} \tilde{\Upsilon}}{(\Upsilon^T \Sigma^{-1} \Upsilon)^{\frac{1}{2}}} \neq 0$$

χ^2-CUSUM algorithm

$$\bar{\tau}^* \sim \frac{\ln \bar{T}}{\mathbf{K}(\theta_1, \theta_0)} \text{ when } \bar{T} \to \infty$$

Geometrical Approach

Direct redundancy

$$Y_k \quad = \quad HX_k \quad + \quad V_k \quad \text{where } R = \sigma^2 I_r$$

Parity vector

$$\zeta_k = CY_k$$

where $CH = 0,\ CC^T = I_{r-n},\ C^T C = P_H^*$.

Residual vector

$$e_k = Y_k - H\hat{X}_k = P_H^* Y_k = C^T \zeta_k$$

$$\|e_k\|^2 = \|\zeta_k\|^2$$

Temporal redundancy

$$\left\{ \begin{array}{rcccccc} X_{k+1} & = & FX_k & + & GU_k & + & W_k \\ Y_k & = & HX_k & + & JU_k & + & V_k \end{array} \right.$$

Dynamic model as a static regression equation

$$\tilde{\mathcal{Y}}^k_{k-l+1} \;=\; \mathcal{O}_l \; X_{k-l+1} + \tilde{\mathcal{V}}^k_{k-l+1}$$

Parity check

$$\zeta_k = v^T \left[\mathcal{Y}^k_{k-l+1} - \mathcal{J}_l(G,J) \; \mathcal{U}^k_{k-l+1} \right] = v^T \; \tilde{\mathcal{Y}}^k_{k-l+1}$$

associated with each vector v in the parity space of order l $(1 \leq l \leq n)$:

$$\mathcal{S}_l = \text{span } \{ v | v^T \mathcal{O}_l(H,F) = 0 \}$$

Generalized parity vector

$$\zeta_k = \tilde{D}(z) Y_k - \tilde{N}(z) U_k$$

where

$$\begin{array}{rcccl} \tilde{N}(z) & = & J & + & H(zI_n - \bar{F})^{-1}(G - FKJ) \\ \tilde{D}(z) & = & I_r & - & H(zI_n - \bar{F})^{-1} FK \end{array}$$

factorize the input-output transfer function $\mathcal{T}_U = \tilde{D}^{-1}\tilde{N}$.

Signature of the changes on the parity vector

$$\begin{array}{rcl} \varrho(k,t_0) & = & \mathcal{H}_x(z) \Upsilon_x(k,t_0) + \mathcal{H}_y(z) \Upsilon_y(k,t_0) \\ \text{where } \mathcal{H}_x(z) & = & \left[I_r - H(zI_n - \bar{F})^{-1} FK \right] H(zI_n - F)^{-1} \Gamma \\ \mathcal{H}_y(z) & = & \left[I_r - H(zI_n - \bar{F})^{-1} FK \right] \Xi \end{array}$$

Geometrical decoupling Solve

$$\mathcal{A} M_2 = 0$$

Geometrical detectability An additive change in a state-space model is weakly detectable if the extended state-space model - which incorporates the change vector in the state - is detectable.

Basic Geometrical/Statistical Links

Regression models The statistical change detection algorithms are based upon the residual e_k, which is used for the generation of the parity check of the geometrical method.

State-space models The signatures of the additive change, on the Kalman filter innovation and on the generalized parity check designed with the Kalman gain, are the same :

$$\begin{array}{rcl} \mathcal{H}_x(z) & = & \mathcal{K}_x(z) \\ \mathcal{H}_y(z) & = & \mathcal{K}_y(z) \end{array}$$

Diagnosis The transformation $\mathcal{A} = M_1^T \bar{P}_2$ associated with the minmax approach to statistical decoupling satisfies

$$\mathcal{A} M_2 = 0$$

The resulting χ^2-test is independent of the solution of this equation.

Detectability in a state-space model The following conditions are equivalent :

$$\begin{array}{ll} \mathcal{T}_{\Upsilon \to Y} = \mathcal{T}_{\Upsilon} & \text{left invertible} \\ \mathcal{T}_{\Upsilon \to \varepsilon} = \left[\begin{array}{cc} \mathcal{K}_x & \mathcal{K}_y \end{array} \right] & \text{left invertible} \\ \mathcal{T}_{\Upsilon \to \zeta} = \left[\begin{array}{cc} \mathcal{H}_x & \mathcal{H}_y \end{array} \right] & \text{left invertible} \end{array}$$

and are equivalent to the statistical and geometrical detectability conditions.

8

Nonadditive Changes - Scalar Signals

In this chapter, we investigate the problem of detecting *nonadditive changes in scalar signals*. Nonadditive changes are considered in the four following models :

1. conditional probability distributions;

2. AR models;

3. ARMA models;

4. nonlinear ARMA models.

As we explained in section 6.2, conditional distributions for *dependent* processes are the most general statistical model. The issue of detecting nonadditive changes (as defined in section 6.1) in this model plays a central role in this and the following chapter. Solutions to change detection problems for the two other models, AR and ARMA, are obtained as particular cases of this one. The nonlinear case is treated separately, even though its solution is based upon one of the key tools used in the other cases, namely the local approach. The particular case of nonadditive changes in *independent* sequences is also covered in this and the following chapter.

This chapter is mainly devoted to the introduction of the principal ideas for designing *on-line* nonadditive change detection algorithms. However, in section 8.7, we discuss some off-line algorithms and their connections to on-line algorithms.

The main **goals** of this chapter are as follows. First, we extend the GLR and CUSUM algorithms to the detection of nonadditive changes in the first three above-mentioned models. Starting with the GLR approach and its complexity, the second goal is to introduce simplifications that are useful in reducing this complexity. We investigate these questions in sections 8.2 and 8.3, distinguishing several levels of available *a priori* information. We first consider the case of known parameters before and after change, namely the case of simple hypotheses. Next we investigate several cases of composite hypotheses, corresponding to different levels of available *a priori* information about the parameter after change. In section 8.4, the third goal, is to introduce non-likelihood-based algorithms for solving change detection problems in nonlinear models; these can be viewed as another simplification of the GLR approach. The last goal is to clarify the detectability issue, using basically the concept of Kullback information as discussed in section 6.3; this is done in section 8.5.

The **tools** for reaching these goals can be summarized as follows. As we said before, we first put everything in the framework of conditional densities. Next the key issue is that *the transformation from observations to innovations used for additive changes is not sufficient for detecting nonadditive or spectral changes*, as we explain in section 8.1. Then we introduce the local approach for change detection, which has not been discussed yet in this book. The *efficient score* is shown to be both the sufficient statistic for small nonadditive changes and asymptotically Gaussian distributed. Thus, the transformation from

observations to efficient scores transforms the nonadditive change detection problem into the problem of detecting changes in the mean value of a vector Gaussian process, namely into the basic problem of chapter 7. This is explained in subsection 8.1.3, and investigated in sections 8.2, 8.3 and 8.4 for each of the four above-mentioned models. Furthermore, we also introduce non-likelihood-based detection algorithms, as we discussed in section 6.3. This is done for a large class of conditional probabilities in section 8.4. Then in section 8.6 we discuss the problem of the *implementation with unknown model parameters of algorithms that are designed with known model parameters*, namely the problems of choice of identification algorithms when implementing, in real situations where the models are unknown, algorithms designed with the aid of known models. We discuss what we call one- and two-model approaches, according to the number of models whose parameters must be estimated to compute the decision function.

8.1 Introducing the Tools

In this section, we introduce nonadditive changes in the four types of models. In subsection 8.1.2 we also introduce the key concepts that are to be used for solving the corresponding detection problem, namely sufficient statistics and the local approach. These concepts are also useful when discussing the detectability issue in section 8.5. In subsection 8.1.3, we discuss in detail the use of the local approach for change detection. As we noted in section 6.3, the local approach does not provide any new information in the case of additive changes; thus, the use of the local approach for change detection appears in this book for the first time in this chapter. This new approach will be used in *both* chapters 8 and 9, and for both the design of the algorithms and the investigation of their properties.

8.1.1 Nonadditive Changes

We consider sequences of *scalar* observations $(y_k)_k$ (with dimension $r = 1$). In this chapter and in chapter 9, we investigate nonadditive or spectral changes. These are changes in the variance, correlations, spectral characteristics, or dynamics of the signal or system.

We first describe the four parametric models. Then we describe three methods of generating changes in a parameterized conditional density.

8.1.1.1 The Four Models

We consider the four following models :

- **Conditional distribution** : We assume that there exists a parameterized conditional probability density $p_\theta(y_k|\mathcal{Y}_1^{k-1})$ which serves as a model for the observed signal. The problem is to detect changes in the vector parameter θ.
- **AR models** :

$$y_k = \sum_{i=1}^{p} a_i y_{k-i} + v_k \tag{8.1.1}$$

  where $(v_k)_k$ is a Gaussian white noise sequence with variance $R = \sigma^2$. The conditional probability distribution of such a sequence of observations $(y_k)_k$ is denoted by $p_\theta(y_k|\mathcal{Y}_1^{k-1})$, where θ is the vector containing the AR coefficients and the standard deviation σ. The problem is to detect changes in the vector parameter θ, from θ_0 to θ_1, where

$$\theta_l^T = \begin{pmatrix} a_1^l & \dots & a_p^l & \sigma_l \end{pmatrix}, \quad l = 0, 1 \tag{8.1.2}$$

- **ARMA models** :

$$y_k = \sum_{i=1}^{p} a_i y_{k-i} + \sum_{j=0}^{q} b_j v_{k-j} \tag{8.1.3}$$

where $(v_k)_k$ is again a Gaussian white noise sequence with variance $R = \sigma^2$, and $b_0 = 1$. The conditional probability distribution of such a sequence of observations $(y_k)_k$ is denoted by $p_\theta(y_k|\mathcal{Y}_1^{k-1})$, where θ is the vector containing the AR and MA coefficients and the standard deviation σ. The problem is to detect changes in the vector parameter θ, from θ_0 to θ_1, where

$$\theta_l^T = \begin{pmatrix} a_1^l & \dots & a_p^l & b_1^l & \dots & b_q^l & \sigma_l \end{pmatrix}, \quad l = 0, 1 \tag{8.1.4}$$

Changes in these two models are of interest in several types of signals, such as continuous speech signals, seismic data, and biomedical signals.

- **Nonlinear ARMA models** : We refer to processes $(Y_k)_k$, which can be modeled with the aid of the following Markov representation with finite state-space :

$$\begin{cases} \mathbf{P}(X_k \in B|X_{k-1}, X_{k-2}, \dots) &= \int_B \pi_\theta(X_{k-1}, dx) \\ y_k &= f(X_k) \end{cases} \tag{8.1.5}$$

where $\pi_\theta(X, dx)$ is the transition probability of the Markov process $(X_k)_k$ and where f is a nonlinear function. An AR(p) process can be written in the form (8.1.5) with a linear f, using the Markov chain :

$$\begin{aligned} X_k &= \check{\mathcal{Y}}_{k-p}^{k-1} \\ &= \begin{pmatrix} y_{k-1} & y_{k-2} & \dots & y_{k-p} \end{pmatrix}^T \end{aligned} \tag{8.1.6}$$

The problem is to detect changes in the parameter θ of the transition probability π_θ. This problem statement and the corresponding solution presented in section 8.4 are of interest for solving the vibration monitoring problem presented in example 1.2.5 of chapter 1.

Note that we discuss in section 8.3 how a change detection problem in an input-output dynamic model can be solved in the framework of the general case of conditional density.

8.1.1.2 Three Methods of Generating Changes

We consider parameterized families of conditional probability distributions. From the statistical point of view, this model is the most general. In this subsection, we outline three possible ways of modeling both additive and nonadditive changes in such models.

To be able to use the general scheme depicted in figure 6.4, it would be useful to characterize all the conditional densities $p_\theta(Y_k|\mathcal{Y}_1^{k-1})$, which can be written under the form of an innovation model :

$$Y_k = \mathcal{T}_\theta(\mathcal{V}_{k-\infty}^{k-1}) \tag{8.1.7}$$

and the condition under which this function $\mathcal{T}_\theta$ is invertible. The only available results on this issue seem to be the following. This innovation model obviously exists for all linear processes for which Wold's decomposition exists [Shiryaev, 1984].

Let $(Y_k)_{k\geq 1}$ be an observed random process. Assume that $p_{\theta_0}(Y_k|\mathcal{Y}_1^{k-1})$ is the conditional density before change and that $p_{\theta_1}(Y_k|\mathcal{Y}_1^{k-1})$ is the conditional density after change. Such a process can be generated in at least three ways, which we describe now.

The first method (figure 8.1) This method consists of a single "generator" of observations with conditional density $p_\theta(Y_k|\mathcal{Y}_1^{k-1})$. The parameter vector θ_0 of this "generator" is replaced with a new value θ_1 at an unknown time t_0. The "memory" $\mathcal{Y}_1^{t_0-1}$ of this generator is used as the initial condition for the observations after the change. Roughly speaking, this change results in a smooth transition between the observed behavior before and after the change. The conditional density of the set of such observations $\mathcal{Y}_1^k$ under the assumption that $k \geq t_0$ is as follows :

$$p(\mathcal{Y}_1^k|k \geq t_0) = p_{\theta_0}(Y_1) \left[\prod_{i=2}^{t_0-1} p_{\theta_0}(Y_i|\mathcal{Y}_1^{i-1})\right] \left[\prod_{i=t_0}^{k} p_{\theta_1}(Y_i|\mathcal{Y}_1^{i-1})\right] \tag{8.1.8}$$

Note that this method corresponds to the three rows of figure 6.4.

The second method (figure 8.2) This approach consists of two generators with parameters θ_0 and θ_1. These generators produce two stationary processes $(W_k)_k$ and $(V_k)_k$ with conditional densities $p_{\theta_0}(W_k|\mathcal{W}_1^{k-1})$ and $p_{\theta_1}(V_k|\mathcal{V}_1^{k-1})$, which are independent from each other. In this case, the change consists of the following :

$$Y_k = \begin{cases} W_k & \text{if } k < t_0 \\ V_k & \text{if } k \geq t_0 \end{cases} \tag{8.1.9}$$

The "memory" of observations is not kept after change. In contrast to the previous method, we have not a smooth transition, but an abrupt change between the observed behavior before and after the change. Using the stationarity and the fact that V_k is completely unknown before t_0, the conditional density of the set of such observations $\mathcal{Y}_1^k$ under the assumption that $k \geq t_0$ is as follows :

$$p(\mathcal{Y}_1^k|k \geq t_0) = p_{\theta_0}(Y_1) \left[\prod_{i=2}^{t_0-1} p_{\theta_0}(Y_i|\mathcal{Y}_1^{i-1})\right] \; p_{\theta_1}(Y_{t_0}) \left[\prod_{i=t_0+1}^{k} p_{\theta_1}(Y_i|\mathcal{Y}_{t_0}^{i-1})\right] \tag{8.1.10}$$

and is thus different from that of the previous method.

The third method (figure 8.3) The third method is based upon a standard point of view for the detection of a signal in noise with unknown arrival time. It consists of two generators with parameters θ_0 and θ_1 which produce two processes $(W_k)_k$ and $(V_k)_k$ as before. In contrast to the two previous methods, before the change time t_0 the observations contain only the noise W. From time t_0, the observations contain the sum of the noise W and the signal V. From the point of view of memory, this method is in some sense the superposition of the two previous ones.

In this chapter, we use the *first* and *second* methods of generating changes, most of the time we use the first one, but we discuss, for some particular models, the effect of the first and second methods on the design of the detection algorithm.

8.1.2 Sufficient Statistics

Let us now introduce the key concepts that we use for solving these nonadditive change detection problems, namely sufficient statistics and the local approach. In this subsection we discuss only the issue of sufficient statistics. In subsection 8.1.3 we discuss in detail the local approach to change detection.

It results from section 4.1 that a sufficient statistic is a particular function of the raw observations which concentrates all of the available information about the unknown parameter θ of a parametric family $\mathcal{P} = \{\mathbf{P}_\theta\}$. In subsequent discussions, we consider the two simple hypotheses $\mathbf{H}_0 : \{\theta = \theta_0\}$ and $\mathbf{H}_1 : \{\theta = \theta_1\}$.

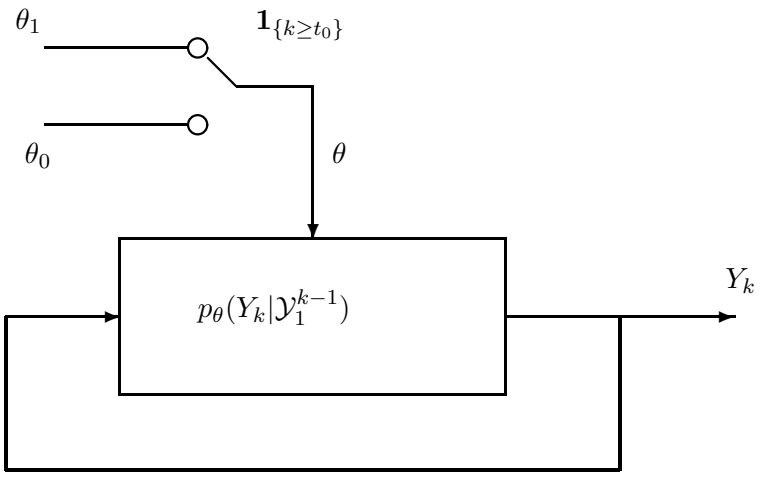

Figure 8.1 First method of generating data with changes.

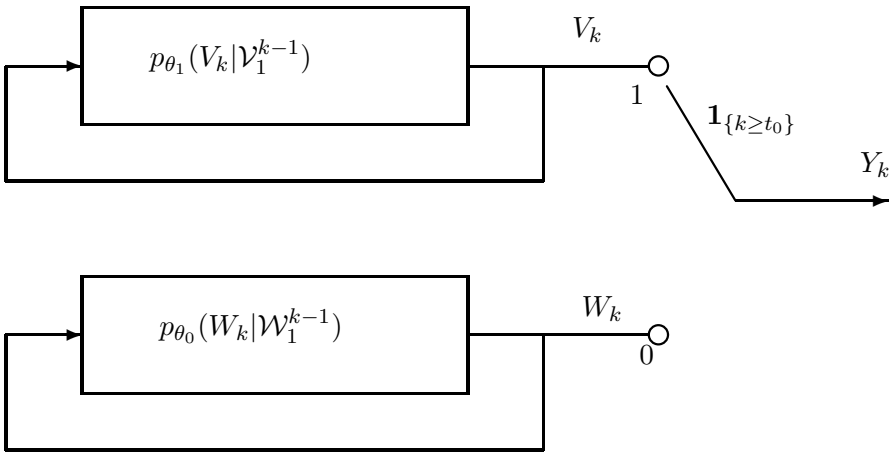

Figure 8.2 Second method of generating data with changes.

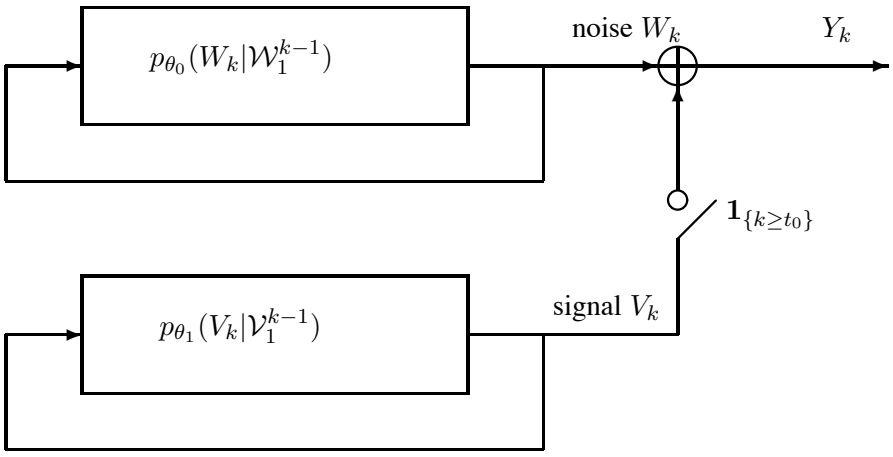

Figure 8.3 Third method of generating data with changes.

8.1.2.1 Insufficiency of the Innovation

In the case of additive changes, which are basically changes in the mean value of the observed signal, or equivalently in the mean value of the conditional probability distribution, we showed in chapter 7 that changes are left unchanged by the transformation from observations to innovations. Nonadditive changes are more complex, in the sense that *the intuitively obvious idea of monitoring deviations either from zero–mean or from whiteness in the sequence of innovations is not convenient for solving a nonadditive change detection problem*. A nonnegligible set of nonadditive changes results in absolutely no change in either the mean value or the variance of the innovation process. This can be seen from the likelihood ratio, as we discuss now.

8.1.2.2 Likelihood Ratio

In chapter 4, the log-likelihood ratio was shown to be a sufficient statistic. Let us write this function for the observations $\mathcal{Y}_1^k$:

$$S_k = \sum_{i=1}^{k} s_i \tag{8.1.11}$$

where

$$s_i = \ln \frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})} \tag{8.1.12}$$

and where $p_\theta(y_1|\mathcal{Y}_1^0) = p_\theta(y_1)$.

Example 8.1.1 (ARMA case). *In the case of ARMA models, we use the following notations*

$$A^T = \begin{pmatrix} 1 & -a_1 & \dots & -a_p \end{pmatrix} \tag{8.1.13}$$

$$B^T = \begin{pmatrix} b_1 & \dots & b_q \end{pmatrix} \tag{8.1.14}$$

for the sets of AR and MA parameters, and

$$(\check{\mathcal{Y}}_{k-p}^k)^T = \begin{pmatrix} y_k & y_{k-1} & \dots & y_{k-p} \end{pmatrix} \tag{8.1.15}$$

$$(\check{\mathcal{E}}_{k-q}^{k-1})^T = \begin{pmatrix} \varepsilon_{k-1} & \varepsilon_{k-2} & \dots & \varepsilon_{k-q} \end{pmatrix} \tag{8.1.16}$$

for the sets of past observations and innovations in backward order. The conditional probability density of the observation y_k is given by

$$\begin{aligned} p_\theta(y_k|\mathcal{Y}_1^{k-1}) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(A^T\check{\mathcal{Y}}_{k-p}^k - B^T\check{\mathcal{E}}_{k-q}^{k-1})^2} \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\varepsilon_k^2} \end{aligned} \tag{8.1.17}$$

Thus, the log-likelihood ratio increment is

$$s_k = \frac{1}{2}\ln\frac{\sigma_0^2}{\sigma_1^2} + \frac{(\varepsilon_k^0)^2}{2\sigma_0^2} - \frac{(\varepsilon_k^1)^2}{2\sigma_1^2} \tag{8.1.18}$$

which reduces to

$$s_k = \frac{(\varepsilon_k^0)^2 - (\varepsilon_k^1)^2}{2\sigma^2} \tag{8.1.19}$$

when the input variance does not change.

It results from this formula that the likelihood ratio is a function of the residuals of two whitening filters, *and not only one, as in the case of additive changes in chapter 7. This fact has strong consequences both on the design of the algorithms and on the investigation of their properties, as we explain later.*

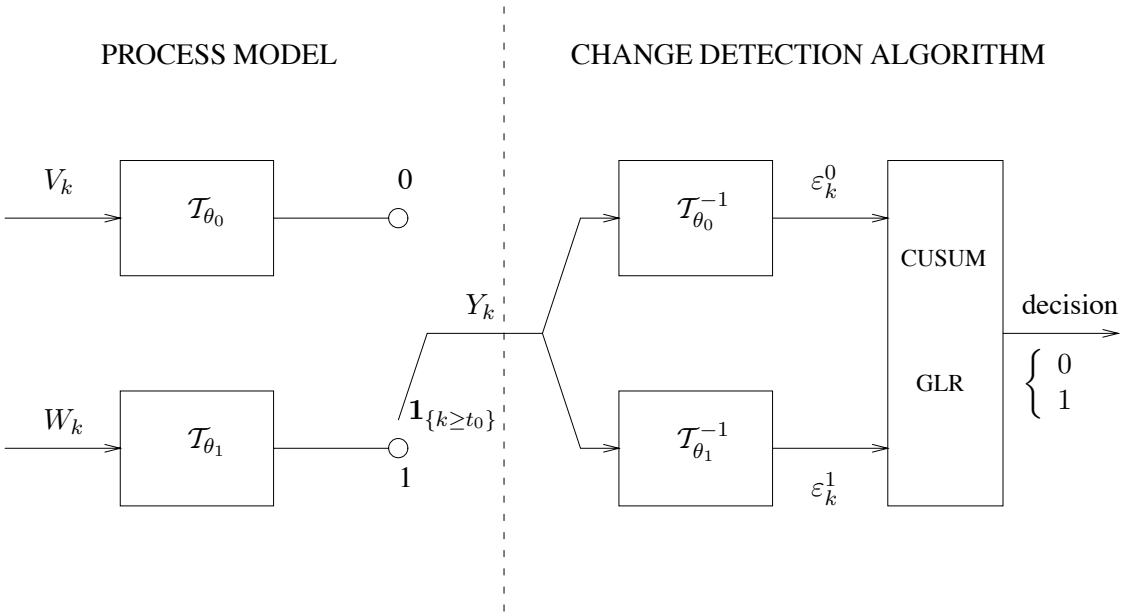

Figure 8.4 Nonadditive change detection using conditional density.

8.1.2.3 Decision Function

In the case of nonadditive changes, the key new feature, with respect to additive changes, lies in the *dependency in the sequence of log-likelihood ratio increments* $(s_k)_k$ in the CUSUM and GLR algorithms. The reason for this dependency is obvious from figure 8.4. Recall that, as we discussed in section 6.1, this picture is only a convenient guideline for the AR and ARMA cases, because the inverse filtering operations should be thought of only as a relevant way of computing the log-likelihood function in these cases. Both CUSUM and GLR algorithms use the two inverse filtering operations corresponding to the parameters before and after change. Before change, the signal is generated with the aid of the filter $\mathcal{T}_{\theta_0}(z)$, and thus the inverse filter $\mathcal{T}_{\theta_1}^{-1}(z)$ results in a dependent process ε. The same is true after change : The inverse filter $\mathcal{T}_{\theta_0}^{-1}(z)$ results in a dependent process when applied to the signal generated with the aid of the filter $\mathcal{T}_{\theta_1}(z)$. This dependency results in a more complex situation, for example, for the analysis of the theoretical properties of the CUSUM and GLR algorithms.

Another new feature of likelihood-based decision functions in this chapter is the complexity of the log-likelihood ratio in complex models. We investigate two ways for reducing this complexity. The first is the local approach, which is discussed in detail in subsection 8.1.3. The second uses statistics that are simpler than the likelihood function but nevertheless efficient from the statistical inference point of view. This results in what we call non-likelihood-based algorithms, and is discussed in section 8.4.

Now let us outline one important practical issue. In subsections 8.2.2 and 8.3.1, we derive the decision rules under the unrealistic assumption of known models before and after change. The resulting tests can nevertheless provide us with relevant algorithms in practice, simply by replacing the values of all the model parameters by their estimates. The corresponding implementation issues are discussed in section 8.6, where we introduce what we call one- and two-model approaches. The one-model approach results in an on-line estimation of the model parameters before change - using, for example, a growing or a sliding window - and testing deviations from this reference model, possibly using some *a priori* information about the

type of change. Algorithms for this purpose are described in subsections 8.2.3 and 8.3.2. In the case of unknown models before and after change, the two-model approach consists of estimating the two sets of model parameters in two different windows of data - using, for example, a growing and a sliding window, or two sliding windows of different sizes - and using one of the algorithms that we describe in subsections 8.2.2 and 8.3.1.

8.1.3 Local Approach to Change Detection

We investigate in detail the use of the asymptotic local approach for designing nonadditive change detection algorithms. The local approach was introduced in subsections 4.2.3 and 4.2.9. Here we follow the main lines of these developments. First, we take an off-line point of view of fixed size sample, and then explain how to transpose these ideas to on-line or sequential detection. We follow [Nikiforov, 1978, Nikiforov, 1980, Benveniste *et al*., 1987, Zhang *et al*., 1994].

8.1.3.1 Local Expansion and Efficient Score

We consider a parametric family of distributions $\mathcal{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta\subset\mathbf{R}^\ell}$, and we assume the two following hypotheses :

$$\begin{aligned} \mathbf{H}_0 &= \{\theta = \theta_0\} \\ \mathbf{H}_1 &= \left\{\theta = \theta_N = \theta_0 + \frac{\nu}{\sqrt{N}}\Upsilon\right\} \end{aligned} \tag{8.1.20}$$

The log-likelihood ratio for a sample of size N is

$$S_1^N(\theta_0, \theta_N) = \ln \frac{p_{\theta_N}(\mathcal{Y}_1^N)}{p_{\theta_0}(\mathcal{Y}_1^N)} \tag{8.1.21}$$

When N goes to infinity, we assume that S can be written as

$$S_1^N(\theta_0, \theta_N) \approx \nu\Upsilon^T\Delta_N(\theta_0) - \frac{\nu^2}{2}\Upsilon^T\ \mathbf{I}_N(\theta_0)\ \Upsilon \tag{8.1.22}$$

where Δ_N is related to the efficient score $\mathcal{Z}_N$ (4.1.19) for the random process $(y_k)_{1\le k\le N}$:

$$\Delta_N(\theta_0) = \frac{1}{\sqrt{N}}\mathcal{Z}_N(\theta_0) = \frac{1}{\sqrt{N}}\ \frac{\partial \ln p_\theta(\mathcal{Y}_1^N)}{\partial\theta}\bigg|_{\theta=\theta_0} \tag{8.1.23}$$

and $\mathbf{I}_N$ is its covariance or, equivalently, the Fisher information matrix. Conditions under which (8.1.22) is true were discussed in subsection 4.2.3. In other words, we consider a locally asymptotic normal (LAN) family of distributions. As we explained in subsection 4.2.9, the asymptotically optimal test for local hypotheses is based upon the efficient score, which is the relevant sufficient statistic in this case.

These ideas are used in the present chapter for designing *on-line* change detection algorithms. In chapter 9, they are used for solving complex change detection problems both on-line and *off-line*, in a fixed size sample framework, as we discussed for Shewhart charts in chapter 2. This is the case of the vibration monitoring problem introduced in example 1.2.5.

Example 8.1.2 (AR model). *Let us consider the case of an AR model (8.1.1), with vector parameter θ defined in (8.1.2). We showed in (4.1.101) that the efficient score at $\theta = \theta_0$ is*

$$\mathcal{Z}_N(\theta_0) = \frac{\partial \ln p_\theta(\mathcal{Y}_1^N)}{\partial\theta}\bigg|_{\theta=\theta_0} = \begin{pmatrix} \frac{1}{\sigma_0^2}\sum_{i=1}^N \check{\mathcal{Y}}_{i-p}^{i-1}\varepsilon_i^0 \\ \\ \frac{1}{\sigma_0}\sum_{i=1}^N \left[\frac{(\varepsilon_i^0)^2}{\sigma_0^2} - 1\right] \end{pmatrix} \tag{8.1.24}$$

and its covariance matrix, or equivalently the Fisher information matrix, is (4.1.102)

$$\mathbf{I}(\theta_0) = \begin{pmatrix} \frac{1}{\sigma_0^2}\mathbf{T}_p(\theta_0) & 0 \\ 0 & \frac{2}{\sigma_0^2} \end{pmatrix} \tag{8.1.25}$$

where $\mathbf{T}_p(\theta_0)$ is the Toeplitz matrix of order p. As we explain later, the inverse *of the Fisher information matrix plays a key role in the design of the local decision functions for spectral changes. The following formula is thus useful :*

$$\mathbf{I}^{-1}(\theta) = \begin{pmatrix} \sigma^2\mathbf{T}_p^{-1}(\theta) & 0 \\ 0 & \frac{\sigma^2}{2} \end{pmatrix} \tag{8.1.26}$$

where the inverse of the Toeplitz matrix is computed with the aid of the Göhberg-Semmencul formula :

$$\mathbf{T}_p^{-1}(\theta) = \mathbf{T}_p^{-1}(A) = T_1 T_1^T - T_2 T_2^T \tag{8.1.27}$$

where A is the vector of AR parameters (8.1.13) and

$$T_1(A) = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ -a_1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -a_{p-1} & -a_{p-2} & \dots & -a_1 & 1 \end{pmatrix} \tag{8.1.28}$$

$$T_2(A) = \begin{pmatrix} -a_p & 0 & \dots & 0 & 0 \\ -a_{p-1} & -a_p & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -a_1 & -a_2 & \dots & -a_{p-1} & -a_p \end{pmatrix} \tag{8.1.29}$$

8.1.3.2 Using Local Expansion for Change Detection

Now we explain the main idea underlying the design of *on-line* change detection algorithms using the local expansion. Let us first assume that $\theta \in \mathbf{R}$ is a *scalar parameter*, and consider two hypotheses $\{\theta = \theta_0\}$ and $\{\theta = \theta_1\}$. As we explained when introducing chapter 2, the main idea underlying the design of change detection algorithms lies in the following property :

$$\begin{array}{rcl} \mathbf{E}_{\theta_0}(s_i) & < & 0 \\ \mathbf{E}_{\theta_1}(s_i) & > & 0 \end{array} \tag{8.1.30}$$

where s_i is the log-likelihood ratio for the observation y_i. Similarly, let us investigate the mean value of the efficient score under both hypotheses. We assume that

$$\begin{array}{rcl} \theta_0 & = & \theta^* - \dfrac{\nu}{2} \\ \theta_1 & = & \theta^* + \dfrac{\nu}{2} \end{array} \tag{8.1.31}$$

where ν is a small positive number. Let z_i^* be the contribution of the observation y_i to the efficient score computed at θ^* :

$$z_i(\theta^*) = z_i^* = \left.\frac{\partial \ln p_\theta(y_i|\mathcal{Y}_1^{i-1})}{\partial\theta}\right|_{\theta=\theta^*} \tag{8.1.32}$$

We first show that, up to second-order terms, the log-likelihood ratio is equivalent to the efficient score. Actually, according to the asymptotic expansion (8.1.22), we have

$$S_1^N\left(\theta^*, \theta^* + \frac{\nu}{2}\right) \approx \frac{\nu}{2} \left.\frac{\partial \ln p_\theta(\mathcal{Y}_1^N)}{\partial \theta}\right|_{\theta=\theta^*} - \frac{\nu^2}{8}\, \mathbf{I}_N(\theta^*) \tag{8.1.33}$$

Therefore, for the above-mentioned hypotheses, we can write the log-likelihood ratio as

$$\begin{aligned} S_1^N\left(\theta^* - \frac{\nu}{2}, \theta^* + \frac{\nu}{2}\right) &\approx \nu \left.\frac{\partial \ln p_\theta(\mathcal{Y}_1^N)}{\partial \theta}\right|_{\theta=\theta^*} \\ &\approx \nu \sum_{i=1}^{N} z_i^* \end{aligned} \tag{8.1.34}$$

From this, it is obvious that, for small changes, the efficient score has approximately the same property as the likelihood ratio, namely

$$\begin{aligned} \mathbf{E}_{\theta_0}(z_i^*) &< 0 \\ \mathbf{E}_{\theta_1}(z_i^*) &> 0 \end{aligned} \tag{8.1.35}$$

In other words, *the change in the parameter θ is reflected as a change in the sign of the expectation of the efficient score*. And we use the efficient score in exactly the same way as we use the log-likelihood ratio.

Now, let us return to the case of a *vector parameter* $\theta \in \mathbf{R}^\ell$, and consider the two simple hypotheses characterized by

$$\begin{aligned} \theta_0 &= \theta^* - \frac{\nu}{2}\Upsilon \\ \theta_1 &= \theta^* + \frac{\nu}{2}\Upsilon \end{aligned} \tag{8.1.36}$$

where Υ is the *unit* vector of the change direction. The generalization of the previous discussion is straightforward, and the log-likelihood ratio in this case is

$$\begin{aligned} S_1^N\left(\theta^* - \frac{\nu}{2}\Upsilon, \theta^* + \frac{\nu}{2}\Upsilon\right) &\approx \nu\, \Upsilon^T \left.\frac{\partial \ln p_\theta(\mathcal{Y}_1^N)}{\partial \theta}\right|_{\theta=\theta^*} \\ &\approx \nu\, \Upsilon^T \sum_{i=1}^{N} Z_i^* \end{aligned} \tag{8.1.37}$$

where Z_i^* is the vector of efficient score. Therefore, *the change in θ is reflected as a change in the sign of the scalar product* $\Upsilon^T \sum_{i=1}^N Z_i^*$.

In fact, for a rigorous mathematical treatment, ν should rather be $\frac{1}{\sqrt{N}}\,\nu$. The interested reader can find precise mathematical investigation of LAN properties in subsections 4.2.3 and 4.2.9.

8.1.3.3 Asymptotic Normality of the Efficient Score

Up to now, as in chapter 4, we have discussed local asymptotic expansion of the LR for fixed sample size. All the results of subsection 4.2.9 for local hypotheses testing can be used for the design of change detection algorithms, provided that these algorithms use samples of data with large *fixed* size. To use the local approach in algorithms working with samples of data having *random* size, such as CUSUM-type and GLR algorithms, it is necessary to know the asymptotic behavior of the cumulative sum of efficient scores in (8.1.34). Let us now show that this asymptotic behavior is Gaussian. More precisely, we show that the

cumulative sum of efficient scores converges to a Brownian motion, and thus the change in the parameter θ is reflected as a change in the drift of this Brownian motion.

For this purpose, let us consider the continuous time interval $[0,1]$ and let $(W_t)_{t\in[0,1]}$ be a normalized ℓ-dimensional Brownian motion. For $t \in [0,1]$, we introduce the cumulative sum :

$$S_{N,t}(\theta^*) = \frac{1}{\sqrt{N}} \sum_{i=1}^{[Nt]} Z_i^* \tag{8.1.38}$$

where $[Nt]$ is the integer part of Nt. The following central limit theorem can be proven [P.Hall and Heyde, 1980, Picard, 1985, Benveniste *et al.*, 1990]. When $N \to \infty$

$$\begin{array}{lll} \text{under} \quad p_{\theta^*} \ : & \mathbf{I}_N^{-\frac{1}{2}}(\theta^*) \ S_{N,t}(\theta^*) & \rightsquigarrow \quad (W_t)_{t\in[0,1]} \\ \text{under} \quad p_{\theta^*+\frac{\nu}{\sqrt{N}}\Upsilon} \ : & \mathbf{I}_N^{-\frac{1}{2}}(\theta^*) \ (S_{N,t}(\theta^*) - \nu \ \mathbf{I}_N(\theta^*) \ \Upsilon \ t) & \rightsquigarrow \quad (W_t)_{t\in[0,1]} \end{array} \tag{8.1.39}$$

The key reasons that the mean value is $\nu \ \mathbf{I} \ \Upsilon \ t$ and that the variance is equal to $\mathbf{I}$ were explained in subsection 4.1.2. It should be clear that, using this asymptotic normality, the resulting test statistic involves the scalar product $\Upsilon \sum_i Z_i^*$.

Note that here we consider the situation where the model parameter is equal to the reference value θ^* before change and to $\theta^* + \nu \ \Upsilon$ after change, which is not the same as in (8.1.31). We discuss this distinction in more detail next. Remember also that for stationary processes the Fisher information matrix $\mathbf{I}_N$ is constant and equal to $\mathbf{I}$.

8.1.3.4 Design of the Algorithms

We now explain how this framework can be used to design local change detection algorithms. For reasons that become clear later, we find it useful to distinguish the following situations of alternative simple and composite hypotheses.

- **Local linear hypotheses** : More precisely, the parameter is assumed to be

$$\theta(k) = \begin{cases} \theta_0 = \theta^* - \frac{\nu}{2} \ \Upsilon & \text{when} \quad k < t_0 \\ \theta_1 = \theta^* + \frac{\nu}{2} \ \Upsilon & \text{when} \quad k \geq t_0 \end{cases} \tag{8.1.40}$$

  where $\nu > 0$ is again a *small* positive number and Υ is again the unit vector of the change direction. In this case, it results from the previous discussion that the relevant algorithm is the CUSUM and that the design of the CUSUM algorithm is based upon *only* the assumption that the likelihood ratio can be replaced by the first term (8.1.37) of the expansion (8.1.22). Note that neither the second term of the expansion nor the asymptotic normality of the cumulative sum of efficient scores is necessary for the design of the CUSUM algorithm here.

- **Local quadratic hypothesis with fixed Kullback information** : We assume the following hypotheses before and after the change :

$$\theta(k) = \begin{cases} \theta_0 & \text{when} \quad k < t_0 \\ \theta \ : (\theta - \theta_0)^T \ \mathbf{I}(\theta_0) \ (\theta - \theta_0) = b^2 & \text{when} \quad k \geq t_0 \end{cases} \tag{8.1.41}$$

  where $b > 0$ is a known *small* positive number. In this case, as we show next, there are two possible approaches for designing change detection algorithms, namely the GLR and the (invariant) χ^2-CUSUM introduced in subsection 7.2.1, and it is necessary to use the *second* term of the expansion (8.1.22). The GLR algorithm is designed using *only* this second-order expansion. The design of the χ^2-CUSUM is also based upon this second-order expansion, but also uses the asymptotic normality.

- **Local composite hypothesis** : Here we assume that

$$\theta(k) = \begin{cases} \theta_0 & \text{when} \quad k < t_0 \\ \theta \ : \|\theta - \theta_0\| = \nu & \text{when} \quad k \geq t_0 \end{cases} \tag{8.1.42}$$

  where $\nu > 0$ is again *small* but unknown. In this case, the relevant algorithm is the GLR.

Note that it may be possible to consider other situations for the second hypothesis, as in subsection 7.2.1. Nevertheless, the main ideas for the design of local nonadditive change detection algorithms can be emphasized with only these two cases.

8.1.3.5 Properties of the Algorithms

The investigation of the properties of nonadditive change detection algorithms is much more difficult than in the case of additive changes. In the case of multiplicative changes, we concentrate on CUSUM-type algorithms as far as the computation of the ARL function is concerned. As we discussed in chapter 4 and in subsection 6.3.2, there exist two possible methods of using the local approach for computing the ARL function of the change detection algorithms. The first can be used when the algorithm is designed with the aid of the heuristic use of the efficient score as the increment of the decision function of the CUSUM algorithm *and* when the increment sequence is i.i.d. Then, the ARL function can be computed by using the solution of Fredholm integral equation, or the Taylor expansion of the moment generating function (mgf) of this increment, as we explained in chapter 5. The second method consists of using the asymptotic normality of the cumulative sum of efficient scores and investigating the properties of the corresponding algorithm for detecting a change in the drift of a Brownian motion. The first method is discussed in detail in chapter 5; we now describe the main features of the second method.

It results from the *asymptotic normality* summarized in (8.1.39) that the change in the vector parameter θ of the conditional probability density $p_\theta(y_k|\mathcal{Y}_1^{k-1})$ is transformed into a change in the drift parameter of a normalized Brownian motion. In some cases, as explained before, it is of interest to use this transformation of the initial problem for the design of the change detection algorithms. The main interest of the Brownian motion as limit of cumulative sum of efficient scores is related to the fact that, for this process, there exist results about exit times and computations of ARL functions, as we described in chapters 4 and 5. The Brownian motion approximation of the cumulative sum of observations has been used for investigating the properties of CUSUM change detection algorithms in [Reynolds, 1975, R.Johnson and Bagshaw, 1974, Bagshaw and R.Johnson, 1975a, Bagshaw and R.Johnson, 1977].

Let us discuss the two change detection problems for a normalized Brownian motion corresponding to the three above-mentioned cases. Here we do *not* distinguish between the local quadratic and local composite hypotheses. Let $t_0 \in (0, 1)$ be the change time.

- Local linear hypotheses : In this case, we assume that the Brownian motion has drift $-\frac{\nu}{2}\ \mathbf{I}(\theta^*)\ \Upsilon$ before the time t_0, and drift $+\frac{\nu}{2}\ \mathbf{I}(\theta^*)\ \Upsilon$ after t_0. In other words, we consider the following Brownian motion with time-varying drift :

$$d\underline{W}_t = -\mathbf{1}_{\{t<t_0\}}\ \frac{\nu}{2}\ \mathbf{I}(\theta^*)\ \Upsilon\ dt + \mathbf{1}_{\{t\geq t_0\}}\ \frac{\nu}{2}\ \mathbf{I}(\theta^*)\ \Upsilon\ dt + \mathbf{I}^{\frac{1}{2}}(\theta^*)\ dW_t \tag{8.1.43}$$

  The discrete time counterpart of this model of change is the basic problem discussed in chapter 7, namely the problem of detecting a change in the mean value of a Gaussian process.

- Local composite hypothesis : We consider the following Brownian motion with time-varying drift :

$$d\underline{W}_t = \mathbf{1}_{\{t\geq t_0\}}\ \nu\ \mathbf{I}(\theta_0)\ \Upsilon\ dt + \mathbf{I}^{\frac{1}{2}}(\theta_0)\ dW_t \tag{8.1.44}$$

  where ν and/or Υ are unknown.

Since these problems are investigated further in chapter 9 for multidimensional signals, the properties of the nonadditive change detection algorithms, for both scalar and multidimensional signals, are investigated in that chapter.

8.1.3.6 Summary

The use of the local approach for designing algorithms for detecting nonadditive changes in a conditional distribution $p_{\theta^*}(y_n|\mathcal{Y}_1^{n-1})$ can be summarized as follows :

- Compute the efficient score Z_n^* using the formula :

$$Z_n^* = \left. \frac{\partial \ln p_\theta(y_n|\mathcal{Y}_1^{n-1})}{\partial \theta} \right|_{\theta=\theta^*} \tag{8.1.45}$$

- Work with the process $(Z_n^*)_n$ of efficient scores as if it was an independent Gaussian sequence :
  - In the case of local linear hypotheses, the scalar product $\Upsilon^T Z_n^*$ has mean negative before and positive after change; the CUSUM algorithm is based upon this statistic.
  - In the case of a local composite hypothesis, the efficient score has a mean of zero before and nonzero after change; the GLR and CUSUM algorithms are based upon a quadratic form of the efficient scores.
- Apply any of the algorithms described in subsection 7.2.1, according to the amount of *a priori* information about the change magnitude ν and direction Υ.

We follow these steps in subsections 8.2.4, 8.3.2, and 8.3.3 for the general, AR, and ARMA cases. In the case of nonlinear ARMA models in section 8.4, we follow similar steps, but start from a function that is not the likelihood function.

8.2 Conditional Densities and Likelihood Ratio

In this section, we investigate in detail the design of nonadditive change detection algorithms in the basic case of conditional probability distributions, and considering both simple and composite hypotheses. From the discussions in chapters 2 and 7, the two main tools we investigate in this chapter are the CUSUM and GLR algorithms. As in chapter 2, these algorithms use the following basic property of the likelihood ratio :

$$\begin{aligned} \mathbf{E}_{\theta_0}\left[\ln \frac{p_{\theta_1}(y_k|\mathcal{Y}_1^{k-1})}{p_{\theta_0}(y_k|\mathcal{Y}_1^{k-1})}\right] &< 0 \\ \mathbf{E}_{\theta_1}\left[\ln \frac{p_{\theta_1}(y_k|\mathcal{Y}_1^{k-1})}{p_{\theta_0}(y_k|\mathcal{Y}_1^{k-1})}\right] &> 0 \end{aligned} \tag{8.2.1}$$

We thus extend the CUSUM and GLR algorithms to this case, and also introduce other algorithms of interest. Furthermore, we describe the algorithms that result from the use of the local approach in the case of an unknown parameter after change. Before proceeding, we emphasize several key issues that arise when dealing with conditional densities and no longer with ordinary densities as in part I. These topics are also discussed in [Bansal and Papantoni-Kazakos, 1986].

8.2.1 Key Issues Concerning Conditional Densities

To outline the differences between the simplest case investigated in part I and the present case of conditional distributions, we first discuss the consequences of the first two methods for generating changes on the design of the decision functions. Second, we show that the different derivations of the CUSUM algorithm, described in section 2.2, lead now to *different* algorithms. It should be clear that this type of problem is not specific for the CUSUM algorithm, but reflects the key difficulties when detecting nonadditive changes.

8.2.1.1 Generation of Changes

Let us recall that there are several ways of generating nonadditive changes, which result in different forms of the joint probability distribution of the observed signal samples, as we explained in section 8.1. It is of interest to outline the differences in the algorithms that result from this issue. We investigate this point, concentrating on the first two methods of generating a nonadditive change, and thus on formulas (8.1.8) and (8.1.10) for the probability densities of a sample of observations, namely :

$$p(\mathcal{Y}_1^k|k \geq t_0) = p_{\theta_0}(y_1) \left[\prod_{i=2}^{t_0-1} p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})\right] \left[\prod_{i=t_0}^{k} p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})\right] \tag{8.2.2}$$

$$p(\mathcal{Y}_1^k|k \geq t_0) = p_{\theta_0}(y_1) \left[\prod_{i=2}^{t_0-1} p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})\right] \; p_{\theta_1}(y_{t_0}) \left[\prod_{i=t_0+1}^{k} p_{\theta_1}(y_i|\mathcal{Y}_{t_0}^{i-1})\right] \tag{8.2.3}$$

As we show later, the consequences of the differences between these two methods on the design of the decision functions is closely connected to the differences between the various derivations of the CUSUM algorithm. We thus discuss these two issues together.

8.2.1.2 Different CUSUM Derivations

Let us recall the three main derivations of the CUSUM algorithm.

Intuitive derivation The intuitive derivation of the CUSUM algorithm, which uses the basic inequalities (8.2.1), can be written in the following *recursive* manner :

$$\begin{aligned} t_a &= \min\{k \; : g_k \geq h\} \\ g_k &= \left[g_{k-1} + \ln \frac{p_{\theta_1}(y_k|\mathcal{Y}_1^{k-1})}{p_{\theta_0}(y_k|\mathcal{Y}_1^{k-1})}\right]^+ \end{aligned} \tag{8.2.4}$$

In this formulation, we make use of the following likelihood ratio :

$$\breve{\Lambda}_j^k = \frac{p_{\theta_1}(\mathcal{Y}_j^k|\mathcal{Y}_1^{j-1})}{p_{\theta_0}(\mathcal{Y}_j^k|\mathcal{Y}_1^{j-1})} \tag{8.2.5}$$

$$= \prod_{i=j}^{k} \frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})} \tag{8.2.6}$$

Off-line derivation The nonadditive counterpart of the off-line change detection point of view described in subsection 2.2.3 consists of testing between the following hypotheses :

$$\mathbf{H}_0 = \{\mathcal{L}(\mathcal{Y}_1^k) = \mathbf{P}_{\theta_0}\} \text{ and } \mathbf{H}_j = \left\{ \begin{array}{lcl} \mathcal{L}(\mathcal{Y}_1^{j-1}) & = & \mathbf{P}_{\theta_0} \\ \mathcal{L}(\mathcal{Y}_j^k) & = & \mathbf{P}_{\theta_1} \end{array} \right. \tag{8.2.7}$$

for $j \geq 1$. The decision rule is then

$$\begin{array}{rcl} t_a & = & \min\{k \ : g_k \geq h\} \\ g_k & = & \max\limits_{1 \leq j \leq k} \ln \Lambda_1^k(j) \end{array} \tag{8.2.8}$$

where $\Lambda_1^k(j)$ is the likelihood ratio for testing between the hypotheses $\mathbf{H}_0$ and $\mathbf{H}_j$ using the observations $\mathcal{Y}_1^k$.

Open-ended tests Now, let us follow Lorden's idea and define the following set of open-ended tests :

$$t_a = \min_{j=1,2,\ldots} \{T_j\}$$

$$T_j = \min\{k \geq j \ : \ln \breve{\Lambda}_j^k \geq h\} \tag{8.2.9}$$

$$\breve{\Lambda}_j^k = \frac{p_{\theta_1}(\mathcal{Y}_j^k)}{p_{\theta_0}(\mathcal{Y}_j^k)} \tag{8.2.10}$$

First generating method Under assumption (8.2.2), the likelihood ratios involved in the decision functions of the three previous algorithms are

$$\begin{array}{rcl} \breve{\Lambda}_j^k & = & \dfrac{p_{\theta_1}(\mathcal{Y}_j^k|\mathcal{Y}_1^{j-1})}{p_{\theta_0}(\mathcal{Y}_j^k|\mathcal{Y}_1^{j-1})} = \displaystyle\prod_{i=j}^{k} \frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})} \\ \Lambda_1^k(j) & = & \dfrac{p_{\theta_1}(\mathcal{Y}_j^k|\mathcal{Y}_1^{j-1})}{p_{\theta_0}(\mathcal{Y}_j^k|\mathcal{Y}_1^{j-1})} \\ \breve{\Lambda}_j^k & = & \dfrac{p_{\theta_1}(\mathcal{Y}_j^k)}{p_{\theta_0}(\mathcal{Y}_j^k)} \end{array} \tag{8.2.11}$$

Second generating method Here we assume that the two sequences of observations $\mathcal{Y}_1^{j-1}$ and $\mathcal{Y}_j^k$ are mutually independent. Under assumption (8.2.3), the likelihood ratios involved in the decision functions of the three previous algorithms are

$$\begin{array}{rcl} \breve{\Lambda}_j^k & = & \dfrac{p_{\theta_1}(\mathcal{Y}_j^k|\mathcal{Y}_1^{j-1})}{p_{\theta_0}(\mathcal{Y}_j^k|\mathcal{Y}_1^{j-1})} = \displaystyle\prod_{i=j}^{k} \frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})} \\ \Lambda_1^k(j) & = & \dfrac{p_{\theta_1}(\mathcal{Y}_j^k)}{p_{\theta_0}(\mathcal{Y}_j^k|\mathcal{Y}_1^{j-1})} \\ \breve{\Lambda}_j^k & = & \dfrac{p_{\theta_1}(\mathcal{Y}_j^k)}{p_{\theta_0}(\mathcal{Y}_j^k)} \end{array} \tag{8.2.12}$$

Discussion It is obvious that each of the three algorithms differs according to which manner of generating changes is assumed. The main reason for these differences lies in the densities that are conditioned in different ways. Moreover, for a given derivation, the method of generating changes is reflected in the transition around the change point. The importance of the differences between all these possible algorithms obviously depends upon the actual correlations that exist in the observations. In the case of a *stationary* AR(p) process, these differences are negligible when $k \gg p$.

We now use these formulas and our general likelihood framework for change detection to derive the nonadditive change detection algorithms appropriate for the general case of conditional densities. From now on, we assume the *first method* of generating changes and we mainly use the intuitive derivation of CUSUM algorithms, basically because it is recursive.

8.2.2 Simple Hypotheses or Known θ_0 and θ_1

As we explained in chapters 2 and 7, in the case of known parameters before and after change, the relevant change detection algorithm is the CUSUM algorithm, which is equivalent to the GLR algorithm in this case. Therefore, we describe this algorithm in the general case of changes in the parameter vector of a conditional probability density. We also introduce, in this general framework, a quite similar algorithm that has proven useful in the case of AR models, and called the divergence algorithm.

8.2.2.1 CUSUM Algorithm

Here we follow [Nikiforov, 1980, Nikiforov, 1983]. The CUSUM test between the two models before and after change is based upon the likelihood ratio. As shown in (2.2.11), when seen as a repeated SPRT, the CUSUM algorithm can be written as

$$\begin{aligned} t_a &= \min\{k : g_k \geq h\} \\ g_k &= \left(S^k_{k-N_k+1}\right)^+ \qquad (8.2.13) \\ S^k_j &= \ln \frac{p_{\theta_1}(\mathcal{Y}^k_j|\mathcal{Y}^{j-1}_1)}{p_{\theta_0}(\mathcal{Y}^k_j|\mathcal{Y}^{j-1}_1)} \qquad (8.2.14) \end{aligned}$$

where N_k is the number of observations since the last vanishing of g_k :

$$N_k = N_{k-1}\, \mathbf{1}_{\{g_{k-1}>0\}} + 1 \qquad (8.2.15)$$

This algorithm can be written in the following recursive manner :

$$\begin{aligned} g_k &= (g_{k-1} + s_k)^+ \qquad (8.2.16) \\ s_k &= \ln \frac{p_{\theta_1}(y_k|\mathcal{Y}^{k-1}_1)}{p_{\theta_0}(y_k|\mathcal{Y}^{k-1}_1)} \qquad (8.2.17) \end{aligned}$$

Under the assumption of a change generated by the first method (8.2.2), S^k_j should be computed as

$$S^k_j = \sum_{i=j}^{k} \ln \frac{p_{\theta_1}(y_i|\mathcal{Y}^{i-1}_1)}{p_{\theta_0}(y_i|\mathcal{Y}^{i-1}_1)} = \sum_{i=j}^{k} s_i \qquad (8.2.18)$$

The *conditional* expectations of the log-likelihood ratio increment s_i before and after change are as follows :

$$\mathbf{E}_{\theta_0}(s_i|\mathcal{Y}_1^{i-1}) = \mathbf{E}_{\theta_0}\left[\ln \frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})}\middle| \mathcal{Y}_1^{i-1}\right] \tag{8.2.19}$$

$$\mathbf{E}_{\theta_1}(s_i|\mathcal{Y}_1^{i-1}) = \mathbf{E}_{\theta_1}\left[\ln \frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})}\middle| \mathcal{Y}_1^{i-1}\right] \tag{8.2.20}$$

The *unconditional* expectations of the increment s_i are asymptotically given by

$$\begin{aligned}
\lim_{i\to\infty} \mathbf{E}_{\theta_0}(s_i) &= \lim_{n\to\infty} \frac{1}{n}\mathbf{E}_{\theta_0}(S_1^n) \\
&= \lim_{n\to\infty} \frac{1}{n}\int \sum_{i=1}^{n} \ln \frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})} p_{\theta_0}(\mathcal{Y}_1^n) d\mathcal{Y}_1^n \\
&= -\mathbf{K}(\theta_0,\theta_1) \qquad (8.2.21) \\
\lim_{i\to\infty} \mathbf{E}_{\theta_1}(s_i) &= \lim_{n\to\infty} \frac{1}{n}\mathbf{E}_{\theta_1}(S_1^n) \\
&= \lim_{n\to\infty} \frac{1}{n}\int \sum_{i=1}^{n} \ln \frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})} p_{\theta_1}(\mathcal{Y}_1^n) d\mathcal{Y}_1^n \\
&= +\mathbf{K}(\theta_1,\theta_0) \qquad (8.2.22)
\end{aligned}$$

In these two expressions, the first equality comes from the fact that when n goes to infinity, the effect of the initial conditions disappears. The last equality comes from definition (4.1.43) of the Kullback information for a random process. These expressions are investigated further in the particular case of an AR model in section 8.3.

The comparisons between (8.2.19) and (8.2.20) on the one hand and between (8.2.21) and (8.2.22) on the other show that the CUSUM algorithm is neither conditionally nor unconditionally symmetric, which can be undesirable in some cases, as we discuss now.

8.2.2.2 Divergence Algorithm

Here we follow [Basseville and Benveniste, 1983b]. It turns out, in some applications, that the symmetry of the test statistics is a desirable behavior. But it is intuitively obvious that this symmetry may be difficult to obtain, because it is well known that a change that involves an increase in the input variance is much easier to detect than a change that involves a decrease in this variance. This issue of symmetry is thus more critical for spectral changes than for additive changes.

In this paragraph, we describe an algorithm that was originally introduced in the AR case with this motivation, which we call the divergence algorithm. The divergence algorithm is based upon a decision function that is quite similar to (8.2.13) :

$$\begin{aligned}
t_a &= \min\{k : \tilde{g}_k \geq h\} \\
\tilde{g}_k &= \left(\tilde{S}^k_{k-\tilde{N}_k+1}\right)^+ \\
\tilde{S}^k_j &= \sum_{i=j}^{k} \tilde{s}_i \\
&= \sum_{i=j}^{k} \left\{ \ln \frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})} - \mathbf{E}_{\theta_0}\left[\ln \frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})}\middle| \mathcal{Y}_1^{i-1}\right] - \nu \right\}
\end{aligned} \tag{8.2.23}$$

where $\tilde{N}_k = \tilde{N}_{k-1}\ \mathbf{1}_{\{\tilde{g}_{k-1}>0\}} + 1$, namely $\tilde{N}_k$ is the number of observations since the last vanishing of $\tilde{g}_k$. We again assume here the first method of generating changes. Note also that the increment in the cumulative sum is the same here as in (8.2.18), except that we have subtracted the conditional mean value of the increment of the CUSUM algorithm before change, and also a constant quantity ν chosen such that

$$\begin{array}{rcl} \mathbf{E}_{\theta_0}(\tilde{s}_i) & < & 0 \\ \mathbf{E}_{\theta_1}(\tilde{s}_i) & > & 0 \end{array} \tag{8.2.24}$$

The constant ν is thought of as being a kind of minimum magnitude of spectral change to be detected. The choice of this constant in practice is discussed in section 8.6 and in chapter 10.

Now, because of the definition of $\tilde{s}_i$ itself, the *conditional* expectation of $\tilde{s}_i$ before change turns out to be constant :

$$\mathbf{E}_{\theta_0}(\tilde{s}_i|\mathcal{Y}_1^{i-1}) = -\nu \tag{8.2.25}$$

On the other hand, the conditional expectation of the increment after the change is

$$\begin{aligned} \mathbf{E}_{\theta_1}(\tilde{s}_i|\mathcal{Y}_1^{i-1}) &= +\mathbf{E}_{\theta_1}\left[\ln\frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})}\middle|\,\mathcal{Y}_1^{i-1}\right] \\ &\quad -\mathbf{E}_{\theta_0}\left[\ln\frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})}\middle|\,\mathcal{Y}_1^{i-1}\right] - \nu \qquad (8.2.26) \\ &= +\mathbf{E}_{\theta_1}(s_i|\mathcal{Y}_1^{i-1}) - \mathbf{E}_{\theta_0}(s_i|\mathcal{Y}_1^{i-1}) - \nu \qquad (8.2.27) \end{aligned}$$

The *unconditional* expectations are as follows. Before the change, the unconditional expectation is constant exactly like the conditional one :

$$\mathbf{E}_{\theta_0}(\tilde{s}_i) = -\nu \tag{8.2.28}$$

After the change, the unconditional expectation is asymptotically given by

$$\begin{aligned} \lim_{i\to\infty}\mathbf{E}_{\theta_1}(\tilde{s}_i) &= +\lim_{n\to\infty}\frac{1}{n}\mathbf{E}_{\theta_1}(\tilde{S}_1^n) \\ &= +\lim_{n\to\infty}\frac{1}{n}\int\sum_{i=1}^{n}\ln\frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})}p_{\theta_1}(\mathcal{Y}_1^n)d\mathcal{Y}_1^n \qquad (8.2.29) \\ &\quad -\lim_{n\to\infty}\frac{1}{n}\int\sum_{i=1}^{n}\mathbf{E}_{\theta_0}\left[\ln\frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})}\middle|\,\mathcal{Y}_1^{i-1}\right]p_{\theta_1}(\mathcal{Y}_1^n)d\mathcal{Y}_1^n - \nu \end{aligned}$$

The first term on the right side of the last equation is $\mathbf{K}(\theta_1,\theta_0)$ for the same reasons that act in (8.2.22). But, to our knowledge, the second term cannot be computed in the general case of conditional distributions. This unconditional expectation is investigated further in the case of an AR model in section 8.3, where we give a closed form expression. This provides us with a relevant choice for ν, when the two models before and after the change are known, in order that the slopes of $\tilde{s}_i$ before and after the change are *symmetric*.

8.2.3 Composite Hypotheses or Known θ_0 and Unknown θ_1

We now investigate the design of nonadditive change detection algorithms in the case of composite hypotheses corresponding to an unknown parameter after the change. We recall that the previous algorithms designed in the ideal case of known parameters before and after the change can also be implemented when the parameters after the change are unknown, using estimated values, as we discuss in section 8.6.

As we explained in preceding chapters, the key tool in this case is the GLR algorithm. But it is also of interest to outline the problems that can arise from the use of a nonsufficient statistic when designing the decision function. For this reason, we begin this subsection with an introduction to a statistic that is widely used in practice in the AR case. Then we describe the GLR algorithm and discuss its complexity. Finally, we explain what we call the one- and two-model approaches.

8.2.3.1 Monitoring Shifted Log-likelihood Function

In this subsection, we describe a very natural and widely used algorithm for change detection, working without any information about the model after the change. And we show why this decision function is not convenient.

To our knowledge, this algorithm was introduced in the AR case independently in [R.Jones *et al.*, 1970, Borodkin and Mottl', 1976, Segen and Sanderson, 1980] for the purpose of automatic segmentation of EEG signals. In [Segen and Sanderson, 1980], it is shown that this decision function is based upon the shifted log-likelihood function

$$\begin{aligned} \mathcal{S}_k &= \sum_{i=1}^{k} \eta_i \\ \text{where } \eta_i &= -\ln p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1}) + \mathbf{E}_{\theta_0}[\ln p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})|\mathcal{Y}_1^{i-1}] \end{aligned} \tag{8.2.30}$$

In the AR case, this statistic results in monitoring the variance of the innovation, as we discuss in section 8.3.

The conditional and unconditional expectations of η_i before the change are

$$\mathbf{E}_{\theta_0}(\eta_i|\mathcal{Y}_1^{i-1}) = \mathbf{E}_{\theta_0}(\eta_i) = 0 \tag{8.2.31}$$

But the main problem with η_i is that its conditional and unconditional expectations after the change can also be zero for a nonnegligible set of nonadditive changes, which is highly nondesirable from both the intuitive and the statistical detectability points of view. Actually the *conditional* expectation is

$$\begin{aligned} \mathbf{E}_{\theta_1}(\eta_i|\mathcal{Y}_1^{i-1}) &= +\mathbf{E}_{\theta_0}\left[\ln p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})|\mathcal{Y}_1^{i-1}\right] - \mathbf{E}_{\theta_1}\left[\ln p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})|\mathcal{Y}_1^{i-1}\right] \\ &= +\mathbf{E}_{\theta_0}\left[\ln p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})|\mathcal{Y}_1^{i-1}\right] - \mathbf{E}_{\theta_1}\left[\ln p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})|\mathcal{Y}_1^{i-1}\right] \\ &\quad +\mathbf{E}_{\theta_1}\left[\ln \frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})}\middle| \mathcal{Y}_1^{i-1}\right] \end{aligned} \tag{8.2.32}$$

and, denoting by $\mathbf{N}(\theta)$ the Shannon entropy, the *unconditional* expectation is asymptotically

$$\lim_{i\to\infty} \mathbf{E}_{\theta_1}(\eta_i) = \mathbf{N}(\theta_1) - \mathbf{N}(\theta_0) + \mathbf{K}(\theta_1, \theta_0) \tag{8.2.33}$$

again for the same reasons as in (8.2.22).

A *sufficient* condition for η_i having a positive mean after the change is thus

$$\mathbf{N}(\theta_1) \geq \mathbf{N}(\theta_0) \tag{8.2.34}$$

Next, remembering that $\mathbf{K}$ is positive, it is obvious from (8.2.33) that a nonnegligible set of parameters θ_1 results in a zero expectation of η_i after the change when condition (8.2.34) is not fulfilled. In subsection 8.3.2, we show that, in the AR case, the condition (8.2.34) is nothing but an increase in the energy of the input excitation. Finally, recall that we proved in section 4.1 that this statistic is sufficient only for changes in the input variance and not for changes in the spectrum.

8.2.3.2 GLR Algorithm

The GLR approach to the detection of a change in the parameter θ of a conditional probability distribution p_θ, from the known value θ_0 to the *unknown* value θ_1, and occurring at an unknown time t_0, consists of the following detection rule :

$$t_a = \min\{k : g_k \geq h\} \tag{8.2.35}$$

$$g_k = \max_{1\leq j\leq k} \sup_{\theta_1} \sum_{i=j}^{k} \ln \frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})} \tag{8.2.36}$$

As we discussed in section 8.1, in this formula we still assume the first method of generating the change.

In the case of nonadditive changes, the maximization over θ_1 in (8.2.36) is not explicit, as opposed to the case of additive changes in (7.2.46). For this reason, the GLR algorithm (8.2.36) is particularly time-consuming. The number of computations at each time step k grows to infinity with k, because the supremum over θ_1 has to be estimated for each possible change time j between 1 and k. For this reason, the practical implementation of the GLR algorithm is not always possible, and it is of interest to investigate alternative solutions with lower computational cost.

In this chapter, we discuss two possible simplifications of the GLR algorithm, which both cancel the second maximization over the value of the parameter θ_1 after the change. The first simplified algorithm is based upon what we call the two-model approach, and uses the divergence decision function $\tilde{s}_i$ given in (8.2.23); it is described in section 8.6. The second is obtained with the aid of the local approach for change detection, introduced in subsection 8.1.3. Its application is discussed in the subsection 8.2.4 for the CUSUM and GLR algorithms.

8.2.3.3 One- and Two-model Approaches

We complete this discussion about simplification of the GLR algorithm by defining more precisely what we call the one- and two-model approaches. The *one-model approach* to change detection refers to using only one set of model parameters, namely the reference value θ_0, and to testing possible deviations from this reference signature. Typical examples are the shifted log-likelihood ratio algorithm and the GLR algorithm when θ_1 is completely *unknown*. In this approach, the alternative hypothesis about possible values of θ_1 is not simple, because θ_1 is neither assumed known nor estimated. The alternative hypothesis is thus composite, but only a small part of the information about it is used for designing the algorithm. Actually, it should be clear that, according to the level of *a priori* information that is available, it is possible to specify which type of deviations from the reference model are of interest.

On the contrary, as we discussed in chapters 1 and 7, the *two-model approach* refers to using information about two sets of model parameters, one before and one after the change, for designing the decision function or improving its performance. Typical examples are the CUSUM and divergence algorithms. From the point of view of implementation, this two-model approach assumes that the two parameters θ_0 and θ_1 before and after the change are both either known or estimated, and thus corresponds basically to two different simple hypotheses.

Our discussion of the one- and two-model approaches concerns mainly the design of the decision functions. It is important to note that, when the algorithms are considered at the level of tuning of their key parameters, and especially of their threshold, the resulting classification between one- and two-model approaches can be different, because the information about the model after the change is often used for tuning the threshold of a decision function designed with the only model before change.

Finally, let us emphasize that there exist strong connections between the one- and two-model approaches on the one hand and composite and simple hypotheses testing problems on the other hand, but not an equivalence. This comes from the difference between change detection and hypotheses testing problems.

8.2.4 Local Approach for Unknown θ_1

We now discuss the use of the local approach presented in subsection 8.1.3 for designing change detection algorithms in the general case of conditional densities. We distinguish several levels of available *a priori* information about θ_1.

8.2.4.1 CUSUM Algorithm for Local Linear Hypotheses

As in the second case investigated in subsection 7.2.1, we assume that the two parameter sets Θ_0 and Θ_1 before and after the change can be locally separated by a hyperplane, and we again start from the decision function corresponding to the case of simple hypotheses. In this case, the CUSUM algorithm can be written as

$$\begin{aligned} t_a &= \min\{k : g_k \geq h\} \\ g_k &= (g_{k-1} + s_k)^+ \end{aligned} \tag{8.2.37}$$

where s_k is the log-likelihood ratio at time k :

$$s_k = \ln \frac{p_{\theta_1}(y_k|\mathcal{Y}_1^{k-1})}{p_{\theta_0}(y_k|\mathcal{Y}_1^{k-1})} \tag{8.2.38}$$

As mentioned before, we can simplify this algorithm in the local hypotheses situation, using the expansion of the log-likelihood ratio around a reference parameter value θ^* :

$$\begin{aligned} S\left(\theta^* - \frac{1}{2}\nu\,\Upsilon, \theta^* + \frac{1}{2}\nu\,\Upsilon\right) &\approx \nu\,\Upsilon^T \left.\frac{\partial \ln p_\theta(\mathcal{Y}_1^N)}{\partial\theta}\right|_{\theta=\theta^*} \\ &\approx \nu\,\Upsilon^T \sum_{i=1}^{N} Z_i^* = \nu\,\Upsilon^T \mathcal{Z}_N^* \end{aligned} \tag{8.2.39}$$

where $\nu > 0$, Υ is the *unit* vector of the *known* change direction and where Z_k^* is the vector of efficient score, defined in (8.1.45) as

$$Z_k^* = \left.\frac{\partial \ln p_\theta(y_k|\mathcal{Y}_1^{k-1})}{\partial\theta}\right|_{\theta=\theta^*} \tag{8.2.40}$$

The resulting modified CUSUM algorithm is then as in (8.2.37) with the increment

$$s_k = \nu\,\Upsilon^T Z_k^* \tag{8.2.41}$$

From now on, and without loss of generality, we simply consider the CUSUM algorithm associated with

$$s_k = \Upsilon^T Z_k^* \tag{8.2.42}$$

using a threshold that is obtained after division of the previous one by the change magnitude ν. As we explained in subsection 7.2.1, the behavior of the increment of the CUSUM algorithm is such that there exists a separating surface in the parameter space such that

$$\mathbf{E}_\theta(s_k) = 0 \tag{8.2.43}$$

We showed in subsection 4.1.2 that the efficient score has the following property :

$$\mathbf{E}_\theta(Z_k^*) \approx \mathbf{I}(\theta^*)(\theta - \theta^*) \tag{8.2.44}$$

in the neighborhood of the reference point θ^*. Therefore, the possibly complex separating surface between Θ_0 and Θ_1 can be approximated by the following hyperplane :

$$\Upsilon^T \, \mathbf{I}(\theta^*)(\theta - \theta^*) = 0 \tag{8.2.45}$$

Now it is clear that, when the available *a priori* information about the parameters is not in terms of θ_0 and θ_1 but in terms of the hyperplane defined by Υ and θ^* as in (8.2.45), the increment of the CUSUM algorithm is given by (8.2.42). Note that we derive the present local linear CUSUM algorithm *without* using the asymptotic normality of the cumulative sum of efficient scores. We investigate this algorithm in detail in the case of AR models in subsection 8.3.2.

8.2.4.2 GLR and CUSUM Algorithms for Local Quadratic Hypothesis

As in the third case investigated in subsection 7.2.1, we now assume that the parameter θ_0 before change is known and that the parameter set Θ_1 after change is the surface of an ellipsoid centered at θ_0 :

$$\theta(k) = \begin{cases} \theta_0 & \text{when} \quad k < t_0 \\ \theta \ : (\theta - \theta_0)^T \mathbf{I}(\theta_0)(\theta - \theta_0) = b^2 & \text{when} \quad k \geq t_0 \end{cases} \tag{8.2.46}$$

where $b > 0$ is small. In other words, we assume that the Kullback information between the models before and after change is constant. In this case, the GLR algorithm can be written as

$$\begin{aligned} t_a &= \min\{k \ : g_k \geq h\} \\ g_k &= \max_{1 \leq j \leq k} \sup_{(\theta - \theta_0)^T \mathbf{I}(\theta_0)(\theta - \theta_0) = b^2} S_j^k(\theta_0, \theta) \\ S_j^k(\theta_0, \theta) &= \ln \frac{\prod_{i=j}^k p_\theta(y_i | \mathcal{Y}_1^{i-1})}{\prod_{i=j}^k p_{\theta_0}(y_i | \mathcal{Y}_1^{i-1})} \end{aligned} \tag{8.2.47}$$

We use the *second-order* expansion (8.1.22) of the log-likelihood ratio for a sample of size N :

$$S_1^N(\theta_0, \theta) \approx (\theta - \theta_0)^T \mathcal{Z}_N(\theta_0) - \frac{N}{2}(\theta - \theta_0)^T \, \mathbf{I}_N(\theta_0)(\theta - \theta_0) \tag{8.2.48}$$

where $\mathcal{Z}_N$ is the efficient score (8.2.39). We thus get

$$\begin{aligned} \sup_{(\theta - \theta_0)^T \mathbf{I}(\theta_0)(\theta - \theta_0) = b^2} S_j^k(\theta_0, \theta) &\approx \sup_{(\theta - \theta_0)^T \mathbf{I}(\theta_0)(\theta - \theta_0) = b^2} \Big[(\theta - \theta_0)^T \mathcal{Z}_j^k(\theta_0) \\ &\quad - \frac{k - j + 1}{2}(\theta - \theta_0)^T \, \mathbf{I}(\theta_0)(\theta - \theta_0) \Big] \\ &= (k - j + 1) \left(b \, \chi_j^k - \frac{b^2}{2} \right) \end{aligned} \tag{8.2.49}$$

where χ_j^k is defined as

$$(\chi_j^k)^2 = (\bar{Z}_j^k)^T \, \mathbf{I}^{-1}(\theta_0) \, (\bar{Z}_j^k) \tag{8.2.50}$$

and

$$\bar{Z}_j^k = \frac{1}{k-j+1} \sum_{i=j}^{k} Z_i(\theta_0) \tag{8.2.51}$$

We obtain (8.2.49) using the following straightforward transformation :

$$(\theta-\theta_0)^T \mathcal{Z}_1^N(\theta_0) - \frac{N}{2}(\theta-\theta_0)^T \mathbf{I}(\theta_0)(\theta-\theta_0) = \tilde{\theta}\ \tilde{\mathcal{Z}}_1^N - \frac{N}{2}\tilde{\theta}^T\tilde{\theta} \tag{8.2.52}$$

where

$$\begin{array}{rcl} \tilde{\theta} & = & R^T(\theta-\theta_0) \\ \tilde{\mathcal{Z}}_1^N & = & R^{-1}\mathcal{Z}_1^N \\ \mathbf{I} & = & RR^T \\ \mathbf{I}^{-1} & = & (R^{-1})^T R^{-1} \end{array} \tag{8.2.53}$$

and the explicit maximization given in (7.2.22) :

$$\sup_{\tilde{\theta}^T\tilde{\theta}=b^2} \left(N\tilde{\theta}^T \bar{\mathcal{Z}}_1^N - \frac{N}{2}\tilde{\theta}^T\tilde{\theta} \right) = N\left(b\, \|\bar{\mathcal{Z}}_1^N\| - \frac{b^2}{2} \right) \tag{8.2.54}$$

where

$$\bar{\mathcal{Z}}_1^N = \frac{1}{N}\tilde{\mathcal{Z}}_1^N \tag{8.2.55}$$

Therefore, the GLR algorithm for the local quadratic hypothesis case is

$$g_k = \max_{1\le j\le k} (k-j+1)\left(b\,\chi_j^k - \frac{b^2}{2} \right) \tag{8.2.56}$$

$$(\chi_j^k)^2 = (\bar{Z}_j^k)^T\, \mathbf{I}^{-1}(\theta_0)\, (\bar{Z}_j^k) \tag{8.2.57}$$

Note that again we do *not* use the asymptotic normality of the efficient score for this derivation.

Now let us use this asymptotic normality to derive what we call the χ^2-CUSUM algorithm for local quadratic hypotheses. As we discussed in (7.2.12) for case 3 of subsection 7.2.1, the relevant sufficient statistic in this case is

$$\tilde{S}_j^k \approx -(k-j+1)\frac{b^2}{2} + \ln G\left[\frac{\ell}{2}, \frac{b^2(k-j+1)^2(\chi_j^k)^2}{4} \right] \tag{8.2.58}$$

where χ_j^k is defined in (8.2.50). As usual, the stopping rule is then

$$t_a = \min\{k\ :\ \max_{1\le j\le k} \tilde{S}_j^k \ge h\} \tag{8.2.59}$$

As we explained for case 3 of subsection 7.2.1, this algorithm cannot be written in a recursive manner, but can be approximated by another algorithm, which can be recursively written because it is based upon a repeated use of the SPRT with lower threshold zero, and sufficient statistic $\bar{Z}_j^k$:

$$\begin{array}{rcl} g_k & = & \left(\tilde{S}_{k-N_k+1}^k \right)^+ \\ N_k & = & N_{k-1}\, \mathbf{1}_{\{g_{k-1}>0\}} + 1 \end{array} \tag{8.2.60}$$

Note that the statistic $\bar{Z}$ can be written in a recursive manner as

$$\begin{array}{rcl} \bar{Z}_k & = & N_k\, \bar{Z}_{k-N_k+1}^k \\ \bar{Z}_k & = & \bar{Z}_{k-1}\, \mathbf{1}_{\{g_{k-1}>0\}} + Z_k \end{array} \tag{8.2.61}$$

This algorithm is described in the case of AR models in subsection 8.3.2. It should be clear that, as we explained in detail for case 3 of section 7.2.1, there exists a connection between this CUSUM algorithm and the GLR algorithm (8.2.56), when the threshold goes to infinity (and practically for reasonable values of the threshold).

Finally, let us outline that we investigate here a situation similar to case 3 of subsection 7.2.1, but that we could have investigated the other cases (4-7) in the same manner.

8.2.4.3 GLR Algorithm for Local Composite Hypothesis

As in case 8 investigated in subsection 7.2.1, we assume that the parameter before the change is known and that the parameter after the change is unknown :

$$\theta(k) = \begin{cases} \theta_0 & \text{when} \quad k < t_0 \\ \theta \ : \|\theta - \theta_0\| = \nu & \text{when} \quad k \geq t_0 \end{cases} \tag{8.2.62}$$

where $\nu > 0$ is small. In this case, the GLR algorithm can be written as

$$\begin{aligned} t_a &= \min\{k \ : g_k \geq h\} \\ g_k &= \max_{1 \leq j \leq k} \sup_{\theta} S_j^k(\theta_0, \theta) \\ S_j^k(\theta_0, \theta) &= \ln \frac{\prod_{i=j}^k p_\theta(y_i|\mathcal{Y}_1^{i-1})}{\prod_{i=j}^k p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})} \end{aligned} \tag{8.2.63}$$

Again we use the second-order expansion (8.2.48) and recover (4.2.96) :

$$\sup_{\theta} S_j^k(\theta_0, \theta) \approx \frac{k-j+1}{2} \ (\chi_j^k)^2 \tag{8.2.64}$$

where χ_j^k is defined in (8.2.50).

8.3 AR/ARMA Models and the Likelihood Ratio

We now describe, for the important cases of AR and ARMA models, all the algorithms introduced in the previous section in the general case of conditional densities. We mainly discuss the case of AR models, and give only, in the last subsection, some comments concerning the key issues in the ARMA case, because, in the framework of conditional distributions that we use here, the generalization from the AR to the ARMA case is straightforward.

As we explained in the examples of chapter 1, spectral change detection algorithms designed with the aid of an AR model excited by a white noise sequence have proven useful for processing real signals which are known to be closer to the output of an ARMA model excited by a white noise together with an impulse sequence than to the output of such a simple AR model. Continuous speech and seismic signals are examples of this situation. This fact is strongly related to the issue of robustness, which we discuss in chapter 10.

As in section 8.2, we first discuss the case of simple hypotheses, and then discuss the case of composite hypotheses.

8.3.1 Simple Hypotheses

In this subsection, we discuss the application, in the case of AR models, of two algorithms for detecting nonadditive changes in the case of simple hypotheses, namely when the models before and after change are assumed to be known. Recall that this assumption of known models is basically used to derive decision functions, and does not prevent their use in real situations of unknown model parameters, as we discuss in section 8.6.

First we describe the CUSUM algorithm for nonadditive changes, as introduced in (8.2.18). Then we describe the application of the divergence algorithm, which is based upon a different measure of disagreement between the two models, but basically also uses a stopping time of the type (8.2.35).

Note that other distance measures between the two models could be used. Such attempts at designing segmentation algorithms are reported in [Mathieu, 1976, Basseville, 1986].

8.3.1.1 CUSUM Algorithm

Here we follow [Lumel'sky, 1972, Bagshaw and R.Johnson, 1977, Nikiforov, 1978, Nikiforov, 1980, Basseville, 1986]. Using the formula of the conditional density of an AR model given in the example of subsection 8.1.2, it is easy to show that the CUSUM algorithm given in (8.2.13) and (8.2.18) reduces to

$$\begin{aligned} t_a &= \min\{k : g_k \geq h\} \\ g_k &= \left(S^k_{k-N_k+1}\right)^+ \\ S^k_j &= \sum_{i=j}^{k} \ln \frac{p_{\theta_1}(y_i|\mathcal{Y}_{i-p}^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_{i-p}^{i-1})} \\ &= \sum_{i=j}^{k} s_i \end{aligned} \tag{8.3.1}$$

where θ is given in (8.1.2) and

$$s_i = \frac{1}{2} \ln \frac{\sigma_0^2}{\sigma_1^2} + \frac{(\varepsilon_i^0)^2}{2\sigma_0^2} - \frac{(\varepsilon_i^1)^2}{2\sigma_1^2} \tag{8.3.2}$$

In other words, the CUSUM algorithm leads us to monitor the difference between the squared normalized innovations.

Writing the residuals of the two models as

$$\varepsilon_k^l = A_l(z)\ y_k \tag{8.3.3}$$

where

$$A_l(z) = 1 - \sum_{i=1}^{p} a_i^l z^{-i} \tag{8.3.4}$$

we consider the coefficients $c_k^{l|j}, (l, j = 0, 1)$ of the following Taylor expansion :

$$\frac{A_l(z)}{A_j(z)} = 1 + \sum_{k=1}^{\infty} c_k^{l|j} z^{-k} \tag{8.3.5}$$

With these notations, formulas (8.2.21) and (8.2.22), giving the unconditional expectations of s_i before and after change respectively, become in the case of an AR model

$$\mathbf{E}_{\theta_0}(s_i) = \frac{1}{2} + \frac{1}{2}\ln\frac{\sigma_0^2}{\sigma_1^2} - \frac{1}{2}\frac{\sigma_0^2}{\sigma_1^2}\left[1 + \sum_{k=1}^{\infty}(c_k^{1|0})^2\right] \tag{8.3.6}$$

$$\mathbf{E}_{\theta_1}(s_i) = -\frac{1}{2} - \frac{1}{2}\ln\frac{\sigma_1^2}{\sigma_0^2} + \frac{1}{2}\frac{\sigma_1^2}{\sigma_0^2}\left[1 + \sum_{k=1}^{\infty}(c_k^{0|1})^2\right] \tag{8.3.7}$$

by a direct computation from (8.3.2). Note that

$$\sum_{k=1}^{\infty}(c_k^{l|j})^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left\|\frac{A_l(e^{i\omega})}{A_j(e^{i\omega})} - 1\right\|^2 d\omega \tag{8.3.8}$$

8.3.1.2 Divergence Algorithm

We now follow [Basseville and Benveniste, 1983b, Basseville, 1986]. For the same reasons as before, the decision function $\tilde{S}_j^k$ introduced in (8.2.23) reduces in the AR Gaussian case to

$$\tilde{S}_j^k = \sum_{i=j}^{k}\tilde{s}_i \tag{8.3.9}$$

where

$$\begin{aligned}\tilde{s}_i &= \ln\frac{p_{\theta_1}(y_i|\mathcal{Y}_{i-p}^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_{i-p}^{i-1})} - \mathbf{E}_{\theta_0}\left[\ln\frac{p_{\theta_1}(y_i|\mathcal{Y}_{i-p}^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_{i-p}^{i-1})}\right] - \nu \\ &= s_i - \frac{1}{2}\ln\frac{\sigma_0^2}{\sigma_1^2} + \frac{1}{2} - \frac{1}{2\sigma_1^2}\,\mathcal{I}(A_0^T\check{\mathcal{Y}}_{i-p}^i, A_1^T\check{\mathcal{Y}}_{i-p}^i) - \nu\end{aligned} \tag{8.3.10}$$

In this expression, we use

$$\mathcal{I}(\alpha,\beta) = \int \frac{1}{\sigma_0\sqrt{2\pi}}e^{-\frac{1}{2\sigma_0^2}(y-\alpha)^2}(y-\beta)^2\,dy \tag{8.3.11}$$

which, after straightforward computations, can be shown to be

$$\mathcal{I}(\alpha,\beta) = \sigma_0^2 + (\beta-\alpha)^2 \tag{8.3.12}$$

Finally,

$$\tilde{s}_i = -\frac{\varepsilon_i^0\varepsilon_i^1}{\sigma_1^2} + \frac{1}{2}\left(\frac{\sigma_0^2}{\sigma_1^2} + 1\right)\frac{(\varepsilon_i^0)^2}{\sigma_0^2} + \frac{1}{2}\left(\frac{\sigma_0^2}{\sigma_1^2} - 1\right) - \nu \tag{8.3.13}$$

The following expression of $\tilde{s}_i$ is also useful :

$$\tilde{s}_i = \frac{(\varepsilon_i^0)^2}{2\sigma_0^2} - \frac{(\varepsilon_i^1)^2}{2\sigma_1^2} - \frac{1}{2} + \frac{\sigma_0^2 + (\varepsilon_i^0 - \varepsilon_i^1)^2}{2\sigma_1^2} - \nu \tag{8.3.14}$$

The main difference with respect to the CUSUM algorithm (8.3.2) lies in the function of the two residuals which is monitored. In the real case of an unknown parameter after change, this parameter must be chosen

by the user, and corresponds to a minimum magnitude of spectral change to be detected, exactly as for the local hypotheses in subsection 8.3.2. This point is discussed further in section 8.6.

The unconditional expectations of $\tilde{s}_i$ given in (8.2.28) and (8.2.29) are here

$$\mathbf{E}_{\theta_0}(\tilde{s}_i) = -\nu \tag{8.3.15}$$

$$\mathbf{E}_{\theta_1}(\tilde{s}_i) = -1 + \frac{1}{2}\left(\frac{\sigma_0^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_0^2}\right) + \frac{1}{2}\left(1 + \frac{\sigma_1^2}{\sigma_0^2}\right)\sum_{k=1}^{\infty}(c_k^{0|1})^2 - \nu \tag{8.3.16}$$

where the second expectation is directly computed from (8.3.14). Therefore, in the case of known parameters before and after change, the following choice of ν leads to a statistic that has symmetric *slopes* before and after change :

$$\nu = \frac{1}{2}\left[-1 + \frac{1}{2}\left(\frac{\sigma_0^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_0^2}\right) + \frac{1}{2}\left(1 + \frac{\sigma_1^2}{\sigma_0^2}\right)\sum_{k=1}^{\infty}(c_k^{0|1})^2\right] \tag{8.3.17}$$

It should be clear that this choice does *not* provide us with a *symmetric* decision function. For a change detection algorithm to have the same statistical behavior - namely, the same delay for detection for a given mean time between false alarms - when designed and run for a change from θ_0 toward θ_1, as in the converse situation, it is necessary that the expectations of the decision function before and after change be equal, up to a sign, to the same symmetric function of θ_0 and θ_1, which is not the case in (8.3.16) even with the choice of ν in (8.3.17).

8.3.2 Composite Hypotheses

In this subsection, we apply, in the AR case again, the results of subsections 8.2.3 and 8.2.4. We describe four possible algorithms for detecting nonadditive changes in the situation where the model before change is known, and the model after change is not known. Recall that, in practice, another possibility for dealing with unknown parameters consists of replacing the unknown parameter values by conveniently estimated ones, as we discuss in section 8.6.

The four algorithms are ordered with respect to the amount of *a priori* information that is available, ranging from no information at all to an information in terms of a confidence ellipsoid. The last two algorithms use the local approach to change detection, which we described in subsection 8.2.4.

8.3.2.1 Monitoring Squared Innovations

This algorithm was introduced for the general case in subsection 8.2.3 using the shifted log-likelihood function, but was originally derived for the AR case. It is based upon the intuitive idea of detecting a nonadditive change by monitoring the innovation of an AR model with the aid of a test for its variance. It should be noted [Mehra and Peschon, 1971] that this test assumes that the innovation sequence is white, and thus a test for whiteness should be done first. This test is based upon the following fact. Under the hypothesis of no change, the cumulative sum

$$\mathcal{S}_k = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{\varepsilon_i^2}{\sigma_0^2} - 1\right) \tag{8.3.18}$$

is asymptotically, when k goes to infinity, distributed as a χ^2 distribution with k degrees of freedom. This property is exploited in [R.Jones *et al.*, 1970] and is frequently used in the engineering literature. The

following is also true :

$$\tilde{\mathcal{S}}_k = \frac{1}{2\sqrt{k}} \sum_{i=1}^{k} \eta_i \tag{8.3.19}$$

$$= \frac{1}{2\sqrt{k}} \sum_{i=1}^{k} \left(\frac{\varepsilon_i^2}{\sigma_0^2} - 1 \right) \tag{8.3.20}$$

is asymptotically distributed as a Gaussian law $\mathcal{L}(0,1)$ under $\mathbf{H}_0$. This fact is used in [Borodkin and Mottl', 1976, Segen and Sanderson, 1980]. In [Segen and Sanderson, 1980], it is shown that this decision function is of the form (8.2.30). Thus, let us transpose the discussions made before to an AR model. In the AR case, the condition (8.2.34) under which η_i has a positive mean after change, reduces to

$$\sigma_1 \geq \sigma_0 \tag{8.3.21}$$

and the *conditional* expectation of the increment of the sum is

$$\mathbf{E}_{\theta_1}(\eta_i | \mathcal{Y}_{i-p}^{i-1}) = \frac{1}{2} \left\{ \frac{\sigma_1^2}{\sigma_0^2} - 1 + \frac{[(A_0 - A_1)^T \check{\mathcal{Y}}_{i-p}^{i-1}]^2}{\sigma_0^2} \right\} \tag{8.3.22}$$

Therefore, in the case of a nonadditive change with decrease in the energy of the excitation, namely $\sigma_1 < \sigma_0$, this mean value can be unpredictably positive, or negative, or even zero, which is clearly undesirable in practice.

Usually, this statistic works satisfactorily in the converse situation, for changes occurring with an increase in energy, and specifically for the detection of spikes in EEG signals. But it is of limited interest in most real situations where the type of nonadditive changes to be detected is not so restricted.

In subsection 4.1.2, we showed that in fact η_i is basically a sufficient statistic for detecting a change in the variance of the excitation only, and not in the AR parameter, which explains its poor behavior in many circumstances.

8.3.2.2 GLR Algorithm

In the case of an AR(p) process, the GLR detection rule is again as in (8.2.36) with

$$g_k = \max_{1 \leq j \leq k} \sup_{\theta_1} \sum_{i=j}^{k} \ln \frac{p_{\theta_1}(y_i | \mathcal{Y}_{i-p}^{i-1})}{p_{\theta_0}(y_i | \mathcal{Y}_{i-p}^{i-1})}$$

$$= \max_{1 \leq j \leq k} \sup_{\theta_1} \sum_{i=j}^{k} \left[\frac{1}{2} \ln \frac{\sigma_0^2}{\sigma_1^2} + \frac{(\varepsilon_i^0)^2}{2\sigma_0^2} - \frac{(\varepsilon_i^1)^2}{2\sigma_1^2} \right] \tag{8.3.23}$$

A clever implementation of the complete GLR algorithm is reported in [Appel and von Brandt, 1983], which reduces the computing time necessary for the optimization with respect to the unknown change time.

8.3.2.3 CUSUM Algorithm for Local Linear Hypotheses

We now describe, again for an AR process, the local linear CUSUM algorithm introduced in subsection 8.2.4. This algorithm is based upon the decision rule (8.2.37) with the increment of the decision function given in (8.2.42). In the AR case, this increment is computed with the aid of the efficient score given in

(4.1.101) :

$$s_k = \Upsilon^T Z_k^* \tag{8.3.24}$$

$$Z_k = \begin{pmatrix} \frac{1}{\sigma^2} \check{\mathcal{Y}}_{k-p}^{k-1} \varepsilon_k \\ \\ \frac{1}{\sigma} \left(\frac{\varepsilon_k^2}{\sigma^2} - 1 \right) \end{pmatrix} \tag{8.3.25}$$

$$\varepsilon_k = A^T \check{\mathcal{Y}}_{k-p}^k \tag{8.3.26}$$

where A is again as in (8.1.13).

8.3.2.4 GLR and χ^2-CUSUM Algorithms for a Local Composite Hypothesis

We now discuss for the AR case the characteristic features of the GLR (8.2.56) and χ^2-CUSUM (8.2.58) algorithms for the local quadratic hypothesis, and of the GLR algorithm (8.2.64) for the local composite hypothesis. These three algorithms are based upon the computation of the following quadratic form :

$$(\chi_j^k)^2 = (\bar{Z}_j^k)^T \, \mathbf{I}^{-1}(\theta_0) \, (\bar{Z}_j^k) \tag{8.3.27}$$

$$= \frac{1}{(k-j+1)^2} \, (\mathcal{Z}_j^k)^T \, \mathbf{I}^{-1}(\theta_0) \, (\mathcal{Z}_j^k) \tag{8.3.28}$$

$$= \frac{1}{(k-j+1)^2} \left\{ \sigma^2 \, (\tilde{\mathcal{Z}}_j^k)^T \, \mathbf{T}_p^{-1}(\theta_0) \, (\tilde{\mathcal{Z}}_j^k) + \frac{1}{2} \left[\sum_{i=j}^{k} \left(\frac{\varepsilon_i^2}{\sigma^2} - 1 \right) \right]^2 \right\} \tag{8.3.29}$$

$$\tilde{\mathcal{Z}}_j^k = \sum_{i=j}^{k} \frac{1}{\sigma^2} \, \check{\mathcal{Y}}_{i-p}^{i-1} \, \varepsilon_i \tag{8.3.30}$$

where we make use of the Fisher information given in (4.1.102) and of the inversion formula (8.1.27) for Toeplitz matrices.

8.3.3 ARMA Models and the Likelihood Ratio

We now discuss the key new issues in the ARMA case with respect to the AR case for the design of nonadditive change detection algorithms. It is obvious that the decision rules are the same as before. We concentrate on the computation of the increment of the decision functions.

8.3.3.1 CUSUM Algorithm for Simple Hypotheses

In the case of simple hypotheses, the increment of the CUSUM algorithm is

$$s_i = \frac{1}{2} \ln \frac{\sigma_0^2}{\sigma_1^2} + \frac{(\varepsilon_i^0)^2}{2\sigma_0^2} - \frac{(\varepsilon_i^1)^2}{2\sigma_1^2} \tag{8.3.31}$$

where the residual ε_i^l of the ARMA model l $(l = 0, 1)$ is given in the example of section 8.1 :

$$\varepsilon_i^l = A_l^T \check{\mathcal{Y}}_{i-p}^i - B_l^T \check{\mathcal{E}}_{i-q}^{i-1} \tag{8.3.32}$$

8.3.3.2 Composite Hypotheses

We now describe for the ARMA process the algorithms described in the case of composite hypotheses in subsection 8.2.3.

Local linear CUSUM algorithm This algorithm is based upon the decision rule (8.2.37) with the increment of the decision function given in (8.2.42). In the ARMA case, this increment is computed with the aid of the efficient score given in (4.1.106) :

$$s_k = \Upsilon^T Z_k^* \tag{8.3.33}$$

$$Z_k = \begin{pmatrix} \frac{1}{\sigma^2} \check{\mathcal{A}}_{k-p}^{k-1} \varepsilon_k \\ \frac{1}{\sigma^2} \check{\mathcal{B}}_{k-q}^{k-1} \varepsilon_k \\ \frac{1}{\sigma} \left(\frac{\varepsilon_k^2}{\sigma^2} - 1 \right) \end{pmatrix} \tag{8.3.34}$$

where $\check{\mathcal{A}}_{k-p}^{k-1}$ and $\check{\mathcal{B}}_{k-q}^{k-1}$ are the sets of α and β ordered backward, and

$$\begin{aligned} \alpha_{k-i} &= -\frac{\partial \varepsilon_k}{\partial a_i} \\ \beta_{k-j} &= -\frac{\partial \varepsilon_k}{\partial b_j} \end{aligned} \tag{8.3.35}$$

are the outputs of the same AR model :

$$\begin{aligned} \alpha_k &= -\sum_{j=1}^{q} b_j \alpha_{k-j} + y_k \\ \beta_k &= -\sum_{j=1}^{q} b_j \beta_{k-j} + \varepsilon_k \end{aligned} \tag{8.3.36}$$

as we explained in subsection 4.1.2.

GLR and χ^2-CUSUM Algorithms for a Local Composite Hypothesis We now discuss for the ARMA case the characteristic features of the GLR (8.2.56) and χ^2-CUSUM (8.2.58) algorithms for the local quadratic hypothesis, and of the GLR algorithm (8.2.64) for the local composite hypothesis. We use the efficient score and the Fisher information matrix for an ARMA process given in section 4.1.2.

These three algorithms are based upon the computation of the following quadratic form :

$$\begin{aligned} (\chi_j^k)^2 &= (\bar{Z}_j^k)^T \, \mathbf{I}^{-1}(\theta_0) \, (\bar{Z}_j^k) \\ &= \frac{1}{(k-j+1)^2} \, (\mathcal{Z}_j^k)^T \, \mathbf{I}^{-1}(\theta_0) \, (\mathcal{Z}_j^k) \\ &= \frac{1}{(k-j+1)^2} \left\{ \sigma^2 \, (\tilde{\mathcal{Z}}_j^k)^T \, \tilde{\mathbf{I}}^{-1}(\theta_0) \, (\tilde{\mathcal{Z}}_j^k) + \frac{1}{2} \left[\sum_{i=j}^{k} \left(\frac{\varepsilon_i^2}{\sigma^2} - 1 \right) \right]^2 \right\} \end{aligned} \tag{8.3.37}$$

$$\tilde{\mathcal{Z}}_j^k = \begin{pmatrix} \sum_{i=j}^{k} \frac{1}{\sigma^2} \check{\mathcal{A}}_{i-p}^{i-1} \, \varepsilon_i \\ \sum_{i=j}^{k} \frac{1}{\sigma^2} \check{\mathcal{B}}_{i-q}^{i-1} \, \varepsilon_i \end{pmatrix} \tag{8.3.38}$$

where

$$\begin{aligned}\mathbf{I}(\theta) &= \begin{pmatrix} \tilde{\mathbf{I}}(\theta) & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\sigma^2}\mathbf{E}_\theta[\check{\mathcal{A}}_{k-p}^{k-1}(\check{\mathcal{A}}_{k-p}^{k-1})^T] & \frac{1}{\sigma^2}\mathbf{E}_\theta[\check{\mathcal{A}}_{k-p}^{k-1}(\check{\mathcal{B}}_{k-q}^{k-1})^T] & 0 \\ \frac{1}{\sigma^2}\mathbf{E}_\theta[\check{\mathcal{B}}_{k-q}^{k-1}(\check{\mathcal{A}}_{k-p}^{k-1})^T] & \frac{1}{\sigma^2}\mathbf{E}_\theta[\check{\mathcal{B}}_{k-q}^{k-1}(\check{\mathcal{B}}_{k-q}^{k-1})^T] & 0 \\ 0 & 0 & \frac{2}{\sigma^2} \end{pmatrix}\end{aligned} \tag{8.3.39}$$

8.3.4 Generalization to the Transfer Function

We now show that the likelihood approach can be extended to change detection in dynamic models represented with the aid of transfer functions, as described in subsection 3.2.4. For simplicity we concentrate on the case of an ARX model :

$$A(z)y_k = C(z)u_k + \varepsilon_k \tag{8.3.40}$$

where $(u_k)_k$ is a *known* input sequence, $(\varepsilon_k)_k$ is a white noise sequence with variance σ^2, and $A(z)$ and $C(z)$ are polynomials in z^{-1}. The extension to ARMAX models is straightforward.

We use the following notation :

$$A^T = \begin{pmatrix} 1 & -a_1 & \dots & -a_p \end{pmatrix} \tag{8.3.41}$$

$$C^T = \begin{pmatrix} c_0 & \dots & c_l \end{pmatrix} \tag{8.3.42}$$

$$\theta^T = \begin{pmatrix} a_1 & \dots & a_p & c_0 & \dots & c_l & \sigma \end{pmatrix} \tag{8.3.43}$$

In this case, the conditional density of the observation y_k with respect to past values of both the observations and the known inputs is

$$\begin{aligned} p_\theta(y_k|\mathcal{Y}_{k-p}^{k-1}, \mathcal{U}_{k-l}^{k}) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(A^T\check{\mathcal{Y}}_{k-p}^{k} - C^T\check{\mathcal{U}}_{k-l}^{k})^2} \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\varepsilon_k^2} \end{aligned} \tag{8.3.44}$$

where

$$\varepsilon_k = A^T\check{\mathcal{Y}}_{k-p}^{k} - C^T\check{\mathcal{U}}_{k-l}^{k} \tag{8.3.45}$$

Thus, the log-likelihood ratio increment is

$$s_k = \frac{1}{2}\ln\frac{\sigma_0^2}{\sigma_1^2} + \frac{(\varepsilon_k^0)^2}{2\sigma_0^2} - \frac{(\varepsilon_k^1)^2}{2\sigma_1^2} \tag{8.3.46}$$

The key difference between (8.3.44) and the ARMA case (8.1.17) lies in the method of computing ε_k.

Therefore, the CUSUM algorithm in this case can be written as

$$\begin{aligned} t_a &= \min\{k : g_k \geq h\} \\ g_k &= (g_{k-1} + s_k)^+ \end{aligned} \tag{8.3.47}$$

where s_k is defined above.

8.4 Non-Likelihood-Based Algorithm

It results from this and the preceding chapters that the likelihood ratio is a general and powerful tool for designing change detection algorithms, but it can be complex in some cases. In the two previous sections, we describe a solution for reducing this complexity, based upon what is called the local approach. In this section, we describe another solution, which is valid not only for ARMA models, but also in the much larger subset of conditional probability distributions (8.1.5), which we call nonlinear ARMA models. The main idea consists of using a function of observations other than the likelihood ratio, and of also using the local approach. We show here that a function of observations relevant for monitoring can be built by starting from the structure of a recursive identification algorithm.

Before proceeding, let us emphasize that this solution provides us with a systematic way of associating a change detection algorithm with any recursive identification algorithm adapted to the class of models (8.1.5) for multidimensional signals, and that a diagnosis algorithm can be also designed within this framework. This diagnosis method is described in the case of vibration monitoring in chapter 9. In this section, we discuss only the detection problem for scalar signals.

As we mentioned in section 6.2, we consider the following class of semi-Markov processes :

$$\left\{ \begin{array}{rcl} \mathbf{P}(X_k \in B | X_{k-1}, X_{k-2}, \ldots) & = & \int_B \pi_{\theta^\dagger}(X_{k-1}, dx) \\ y_k & = & f(X_k) \end{array} \right. \tag{8.4.1}$$

where $\pi_\theta(X, dx)$ is the transition probability of the Markov chain $(X_k)_k$, f is a nonlinear function, and $\theta^\dagger$ is the true value of the parameter.

Example 8.4.1 (ARMA model). *An ARMA(p, q) process can be written in the form of (8.4.1) with a linear function f in the following manner. We recall from subsection 3.2.4 that an ARMA model*

$$y_k = \sum_{i=1}^{p} a_i y_{k-i} + \sum_{j=1}^{q} b_j v_{k-j} + v_k \tag{8.4.2}$$

where $(v_k)_k$ is a Gaussian white noise sequence with variance $R = \sigma^2$, can be written as a state-space model in an innovation form as

$$\left\{ \begin{array}{rcl} X_{k+1} & = & FX_k + Gv_k \\ y_k & = & HX_k + v_k \end{array} \right. \tag{8.4.3}$$

Then the extended state $\mathcal{X}_k = \begin{pmatrix} X_{k+1} \\ y_k \end{pmatrix}$ is easily found to be a Markov process, and thus $y_k = f(\mathcal{X}_k) = \begin{pmatrix} 0 & 1 \end{pmatrix} \mathcal{X}_k$ is a semi-Markov process.

The semi-Markovian property is invariant under many operations. A function of a semi-Markov process is a semi-Markov process; the same is true for a sliding time-window of observations of a semi-Markov process and a stable linear combination of such observations. This is useful in the discussion that follows.

Now, let us explain the *genesis* of the non-likelihood-based statistics for change detection, starting from recursive identification. For the class of nonlinear models (8.4.1), there exists [Benveniste *et al.*, 1990] a family of parameter estimation algorithms, which can be written in the following way :

$$\theta_k = \theta_{k-1} + \Delta \, \mathcal{K}(\theta_{k-1}, \mathcal{Y}_1^k) \tag{8.4.4}$$

where Δ is a gain matrix and $\mathcal{K}$ satisfies some regularity conditions so that this algorithm converges toward the true value $\theta^\dagger$ of the parameter θ. The notation $\mathcal{K}(\theta, \mathcal{Y}_1^k)$ stands for a measurable random vector where the dependence on $\mathcal{Y}_1^k$ is only through a function of the observations, which is a semi-Markov process. This becomes clear in the following example.

Example 8.4.2 (ARMA models and the extended least-squares algorithm). *For ARMA models, we use the following notation :*

$$\underline{\theta}^T = (a_1 \quad \dots \quad a_p \quad b_1 \quad \dots \quad b_q) \tag{8.4.5}$$

$$\mathcal{P}_k^T = \left((\check{\mathcal{Y}}_{k-p}^{k-1})^T \quad (\check{\mathcal{E}}_{k-q}^{k-1})^T \right) \tag{8.4.6}$$

$$e_k(\underline{\theta}) = y_k - \underline{\theta}^T \mathcal{P}_k \tag{8.4.7}$$

where $\check{\mathcal{Y}}$ and $\check{\mathcal{E}}$ are the vectors of the observations y and residuals e ordered backward.

The extended least-squares (ELS) algorithm can be written as

$$\underline{\theta}_k = \underline{\theta}_{k-1} + \frac{1}{k} \Sigma_k^{-1} \, \mathcal{P}_k \, e_k(\underline{\theta}_{k-1}) \tag{8.4.8}$$

where

$$\Sigma_k = \Sigma_{k-1} + \frac{1}{k} (\mathcal{P}_k \, \mathcal{P}_k^T - \Sigma_{k-1}) \tag{8.4.9}$$

Because of the above-mentioned property of a semi-Markov process, $\mathcal{P}_k$ is a semi-Markov process. Therefore, the LS algorithm (8.4.8) is of the form (8.4.4) with

$$\mathcal{K}(\underline{\theta}, \mathcal{Y}_1^k) = \mathcal{P}_k \, e_k(\underline{\theta}) = \mathcal{K}(\underline{\theta}, \mathcal{P}_k, y_k) \tag{8.4.10}$$

When $\mathcal{K}$ is a stochastic gradient, with respect to θ, of the likelihood functional $-\mathbf{E}_\theta \ln p_\theta(\mathcal{Y}_1^k)$, then $\mathcal{K}$ is exactly the efficient score in the Gaussian case, as we show in the example after the next. This is the case for the least-squares algorithms. In several other cases, for example, the instrumental variables estimation method or the extended least-squares algorithm, $\mathcal{K}$ is not a stochastic gradient. This has an important consequence for the detection issue later.

The change detection problem consists of detecting changes in the true value $\theta^\dagger$ of the parameter of the model (8.4.1). To reduce the complexity of the detection algorithm, we constrain the decision function to use only the information contained in the statistic :

$$\breve{y}_k^* = \mathcal{K}(\theta^*, \mathcal{Y}_1^k) \tag{8.4.11}$$

for a *fixed* nominal (assumed) value θ^* of θ, and not upon the likelihood function, and we also use a local point of view. In other words, we test between the hypotheses :

$$\mathbf{H}_0 = \left\{ \theta^\dagger = \theta^* \right\} \text{ and } \mathbf{H}_1 = \left\{ \theta^\dagger = \theta^* + \frac{\nu}{\sqrt{N}} \, \Upsilon \right\} \tag{8.4.12}$$

Example 8.4.3 (ARMA models and the ELS algorithm - contd.). *In the ARMA case, we have $\breve{y}_k^* = \mathcal{K}(\underline{\theta}^*, \mathcal{P}_1^k, y_k)$, where $\mathcal{K}$ is defined in (8.4.10) and $(y_k)_k$ is governed by (8.4.2).*

In contrast to the standard situation in stochastic approximation theory, we are now in a situation where the true (but hidden) parameter value is varying, because subject to a change, and the parameter θ^* used in the algorithm is fixed. Also, in this section, we investigate change detection problems, and *not* tracking problems as in section 2.5. In other words, the only use that we make here of the recursive identification algorithm is the design of the monitoring statistic (and possibly the guess of a relevant nominal value θ^*); we do *not* use the change detection algorithm to improve the tracking capability of the recursive identification algorithm, as opposed to what we explained in section 2.5.

Let us now introduce some notation. In analogy with (8.1.38), we consider the following cumulative sum :

$$\breve{S}_{N,t}(\theta^*) = \frac{1}{\sqrt{N}} \sum_{i=1}^{[Nt]} \breve{y}_i^* \tag{8.4.13}$$

for $t \in [0,1]$. Let

$$\kappa(\theta, \theta^*) = \lim_{k\to\infty} \mathbf{E}_\theta(\breve{y}_k^*) = \lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} \breve{y}_i^* \tag{8.4.14}$$

be the asymptotic mean value of the statistic $\breve{y}^*$. Note that $\kappa(\theta^\dagger, \theta^\dagger) = 0$ and, consequently,

$$\frac{\partial}{\partial\theta}\kappa(\theta,\theta^\dagger)\Big|_{\theta=\theta^\dagger} = -\frac{\partial}{\partial\theta^\dagger}\kappa(\theta,\theta^\dagger)\Big|_{\theta^\dagger=\theta} \tag{8.4.15}$$

We also use the notation

$$\dot\kappa(\theta^*) = \frac{\partial}{\partial\theta}\kappa(\theta,\theta^\dagger)\Big|_{\theta=\theta^*,\theta^\dagger=\theta^*} \tag{8.4.16}$$

It turns out that the asymptotic behavior of the cumulative sum (8.4.13) is Gaussian, as stated in the following central limit theorem [Benveniste *et al.*, 1987, Benveniste *et al.*, 1990, Ladelli, 1990]. When $N \to \infty$,

$$\begin{array}{lllll} \text{under} \quad p_{\theta^*} \ : & \breve{\Sigma}_N^{-\frac{1}{2}}(\theta^*) \ \breve{S}_{N,t}(\theta^*) & \rightsquigarrow & (W_t)_{t\in[0,1]} \\ \text{under} \quad p_{\theta^*+\frac{\nu}{\sqrt{N}}\Upsilon} \ : & \breve{\Sigma}_N^{-\frac{1}{2}}(\theta^*) \ \Big[\breve{S}_{N,t}(\theta^*) + \nu\ \dot\kappa(\theta^*)\ \Upsilon t\Big] & \rightsquigarrow & (W_t)_{t\in[0,1]} \end{array} \tag{8.4.17}$$

where the covariance matrix is given by

$$\breve{\Sigma}_N(\theta^*) = \sum_{i=-N}^{N} \mathrm{cov}_{\theta^*}(\breve{y}_i^*, \breve{y}_1^*) \tag{8.4.18}$$

and where $(W_t)_{t\in[0,1]}$ is an ℓ-dimensional Brownian motion. This invariance principle implicitly assumes that the limit $\breve{\Sigma} = \lim_{N\to\infty} \breve{\Sigma}_N$ exists. Furthermore, because of (8.4.11), (8.4.17), and (8.4.18), the resulting detection procedure is left unchanged if $\mathcal{K}$ in (8.4.4) is premultiplied by an invertible matrix gain.

The initial change detection problem is thus transformed into the problem of detecting a change in the drift of a normalized Brownian motion, which we summarize in the following formula :

$$d\breve{W}_t = -\mathbf{1}_{\{t\geq t_0\}}\ \nu\ \dot\kappa(\theta^*)\ \Upsilon\ dt + \breve{\Sigma}^{\frac{1}{2}}(\theta^*)\ dW_t \tag{8.4.19}$$

This means that the solution to the change detection problem consists of working with the process $(\breve{y}_n^*)_n$ of nonnormalized efficient scores as if it was an independent Gaussian sequence with mean zero before change and $-\nu\ \dot\kappa(\theta^*)\Upsilon$ after change, and covariance matrix $\breve{\Sigma}(\theta^*)$. This problem can be solved with the aid of any algorithm of subsection 7.2.1 according to the amount of *a priori* information about ν and Υ.

Consequently, the two possible local approaches for detecting nonadditive changes in the nonlinear ARMA models (8.1.5) can be summarized as

- *Likelihood and efficient score approach :*

$$z_i = \frac{\partial \ln p_\theta(y_i|\mathcal{Y}_1^{i-1})}{\partial\theta} \tag{8.4.20}$$

$$d\underline{W}_t = \mathbf{1}_{\{t\geq t_0\}}\ \nu\ \mathbf{I}(\theta^*)\ \Upsilon\ dt + \mathbf{I}^{\frac{1}{2}}(\theta^*)\ dW_t \tag{8.4.21}$$

- *Non-likelihood approach :*

$$\breve{y}_i = \mathcal{K}(\theta, \mathcal{Y}_1^i) \tag{8.4.22}$$

$$d\breve{W}_t = -\mathbf{1}_{\{t \geq t_0\}}\, \nu\, \dot{\kappa}(\theta^*)\, \Upsilon\, dt + \breve{\Sigma}^{\frac{1}{2}}(\theta^*)\, dW_t \tag{8.4.23}$$

Recall that the efficient scores to be considered are *nonnormalized* with respect to the sample size.

We now investigate three examples. The first two are concerned with AR and ARMA models and are discussed to show that these two local approaches coincide. The third example is concerned with the problem of detecting changes in the AR part of an ARMA model, which is discussed in detail in chapter 9 and is related to the vibration monitoring example of section 1.2.

Example 8.4.4 (AR model and the stochastic gradient least-squares algorithm). *In the AR case, the model (8.1.5) is reduced to*

$$y_k = \underline{\theta}^T\, \mathcal{Y}_{k-p}^{k-1} + v_k \tag{8.4.24}$$

where

$$\underline{\theta}^T = \begin{pmatrix} a_1 & \dots & a_p \end{pmatrix} \tag{8.4.25}$$

and the stochastic gradient least-squares algorithm is nothing but

$$\underline{\theta}_k = \underline{\theta}_{k-1} + \gamma\, \breve{\mathcal{Y}}_{k-p}^{k-1}\, e_k(\underline{\theta}_{k-1}) \tag{8.4.26}$$

where

$$e_k(\underline{\theta}) = y_k - (\breve{\mathcal{Y}}_{k-p}^{k-1})^T\, \underline{\theta} \tag{8.4.27}$$

Therefore,

$$\breve{y}_k = \breve{\mathcal{Y}}_{k-p}^{k-1}\, e_k(\underline{\theta}) \tag{8.4.28}$$

Straightforward computations show that the covariance matrix of this statistic is

$$\breve{\Sigma}(\underline{\theta}) = \sigma^2 \operatorname{cov}(\breve{\mathcal{Y}}_{k-p}^{k-1}) \tag{8.4.29}$$

and the derivative of the mean is

$$\dot{\kappa}(\underline{\theta}) = -\,\mathbf{E}_{\underline{\theta}}\, (\breve{\mathcal{Y}}_{k-p}^{k-1})(\breve{\mathcal{Y}}_{k-p}^{k-1})^T = -\frac{1}{\sigma^2}\, \breve{\Sigma}(\underline{\theta}) \tag{8.4.30}$$

It results from (4.1.101) and (4.1.102) that

$$\begin{pmatrix} \frac{1}{\sigma^2}\breve{y}_k \\ \frac{1}{\sigma}(\frac{e_k^2}{\sigma^2} - 1) \end{pmatrix} = z_k \tag{8.4.31}$$

$$\begin{pmatrix} \frac{1}{\sigma^4}\breve{\Sigma}(\underline{\theta}) & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix} = \mathbf{I}(\theta) \tag{8.4.32}$$

In other words, the two above-mentioned approaches coincide. The key reason is that the stochastic gradient least-squares algorithm minimizes the likelihood *functional* $-\mathbf{E}_\theta \ln p_\theta(\mathcal{Y}_1^k)$.

Example 8.4.5 (ARMA models and the ELS algorithm - contd.). *As we already explained, the extended least-squares (ELS) algorithm can be written as*

$$\underline{\theta}_k = \underline{\theta}_{k-1} + \frac{1}{k}\, \Sigma_k^{-1}\, \mathcal{P}_k\, e_k(\underline{\theta}_{k-1}) \tag{8.4.33}$$

where $\mathcal{P}_k^T$ is defined in (8.4.6). Therefore,

$$\breve{y}_k = \mathcal{P}_k \; e_k(\underline{\theta}) \tag{8.4.34}$$

Straightforward computations show that the covariance matrix of this statistic is

$$\breve{\Sigma}(\underline{\theta}) = \sigma^2 \begin{pmatrix} \mathbf{E}_{\underline{\theta}}[\breve{\mathcal{Y}}_{k-p}^{k-1}(\breve{\mathcal{Y}}_{k-p}^{k-1})^T] & \mathbf{E}_{\underline{\theta}}[\breve{\mathcal{Y}}_{k-p}^{k-1}(\breve{\mathcal{E}}_{k-q}^{k-1})^T] \\ \\ \mathbf{E}_{\underline{\theta}}[\breve{\mathcal{E}}_{k-q}^{k-1}(\breve{\mathcal{Y}}_{k-p}^{k-1})^T] & \mathbf{E}_{\underline{\theta}}[\breve{\mathcal{E}}_{k-q}^{k-1}(\breve{\mathcal{E}}_{k-q}^{k-1})^T] \end{pmatrix} \tag{8.4.35}$$

and the derivative of the mean value is

$$\begin{aligned} \dot{\kappa}(\underline{\theta}) &= -\mathbf{E}_{\underline{\theta}} \left(\mathcal{P}_k \mathcal{P}_k^T \right) - \mathbf{E}_{\underline{\theta}} \left[\mathcal{P}_k \left(\frac{\partial}{\partial \underline{\theta}} \mathcal{P}_k^T \; \underline{\theta} \right)^T \right] \\ &= -\frac{1}{\sigma^2} \breve{\Sigma}(\underline{\theta}) - \mathbf{E}_{\underline{\theta}} \left[\mathcal{P}_k \left(\frac{\partial}{\partial \underline{\theta}} \mathcal{P}_k^T \; \underline{\theta} \right)^T \right] \end{aligned} \tag{8.4.36}$$

It results from the comparison between these two expressions and formula (8.3.39) for the Fisher information matrix of an ARMA process that the local ELS-based detection approach is not *identical to the local likelihood detector.*

Actually, we recover in these two examples what we said before : *Each time we use an identification algorithm that minimizes the* likelihood *functional $-\mathbf{E}_\theta \ln p_\theta(\mathcal{Y})$ through a stochastic gradient (or stochastic Newton) algorithm for a* Gaussian *process, we in fact deal with the* efficient score*, and thus our general and optimal local likelihood ratio approach and the particular local non-likelihood approach coincide in this case*. But, the extended least-squares algorithm is not a stochastic gradient and thus the two detectors are different.

As another illustration of what new information can be obtained by the local non-likelihood approach, let us now discuss the third example. We consider the problem of detecting changes in the AR part of a nonstationary ARMA model having a time-varying MA part. In this problem, the MA parameters are nothing but nuisance parameters which prevent the use of the likelihood ratio approach. The key reason for this is the fact that the Fisher information matrix of an ARMA process is not block diagonal with respect to the AR coefficients on one hand, and the MA coefficients on the other one; thus, these two parts are tightly coupled in the likelihood function. Therefore, we use the instrumental variables identification algorithm, which is known to decouple the two types of coefficients, but which does *not* minimize the likelihood functional. Thus, the associated detection algorithm does not coincide with the local likelihood ratio detector.

Example 8.4.6 (AR part of an ARMA model and the IV algorithm). *The instrumental variables (IV) method for estimating the AR part of an ARMA(p,q) model can be written in the form of (8.4.4) as follows :*

$$\begin{aligned} \underline{\theta}_k &= \underline{\theta}_{k-1} + \frac{1}{k} \Sigma_k^{-1} \; \breve{\mathcal{Y}}_{k-q-p}^{k-q-1} \; e_k(\underline{\theta}_{k-1}) \\ \Sigma_k &= \Sigma_{k-1} + \frac{1}{k} \left[\breve{\mathcal{Y}}_{k-q-p}^{k-q-1} (\breve{\mathcal{Y}}_{k-p}^{k-1})^T - \Sigma_{k-1} \right] \\ e_k(\underline{\theta}) &= y_k - \underline{\theta}^T \breve{\mathcal{Y}}_{k-p}^{k-1} \end{aligned} \tag{8.4.37}$$

where $\underline{\theta}$ is defined in (8.4.25). Therefore, the statistic

$$\breve{y}_k = \breve{\mathcal{Y}}_{k-q-p}^{k-q-1}\, e_k(\underline{\theta}) \tag{8.4.38}$$

has the covariance matrix

$$\begin{aligned}\breve{\Sigma}(\underline{\theta}) &= \sum_{k=-\infty}^{+\infty} \mathbf{E}_{\underline{\theta}}(\breve{y}_k \breve{y}_1^T) \\ &= \sum_{k=-q}^{q} \mathbf{E}_{\underline{\theta}} \left[\breve{\mathcal{Y}}_{k-q-p}^{k-q-1} (\breve{\mathcal{Y}}_{1-q-p}^{-q})^T (y_k - \underline{\theta}^T \breve{\mathcal{Y}}_{k-p}^{k-1})(y_1 - \underline{\theta}^T \breve{\mathcal{Y}}_{1-p}^{0})^T \right]\end{aligned} \tag{8.4.39}$$

Furthermore, the derivative of the mean of $\breve{y}$ is

$$\dot{\kappa}(\underline{\theta}) = -\mathbf{E}_{\underline{\theta}}(\breve{\mathcal{Y}}_{k-p}^{k-1})(\breve{\mathcal{Y}}_{k-q-p}^{k-q-1})^T \tag{8.4.40}$$

The detection algorithm based upon $\breve{y}$ and $\breve{\Sigma}$ is used in the case of multidimensional signals in chapter 9 for solving the vibration monitoring problem of the example 1.2.5.

8.5 Detectability

We now discuss the issue of detectability of nonadditive changes, using the detectability definition in chapter 6, namely in terms of the positivity of the Kullback information between the two distributions of the process before and after change.

The reason we select Kullback information and not Kullback divergence for defining the statistical detectability is that it is useful to distinguish between the detectability of a change from θ_0 to θ_1 and the detectability of a change from θ_1 to θ_0, both because it is of interest to investigate the robustness of an algorithm to an error in the direction of the change, and because it is known that a lack of symmetry of change detection algorithms can occur in practice, especially for spectral changes. However, for a change in the mean in the Gaussian case and for local hypotheses, this distinction between Kullback information and divergence does not hold, because they both rely on the same symmetric quadratic form. This is the reason we investigated the detectability mainly by computing the Kullback divergence in chapter 7.

In this section, we concentrate our discussion of detectability on the case of AR models. We first recall several results that we reported in subsection 4.1.2 and section 8.3 about the computation of the Kullback information in this case. Then we discuss several consequences of the detectability definition.

8.5.1 Kullback Information in AR Models

In subsection 4.1.2 and section 8.3, we derived the following four formulas concerning the Kullback information between two AR models. The first two are frequency-domain formulations and the last two ones are time-domain formulations.

1. From (4.1.108) and (3.2.36), it results that

$$\mathbf{K}(\theta_1, \theta_0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \frac{|A^0(e^{i\omega})|^2}{|A^1(e^{i\omega})|^2} \frac{\sigma_1^2}{\sigma_0^2} - 1 - \ln \left[\frac{|A^0(e^{i\omega})|^2}{|A^1(e^{i\omega})|^2} \frac{\sigma_1^2}{\sigma_0^2} \right] \right\} d\omega \tag{8.5.1}$$

where $A(z)$ is defined in (8.3.4).

2. The comparison between (8.2.22) and (8.3.7) shows that

$$\mathbf{K}(\theta_1, \theta_0) = -\frac{1}{2} - \frac{1}{2} \ln \frac{\sigma_1^2}{\sigma_0^2} + \frac{1}{2} \frac{\sigma_1^2}{\sigma_0^2} \left[1 + \sum_{k=1}^{\infty} (c_k^{0|1})^2 \right] \tag{8.5.2}$$

where the coefficients $c_k^{0|1}$ are defined in (8.3.5).

3. We recall from subsection 4.1.2 that

$$\mathbf{K}(\theta_1, \theta_0) = \frac{1}{2\sigma_0^2} (A^0)^T \mathbf{T}_p(\Phi_1) A^0 - \frac{1}{2} \ln \frac{\sigma_1^2}{\sigma_0^2} - \frac{1}{2} \tag{8.5.3}$$

where Φ is the power spectrum defined in (3.2.36), A is defined in (8.1.13), and $\mathbf{T}_p$ is the $(p+1) \times (p+1)$ Toeplitz matrix filled with the covariances.

4. We explained in subsection 4.1.2 that when the difference between the two vector parameters is small, the following approximation holds :

$$\mathbf{K}(\theta_1, \theta_0) \approx \frac{1}{2} (\theta_1 - \theta_0)^T \, \mathbf{I}(\theta_0) \, (\theta_1 - \theta_0) \tag{8.5.4}$$

where $\mathbf{I}$ is the Fisher information matrix, which is given in (4.1.102) in the AR case, or can be computed analytically in terms of the magnitudes and angles of the poles [Bruzzone and Kaveh, 1984].

The second and fourth formulas for $\mathbf{K}$ seem to be the most tractable, and are used in our discussion.

8.5.2 Discussion

Now let us investigate, in the case of AR models, several consequences of the detectability definition in terms of the Kullback information. We have already discussed the problems that can arise when changes in AR coefficients are associated with changes in the input variance. Now we discuss the situation where the input variance is constant, and concentrate on the detectability of changes in the frequencies. First, for a given Euclidian distance between the poles of two AR(1) models, we show that the amount of Kullback information is greater when the poles are closer to the unit circle, which shows the experimentally obvious fact that changes in damped frequencies are much more difficult to detect than changes in weakly damped frequencies. Second, we compute the Kullback information in an AR($2p$) model corresponding to a given number of changes in one frequency and its complex conjugate, again in the two situations of low and high damping, and explain that a change in a damped frequency can be masked by another weakly damped frequency.

First, we note that, in the case of a constant input variance, the formula (8.5.2) reduces to

$$\mathbf{K}(\theta_1, \theta_0) = \frac{1}{2} \sum_{k=1}^{\infty} (c_k^{0|1})^2 \tag{8.5.5}$$

Let us thus consider the case of an AR(1) model. Straightforward computations show that

$$c_k^{0|1} = (a_1^1)^{k-1} (a_1^1 - a_1^0) \tag{8.5.6}$$

and thus (8.5.5) can be written as

$$\mathbf{K}(\theta_1, \theta_0) = \frac{1}{2} \frac{(a_1^1 - a_1^0)^2}{1 - (a_1^1)^2} \tag{8.5.7}$$

where a_1^0, a_1^1 are as in (8.1.2). Since the pole of an AR(1) model (8.1.1) is exactly the AR coefficient, the Euclidian distance between the poles is thus

$$d^2 = (a_1^1 - a_1^0)^2 \tag{8.5.8}$$

Assume a *given* value d of the Euclidean distance between the poles of the two models. It results from the last two equations that

$$\mathbf{K}(\theta_1, \theta_0) = \frac{1}{2} \frac{d^2}{1 - (a_1^1)^2} \tag{8.5.9}$$

From this, we deduce that, for a given value of the Euclidean distance between the poles of the two models, when the pole after change goes to zero, the Kullback information $\mathbf{K}$ decreases to $\frac{d^2}{2}$; when the pole goes to the unit circle, the Kullback information grows to infinity. Therefore, we obtain the intuitively obvious fact that, for a given value of the Euclidean distance between the poles of the two models, a change between weakly damped frequencies is much more easily detectable than a change in highly damped frequencies.

Second, we consider an AR($2p$) model having p times two pairwise conjugate poles, in which we introduce a *given* amount of change in *one* of the p eigenfrequencies. Computing the resulting Kullback information, we show that the detectability of such a change depends upon the location of the changed pole with respect to the unit circle. First we note that, because of (8.5.5) and the definition of the coefficients $c_k^{0|1}$ in (8.3.5), we only need to compute this Taylor expansion for an AR(2) model, because we assume a change in only one pole (and its complex conjugate). We thus need

$$\frac{1 - 2\rho\cos\omega_0 z^{-1} + \rho^2 z^{-2}}{1 - 2\rho\cos\omega_1 z^{-1} + \rho^2 z^{-2}} = 1 + \sum_{k=1}^{\infty} c_k^{0|1} z^{-k} \tag{8.5.10}$$

Using

$$\frac{1}{1 - 2\rho\cos\omega_1 z^{-1} + \rho^2 z^{-2}} = 1 + \sum_{k=1}^{\infty} \alpha_k^1 z^{-k} \tag{8.5.11}$$

straightforward computations lead to the following formulas :

$$\begin{array}{rcl} c_k^{0|1} & = & \alpha_k^1 - 2\rho\cos\omega_0\ \alpha_{k-1}^1 + \rho^2 \alpha_{k-2}^1 \quad \text{for} \quad k \geq 1 \\ \alpha_k^1 & = & \rho^k\ \sum_{j=0}^{\left[\frac{k}{2}\right]} C_{k-j}^j (-1)^j (2\cos\omega_1)^{k-2j} \quad \text{for} \quad k \geq 1 \\ \alpha_0^1 & = & 1 \\ \alpha_{-1}^1 & = & 0 \\ C_k^j & = & \frac{k!}{j!(k-j)!} \end{array} \tag{8.5.12}$$

which we rewrite as

$$\begin{array}{rcl} \alpha_k^1 & = & \rho^k Q_k(\cos\omega_1) \quad \text{for} \quad k \geq 0 \\ c_k^{0|1} & = & \rho^k \left[Q_k(\cos\omega_1) - 2\cos\omega_0\ Q_{k-1}(\cos\omega_1) + Q_{k-2}(\cos\omega_1) \right] \end{array} \tag{8.5.13}$$

where the kth-order polynomial $Q_k(\cos\omega)$ can be easily checked to satisfy

$$Q_k(\cos\omega) + Q_{k-2}(\cos\omega) = 2\cos\omega\ Q_{k-1}(\cos\omega) \quad \text{for} \quad k \geq 2 \tag{8.5.14}$$

with $Q_0(\cos\omega) = 1$ and $Q_1(\cos\omega) = 2\cos\omega$. Therefore,

$$\begin{array}{rcl} c_k^{0|1} & = & 2\rho^k(\cos\omega_1 - \cos\omega_0)\ Q_{k-1}(\cos\omega_1) \\ & = & 2\rho(\cos\omega_1 - \cos\omega_0)\ \alpha_{k-1}^1 \end{array} \tag{8.5.15}$$

From (8.5.5), we thus get

$$\mathbf{K}(\theta_1,\theta_0) = 2\rho^2(\cos\omega_1-\cos\omega_0)^2 \sum_{k=0}^{\infty}\rho^{2k}\, Q_k^2(\cos\omega_1) \tag{8.5.16}$$

$$= 2\rho^2(\cos\omega_1-\cos\omega_0)^2 \sum_{k=0}^{\infty}(\alpha_k^1)^2 \tag{8.5.17}$$

Now we show that a closed form expression for the Kullback information can be obtained using relation (8.5.14). For this purpose, let us note

$$\begin{array}{lcl} S_0(\omega) & = & \sum_{k=0}^{\infty}\rho^{2k}\, Q_k^2(\cos\omega) \\ S_1(\omega) & = & \sum_{k=1}^{\infty}\rho^{2k}\, Q_k(\cos\omega)\, Q_{k-1}(\cos\omega) \\ S_2(\omega) & = & \sum_{k=2}^{\infty}\rho^{2k}\, Q_k(\cos\omega)\, Q_{k-2}(\cos\omega) \end{array} \tag{8.5.18}$$

We have

$$\mathbf{K}(\theta_1,\theta_0) = 2\rho^2(\cos\omega_1-\cos\omega_0)^2\, S_0(\omega_1) \tag{8.5.19}$$

From (8.5.14), using first premultiplication by ρ^k and raising to the power 2, and then premultiplication by $\rho^k Q_{k-1}$ and $\rho^k Q_{k-2}$, respectively, and summing each of the three resulting equations over k, we deduce the three following relations between the $S_\imath, \imath = 0,1,2$:

$$\begin{array}{rcrcrcl} (1-4\rho^2\cos^2\omega+\rho^4)\, S_0(\omega) & & & + & 2\, S_2(\omega) & = & 1 \\ 2\rho^2\cos\omega\, S_0(\omega) & - & (1+\rho^2)\, S_1(\omega) & & & = & -2\rho^4\cos\omega \\ \rho^4\, S_0(\omega) & - & 2\rho^2\cos\omega\, S_1(\omega) & + & S_2(\omega) & = & 0 \end{array}$$

Straightforward computations then lead to

$$S_0(\omega) = \frac{1+\rho^2+8\rho^6\cos^2\omega}{(1-\rho^2)\,[(1+\rho^2)^2-4\rho^2\cos^2\omega]} \tag{8.5.20}$$

From this and (8.5.19), we deduce that the Kullback information is small whenever $\rho < 1$. For a given amount of change in the frequency ω from ω_0 to ω_1, the information $\mathbf{K}$ decreases to zero with ρ. In other words, a change in a *damped* frequency is difficult to detect.

8.6 Implementation Issues

In this section, we establish a bridge between formal problem statements with different levels of *a priori* information and practical experience where known values of parameters are replaced by estimated ones. We discuss the implementation of the decision functions which we introduced in the preceding sections, and concentrate our discussion on AR models. We do not discuss this issue for the case of the additive changes in chapter 7 because, once the transformation from observations to innovations has been achieved, no other implementation question arises.

In the case of spectral changes, in the real situation of unknown parameters θ_0 before and θ_1 after change, we must explain how to tune the algorithms discussed above, which all assume θ_0 to be known and assume different levels of available *a priori* information about θ_1. The general idea consists of replacing all unknown parameter values by estimated ones. But in some sense spectral parameters are more difficult to estimate than mean values, especially when short delays for detection are required. This point is one of the main motivations for the present discussion.

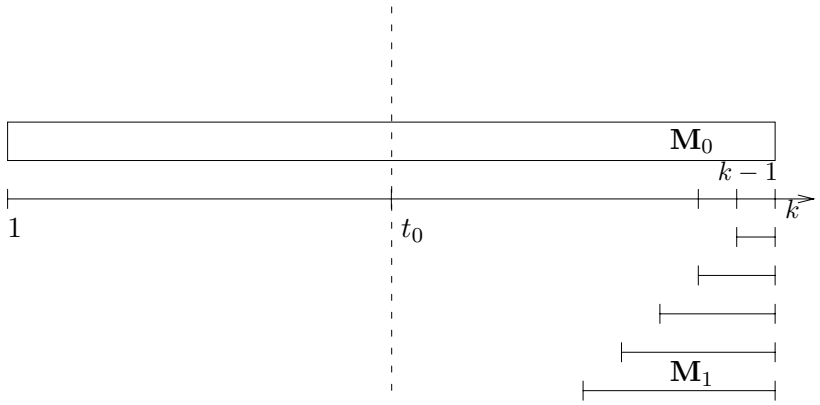

Figure 8.5 Estimation of θ_1 in the GLR algorithm.

First, we note that in our framework of on-line change detection algorithms, replacing the unknown parameter value θ_0 *before* change by its estimate is a standard approach. Many identification algorithms can be used for this purpose. The literature on this topic is extensive [Ljung, 1987, Söderström and Stoïca, 1989, Benveniste *et al*., 1990] and we do not discuss this here. The main issue is rather *on which data* the parameter θ_0 should be estimated – on all the data that are available up to time k or only on a fixed size sample of most recent ones? The former choice is intuitively the best because it implies the use of all of the available information lying in the observations. It often requires the inclusion of a slight forgetting ability in the chosen identification algorithm. This choice is made for the divergence algorithm, and can be made for the other decision functions as well.

Estimating the value of the unknown parameter *after* change is a less straightforward issue, especially in our framework of on-line change detection algorithms! A solution that gets rid of the on-line estimation constraints consists of using prior investigations on other records of data to estimate a set of relevant values for θ_1 and tune the change detection algorithm accordingly. This is discussed in chapter 10. The main problem that arises then is the robustness issue, namely the performance, under actual parameter values θ_0 and θ_1, of an algorithm tuned with the parameter values $\tilde{\theta}_0$ and $\tilde{\theta}_1$. This question is also discussed in chapter 10.

To infer other relevant practical solutions, let us discuss the way by which the GLR algorithm solves this problem. It results from formula (8.3.23) that this algorithm consists of comparing, through the log-likelihood ratio, the parameter θ_0, estimated in the growing time window $\mathbf{M}_0$ of time instants up to time k, to values of θ_1, estimated in all possible time windows $\mathbf{M}_1$ ending at current time k. This is depicted in figure 8.5. We already mentioned that this approach is time-consuming. Moreover, it is well known that AR models are not very reliable when estimated on short data records. This leads to what is called boundary problems for change detection algorithms, which is discussed in section 8.7 for the AR case. To overcome these two drawbacks, one solution consists of keeping from the GLR algorithm only the idea of comparing, with the aid of a convenient spectral distance measure, *a long-term* model corresponding to the absence of a change, and *a short-term* model corresponding to the model after a possible change. This is exactly the key implementation of the two-model approach discussed before, which we suggest for practical use of the divergence decision function (8.2.23). This is depicted in figure 8.6. The parameter θ_0 is estimated in the *growing time window* $\mathbf{M}_0$, and the parameter θ_1 is estimated in the *sliding fixed-size time window* $\mathbf{M}_1$ ending at current time k. A straightforward consequence of this type of implementation is that after each detection, it is necessary to inhibit the computation of the decision function during a time interval with length at least equal to the size of $\mathbf{M}_1$, which is often referred to as a dead zone. The choice of the size of $\mathbf{M}_1$ should thus

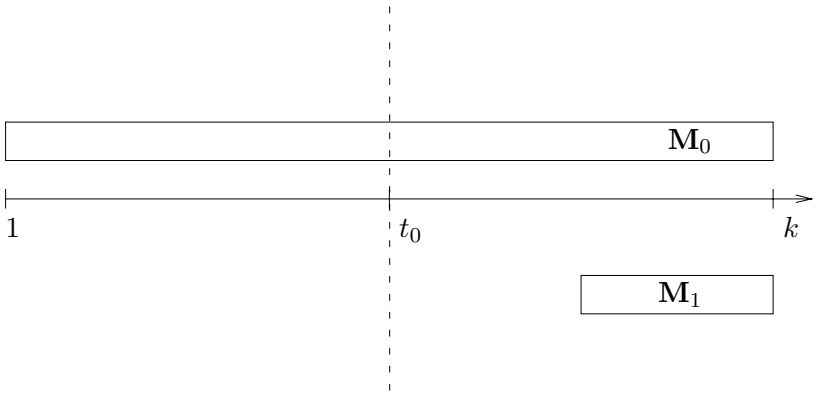

Figure 8.6 Practical implementation of the divergence algorithm.

result from a tradeoff between the precision of the estimation required for the parameter θ_1 and the mean time between changes actually present in the processed signal. This implementation unavoidably introduces a limitation of the resulting algorithm with respect to the presence of frequent changes or equivalently short segments.

8.7 Off-line Algorithms

In practical applications, after the solution of an on-line detection problem, the problem of the *off-line estimation* of the change time often arises. This is the case for the problem of estimating onset times in seismic signals. Onset time estimation is a typical off-line estimation problem. But, because of the length of the signals, their nonstationarity, and the number of seismic waves, whose onset times have to be estimated, off-line estimation algorithms cannot be used on the initial data. As we explain in chapters 1 and 11, a relevant approach in such a situation consists of the two following steps :

- Use an on-line detection algorithm to get a preliminary estimation of the onset time (either the alarm time or the estimated change time discussed for the CUSUM and GLR algorithms).
- Use an off-line change time estimation algorithm for a data window with fixed length and centered at this preliminary estimated time instant. The length of window should be chosen according to the minimum time between two successive onsets. This off-line change time estimation problem is often called *a posteriori* estimation in the literature.

In this section, we discuss the problem of the off-line estimation of a change time in the case of AR processes, because our experience is that, for this problem as for many other signal processing ones, the AR model leads to a satisfactory trade-off between complexity and efficiency of the corresponding algorithms. We first describe the maximum likelihood estimation (MLE) of the change time. Then, using the concept of Kullback information, we discuss the connections between the off-line estimation algorithm and the on-line detection algorithms.

These topics are investigated in [Kligiene and Telksnys, 1983, Deshayes and Picard, 1983, Picard, 1985, Deshayes and Picard, 1986].

8.7.1 Maximum Likelihood Estimation

As we explained in chapters 2 and 7, the MLE estimation of the change time and of the parameters before and after change consists of the following triplicate maximization :

$$(\hat{t}_0, \hat{\theta}_0, \hat{\theta}_1) = \arg \max_{1+\imath \leq k \leq N-\imath} \sup_{\theta_1} \sup_{\theta_0} \left[\ln p_{\theta_0}(\mathcal{Y}_1^{k-1}) + \ln p_{\theta_1}(\mathcal{Y}_k^N | \mathcal{Y}_1^{k-1}) \right] \tag{8.7.1}$$

when the change is generated according to the first method, and in

$$(\hat{t}_0, \hat{\theta}_0, \hat{\theta}_1) = \arg \max_{1+\imath \leq k \leq N-\imath} \sup_{\theta_1} \sup_{\theta_0} \left[\ln p_{\theta_0}(\mathcal{Y}_1^{k-1}) + \ln p_{\theta_1}(\mathcal{Y}_k^N) \right] \tag{8.7.2}$$

when the change is generated according to the second method. In these expressions, $\imath$ is the length of the boundary dead zones in which we cannot compute the likelihood function. Recall that the conditional and unconditional log-likelihood functions are

$$\ln p_\theta(\mathcal{Y}_1^n | \mathcal{Y}_{1-p}^0) = -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (A^T \check{\mathcal{Y}}_{i-p}^i)^2 \tag{8.7.3}$$

and

$$\ln p_\theta(\mathcal{Y}_1^n) = -\frac{n}{2} \ln \sigma^2 + \frac{1}{2} \ln \det \mathbf{T}_p^{-1} - \frac{1}{2\sigma^2} S_1^n(\theta) \tag{8.7.4}$$

$$S_1^n(\theta) = (\mathcal{Y}_1^p)^T \mathbf{T}_p^{-1} \mathcal{Y}_1^p + \sum_{i=p+1}^{n} (A^T \check{\mathcal{Y}}_{i-p}^i)^2 \tag{8.7.5}$$

respectively.

Let us comment further on the computational issues related to this MLE, which is known to be complex. It results from (8.7.1)-(8.7.2) that, at each time k, we must estimate two maximum likelihood values of the parameters θ_0 and θ_1, and then compute two log-likelihood functions before and after change. It is well known that a convenient tradeoff between complexity and efficiency in these types of computations is the solution of the Yule-Walker equations, which is known to provide us with asymptotically efficient estimates of the autoregressive coefficients. On the other hand, fast algorithms exist for inverting the Toeplitz matrix $\mathbf{T}$. Moreover, the computations of the conditional and unconditional log-likelihood functions and their maximization can be done in a completely recursive manner. These algorithms can be found in [Kligiene and Telksnys, 1983, Nikiforov, 1983, Nikiforov and Tikhonov, 1986, Nikiforov *et al*., 1989].

8.7.2 Connection with On-line Algorithms

Let us now discuss the relationships between the properties of the on-line change detection algorithms and the above-mentioned off-line change time estimation algorithm, in order to outline the key common features between these two points of view. Here we follow [Picard, 1985, Deshayes and Picard, 1986] and we explain the available asymptotic results concerning the precision of this MLE estimate of the change time.

We consider the following asymptotic point of view, in order to avoid the degeneracy of the boundary problems :

$$\begin{aligned} & N \to \infty, \;\; t_0(N) \to \infty, \;\; [N - t_0(N)] \to \infty \\ & \lim_{N \to \infty} \|\theta_1(N) - \theta_0(N)\| = 0 \\ & \lim_{N \to \infty} \frac{t_0(N)[N - t_0(N)]}{N} \|\theta_1(N) - \theta_0(N)\|^2 = +\infty \end{aligned} \tag{8.7.6}$$

Under these assumptions, the following holds :

$$\begin{aligned} \tfrac{t_0(N)[N-t_0(N)]}{N}\ [\theta_1(N)-\theta_0(N)]^T\mathbf{I}(\theta_0)[\theta_1(N)-\theta_0(N)]\ \left[\tfrac{\hat{t}_0-t_0(N)}{N}\right] \\ \rightsquigarrow \arg\sup_{t\in\mathbf{R}}\left\{W_t-\tfrac{|t|}{2}\right\} \end{aligned} \tag{8.7.7}$$

where $(W_t)_{t\in\mathbf{R}}$ is a normalized Brownian motion satisfying $W_0 = 0$. Let us comment on this result. The term inside the square brackets is the asymptotic relative error of the change time estimate. From this formula, it is obvious that the scale of this error is

$$\left\{[\theta_1(N)-\theta_0(N)]^T\mathbf{I}(\theta_0)[\theta_1(N)-\theta_0(N)]\right\}^{-1} \approx \mathbf{K}^{-1}(\theta_1(N),\theta_0(N)) \tag{8.7.8}$$

Note first that, in the present asymptotic local framework and as we showed in section 4.1.2, this quantity is nothing but the inverse of the Kullback information between the two models before and after change. Second, as we explained in chapter 5 and also in chapter 9, this quantity plays a key role in the ARL function and more generally in the properties of the optimal change detection algorithms. This fact provides us with the relevant bridge between on-line and off-line points of view for solving change detection problems : From (8.7.7), we deduce that the relative error in the change time estimate decreases when the Kullback information between the two models increases, exactly as (5.2.10) and (9.5.31) show us that the delay for detection for a given false alarm rate also decreases when this information increases.

8.8 Notes and References

Section 8.1

The use of the local approach for designing change detection algorithms was first proposed in [Nikiforov, 1978, Nikiforov, 1980, Nikiforov, 1983]. The usefulness of this approach for designing non-likelihood based change detection algorithms has been recognized in [Basseville *et al.*, 1986, Benveniste *et al.*, 1987, Benveniste *et al.*, 1990, Zhang *et al.*, 1994].

Section 8.2

The problem of detecting a nonadditive change in a conditional distribution was addressed in [Lumel'sky, 1972, Borodkin and Mottl', 1976, Bagshaw and R.Johnson, 1977, Nikiforov, 1978, Segen and Sanderson, 1980, Basseville and Benveniste, 1983b, Nikiforov, 1983].

Section 8.3

The AR case was investigated in [Lumel'sky, 1972, Borodkin and Mottl', 1976, Bagshaw and R.Johnson, 1977, Nikiforov, 1978, Segen and Sanderson, 1980, Basseville and Benveniste, 1983b, Nikiforov, 1983]. The use of the divergence decision function was proposed in [Basseville, 1982, Basseville and Benveniste, 1983b, Basseville, 1986]. The use of the local quadratic CUSUM algorithms was proposed in [Nikiforov, 1978, Nikiforov, 1983] and the local GLR was introduced in [Basseville *et al.*, 1987a, Benveniste *et al.*, 1987, Benveniste *et al.*, 1990].

Section 8.4

The idea of using together a non-likelihood-based statistic and the local approach for solving complex nonadditive change detection problems was introduced in [Basseville *et al.*, 1986] for solving the problem of vibration monitoring with the aid of the instrumental statistic, as described in [Basseville *et al.*, 1987a]. This idea was extended and generalized to other non-likelihood-based statistics in [Benveniste *et al.*, 1987, Benveniste *et al.*, 1990]. An extension of the method, allowing model reduction and biased identification, is reported in [Zhang, 1991, Zhang *et al.*, 1994], together with an application to the monitoring of the combustion chambers of a gas turbine.

Section 8.5

To our knowledge, the detectability of nonadditive changes with the aid of the Kullback information is introduced here for the first time.

Section 8.7

The first investigations of the off-line algorithms for conditional densities and ARMA models were reported in [Kligiene and Telksnys, 1983]. The theoretical investigations of the properties of the off-line algorithms were reported in [Deshayes and Picard, 1983, Picard, 1985, Deshayes and Picard, 1986]. The number of papers on this topic is very large, but we do not cite them because it is not the main purpose of this book.

8.9 Summary

Local Approach

$$S_1^N(\theta_0, \theta_N) \approx \nu \Upsilon^T \Delta_N(\theta_0) - \frac{\nu^2}{2} \Upsilon^T \, \mathbf{I}_N(\theta_0) \, \Upsilon$$

Conditional Distribution

CUSUM algorithm

$$\begin{aligned} t_a &= \min\{k : g_k \geq h\} \\ g_k &= \left(S^k_{k-N_k+1}\right)^+ \\ S^k_j &= \sum_{i=j}^{k} s_i \\ s_i &= \ln \frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})} \end{aligned}$$

Divergence algorithm

$$\begin{aligned} t_a &= \min\{k : \tilde{g}_k \geq h\} \\ \tilde{g}_k &= \left(\tilde{S}^k_{k-\tilde{N}_k+1}\right)^+ \end{aligned}$$

$$\begin{aligned}\tilde{S}_j^k &= \sum_{i=j}^{k} \tilde{s}_i \\ \tilde{s}_i &= \ln \frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})} - \mathbf{E}_{\theta_0}\left[\ln \frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})}\middle| \mathcal{Y}_1^{i-1}\right] - \nu\end{aligned}$$

Shifted log-likelihood function

$$\begin{aligned}\mathcal{S}_k &= \sum_{i=1}^{k} \eta_i \\ \eta_i &= -\ln p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1}) + \mathbf{E}_{\theta_0}[\ln p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})|\mathcal{Y}_1^{i-1}]\end{aligned}$$

GLR algorithm

$$\begin{aligned}t_a &= \min\{k : g_k \geq h\} \\ g_k &= \max_{1\leq j\leq k} \sup_{\theta_1} \sum_{i=j}^{k} \ln \frac{p_{\theta_1}(y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(y_i|\mathcal{Y}_1^{i-1})}\end{aligned}$$

Local linear CUSUM algorithm

$$\begin{aligned}t_a &= \min\{k : g_k \geq h\} \\ g_k &= (g_{k-1} + s_k)^+ \\ s_k &= \Upsilon^T Z_k^* \\ Z_k^* &= \left.\frac{\partial \ln p_\theta(y_k|\mathcal{Y}_1^{k-1})}{\partial \theta}\right|_{\theta=\theta^*}\end{aligned}$$

Local quadratic hypothesis

$$\begin{aligned}g_k &= \max_{1\leq j\leq k} (k-j+1)\left(b\chi_j^k - \frac{b^2}{2}\right) \\ (\chi_j^k)^2 &= (\bar{Z}_j^k)^T\, \mathbf{I}^{-1}(\theta_0)\, (\bar{Z}_j^k)\end{aligned}$$

Local composite hypothesis

$$\begin{aligned}t_a &= \min\{k\ : g_k \geq h\} \\ g_k &= \max_{1\leq j\leq k} \sup_{\theta} S_j^k(\theta_0, \theta) \\ S_j^k(\theta_0, \theta) &\approx \frac{k-j+1}{2}(\chi_j^k)^2\end{aligned}$$

AR Model

CUSUM algorithm

$$s_i = \frac{1}{2}\ln\frac{\sigma_0^2}{\sigma_1^2} + \frac{(\varepsilon_i^0)^2}{2\sigma_0^2} - \frac{(\varepsilon_i^1)^2}{2\sigma_1^2}$$

Divergence algorithm

$$\tilde{s}_i = -\frac{\varepsilon_i^0 \varepsilon_i^1}{\sigma_1^2} + \frac{1}{2}\left(\frac{\sigma_0^2}{\sigma_1^2} + 1\right)\frac{(\varepsilon_i^0)^2}{\sigma_0^2} + \frac{1}{2}\left(\frac{\sigma_0^2}{\sigma_1^2} - 1\right) - \nu$$

Squared innovations

$$\eta_i = \frac{\varepsilon_i^2}{\sigma_0^2} - 1$$

Local linear CUSUM algorithm

$$\begin{aligned} s_k &= \Upsilon^T Z_k^* \\ Z_k &= \begin{pmatrix} \frac{1}{\sigma^2}\breve{\mathcal{Y}}_{k-p}^{k-1}\varepsilon_k \\ \\ \frac{1}{\sigma}\left(\frac{\varepsilon_k^2}{\sigma^2} - 1\right) \end{pmatrix} \\ \varepsilon_k &= A^T \breve{\mathcal{Y}}_{k-p}^{k} \end{aligned}$$

Local CUSUM for composite hypothesis

$$\begin{aligned} (\chi_j^k)^2 &= (\bar{Z}_j^k)^T\, \mathbf{I}^{-1}(\theta_0)\, (\bar{Z}_j^k) \\ &= \frac{1}{(k-j+1)^2}\ (\mathcal{Z}_j^k)^T\, \mathbf{I}^{-1}(\theta_0)\, (\mathcal{Z}_j^k) \\ &= \frac{1}{(k-j+1)^2}\left\{\sigma^2\ (\tilde{\mathcal{Z}}_j^k)^T\, \mathbf{T}_p^{-1}(\theta_0)\, (\tilde{\mathcal{Z}}_j^k) + \frac{1}{2}\left[\sum_{i=j}^{k}\left(\frac{\varepsilon_i^2}{\sigma^2} - 1\right)\right]^2\right\} \\ \tilde{\mathcal{Z}}_j^k &= \sum_{i=j}^{k} \frac{1}{\sigma^2}\ \breve{\mathcal{Y}}_{i-p}^{i-1}\ \varepsilon_i \end{aligned}$$

Non-Likelihood-Based Algorithm

$$\left\{\begin{array}{rcl} \mathbf{P}(X_k \in B | X_{k-1}, X_{k-2}, \ldots) &=& \int_B \pi_{\theta^\dagger}(X_{k-1}, dx) \\ y_k &=& f(X_k) \end{array}\right.$$

Let

$$\breve{y}_k^* = \mathcal{K}(\theta^*, \mathcal{Y}_1^k)$$

where $\mathcal{K}$ is used in

$$\theta_k = \theta_{k-1} + \Delta\, \mathcal{K}(\theta_{k-1}, \mathcal{Y}_1^k)$$

Then

$$\begin{array}{lllllll} \text{under} & p_{\theta^*} & : & \breve{\Sigma}_N^{-\frac{1}{2}}(\theta^*)\ \breve{S}_{N,M}(\theta^*) & \rightsquigarrow & (W_t)_t & \text{when} \quad N \to \infty \\ \text{under} & p_{\theta^* + \frac{\nu}{\sqrt{N}}\Upsilon} & : & \breve{\Sigma}_N^{-\frac{1}{2}}(\theta^*)\ \left[\breve{S}_{N,M}(\theta^*) + \nu\ \dot{\kappa}(\theta^*)\ \Upsilon\ t\right] & \rightsquigarrow & (W_t)_t & \text{when} \quad N \to \infty \end{array}$$

where

$$\begin{aligned}
\breve{S}_{N,M}(\theta^*) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{M} \breve{y}_i^* \\
\kappa(\theta, \theta^*) &= \lim_{k\to\infty} \mathbf{E}_\theta(\breve{y}_k^*) = \lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} \breve{y}_i^* \\
\dot{\kappa}(\theta^*) &= \frac{\partial}{\partial\theta} \kappa(\theta, \theta^\dagger)\Big|_{\theta=\theta^*, \theta^\dagger=\theta^*} \\
\breve{\Sigma}_N(\theta^*) &= \sum_{i=-N}^{N} \mathrm{cov}_{\theta^*}(\breve{y}_i^*, \breve{y}_1^*)
\end{aligned}$$

Thus, for detecting a change in the parameter θ of the initial model, apply to this Gaussian problem any of the algorithms for solving the basic problem of chapter 7.

Off-line Algorithms

$$(\hat{t}_0, \hat{\theta}_0, \hat{\theta}_1) = \arg \max_{1+\imath \le k \le N-\imath} \sup_{\theta_1} \sup_{\theta_0} \left[\ln p_{\theta_0}(\mathcal{Y}_1^{k-1}) + \ln p_{\theta_1}(\mathcal{Y}_k^N | \mathcal{Y}_1^{k-1}) \right]$$

$$\tfrac{t_0(N)[N-t_0(N)]}{N} \; [\theta_1(N) - \theta_0(N)]^T \mathbf{I}(\theta_0) [\theta_1(N) - \theta_0(N)] \; \left[\tfrac{\hat{t}_0 - t_0(N)}{N} \right]$$

$$\rightsquigarrow \arg\sup_{t\in\mathbf{R}} \left\{ W_t - \tfrac{|t|}{2} \right\}$$

9

Nonadditive Changes - Multidimensional Signals

In this chapter, we investigate the problem of detecting *nonadditive changes in multidimensional signals*. First we emphasize that, for the general model of conditional probability distributions, all that is explained in chapter 8 is also valid in the case of multidimensional signals, and therefore we do *not* consider this model anymore here. Nonadditive changes are investigated in the four following models :

1. AR models;
2. ARMA models;
3. state-space models;
4. nonlinear ARMA models.

Note, however, that, as we explained in section 6.2, conditional distributions comprise the most general statistical model and encompass these four models. The central issue of detecting nonadditive changes in this model was addressed in chapter 8, using the likelihood ratio methodology and the local approach. Solutions to change detection problems for the AR and ARMA models are obtained as particular cases of the general statistical model, and the nonlinear case is treated separately. What is new in this chapter is the use of state-space models.

This chapter is mainly devoted to the extension to multidimensional signals of the main ideas developed in chapter 8 for designing *on-line* nonadditive change detection algorithms. However, we discuss some off-line algorithms in section 9.3.

The main **goals** of this chapter are as follows. First we extend the GLR and CUSUM algorithms and the use of the local approach to the detection of nonadditive changes in AR/ARMA models, starting from the *general case* of conditional distributions investigated in chapter 8. The second goal is to investigate one problem related to the detection of nonadditive changes in state-space models as defined in section 6.1. This problem, which is discussed in section 9.3, is related to the important issue of vibration monitoring of mechanical structures and rotating machines presented in example 1.2.5 of chapter 1. It is also equivalent to the problem of detecting changes in the AR part of an ARMA model with nonstationary MA part, and it is one example of use of the general non-likelihood-based methodology described in section 8.4. The third goal is to clarify the detectability issue, which is done in section 9.4. The last goal is to give the available theoretical results concerning the properties of the algorithms presented in *both* chapters 8 and 9.

The **tools** for reaching these goals can be summarized as follows. As we said before, the basic tools that are necessary for nonadditive change detection in the above-mentioned models, namely the likelihood ratio for conditional densities, use of local approach, and non-likelihood-based algorithms, were presented in detail in chapter 8. The new tools used here concern basically the detection of changes in state-space models.

For the problem investigated in section 9.3, two key additional tools are the following. First, the equivalence between ARMA and state-space models, which was described in subsection 3.2.4, is applied to the general non-likelihood-based algorithm, giving rise to what we call instrumental statistics. Second, we investigate the diagnosis problem and introduce the solution that is associated with this general non-likelihood approach. Finally, we address the detectability issue using basically the concept of Kullback information discussed in section 6.3 and chapter 8, and in the case of state-space models discussed in section 9.3, we establish a link between this statistical detectability criterion and a geometric condition as in chapter 7.

9.1 Introducing the Tools

In this section, we first introduce nonadditive changes in the four types of models mentioned before, namely AR and ARMA models, state-space models, and nonlinear ARMA models. We describe the basic tools in subsection 9.1.2. The key concepts to be used for solving the corresponding detection problems and discussing the detectability issue in section 9.4, namely sufficient statistics and the local approach, were introduced in subsection 8.1.2; therefore, they are only briefly summarized here. Then we discuss the issues and introduce the tools that are new in the case of multidimensional signals.

9.1.1 Nonadditive Changes

In this chapter, we consider sequences of *multidimensional* observations $(Y_k)_k$ with dimension r, which we represent using an ARMAX model as in (4.1.92) or a nonlinear ARMA model as in (6.2.11). In this chapter, as in chapter 8, we investigate nonadditive or spectral changes, which are changes in the variance, correlations, spectral characteristics, or dynamics of the signal or system.

We consider the four following models :

- **AR models** :

$$Y_k = \sum_{i=1}^{p} A_i Y_{k-i} + V_k \tag{9.1.1}$$

where $(V_k)_k$ is a Gaussian white noise sequence with covariance matrix R. The conditional probability distribution of such a sequence of observations $(Y_k)_k$ is denoted by $p_\theta(Y_k|\mathcal{Y}_1^{k-1})$, where θ is the matrix containing the AR coefficients and the covariance matrix R. The problem is to detect changes in the parameter θ, from θ_0 to θ_1, where

$$\theta_l^T = \begin{pmatrix} A_1^l & \dots & A_p^l & R_l \end{pmatrix}, \quad l = 0, 1 \tag{9.1.2}$$

- **ARMA models** :

$$Y_k = \sum_{i=1}^{p} A_i Y_{k-i} + \sum_{j=0}^{q} B_j V_{k-j} \tag{9.1.3}$$

where $(V_k)_k$ is again a Gaussian white noise sequence with covariance matrix R, and $B_0 = I_r$. The conditional probability distribution of such a sequence of observations $(Y_k)_k$ is denoted by $p_\theta(Y_k|\mathcal{Y}_1^{k-1})$, where θ is the matrix containing the AR and MA coefficients and the covariance matrix R. The problem is to detect changes in the parameter θ, from θ_0 to θ_1, where

$$\theta_l^T = \begin{pmatrix} A_1^l & \dots & A_p^l & B_1^l & \dots & B_q^l & R_l \end{pmatrix}, \quad l = 0, 1 \tag{9.1.4}$$

Changes in these two models are of interest in several types of signals, such as seismic data, biomedical signals, acoustic signals, vibration measurements.

- **State-space models** :

$$\left\{ \begin{array}{rcl} X_{k+1} & = & FX_k + GU_k + W_k \\ Y_k & = & HX_k + JU_k + V_k \end{array} \right. \tag{9.1.5}$$

where the state X, the measurement Y, and the control U have dimensions n, r, and m, respectively, and where $(W_k)_k$ and $(V_k)_k$ are two independent Gaussian white noises, with covariance matrices Q and R, respectively. The parameter θ here is made of the pair (H, F) and the covariance matrices. The problem is to detect changes in the parameter θ, from θ_0 to θ_1, where

$$\theta_l^T = \left(\begin{array}{cccc} H^l & F^l & Q_l & R_l \end{array} \right), \quad l = 0, 1 \tag{9.1.6}$$

As mentioned in chapter 6, in the present framework of multidimensional signal processing, there exist particular problems of parameterization of state-space models. In subsection 6.2.3, we discussed several nonadditive changes in state-space models, corresponding to only one possible choice of parameterization. Moreover, the key difficulty here is that it can happen that several *different* parameter vectors θ give rise to the *same* likelihood of a given sample of observations. We discuss this issue further next, when we discuss the new issues related to the multidimensional framework.
We mainly investigate one specific example of nonadditive changes, with no change in the covariance matrices Q and R. This change is concerned with modifications in the observed components of the eigenstructure of F, and is related to the vibration monitoring problem. It is also equivalent to the problem of detecting changes in the AR part of an ARMA model with a nonstationary MA part.

- **Nonlinear ARMA models** :

$$\left\{ \begin{array}{rcl} \mathbf{P}(X_k \in B | X_{k-1}, X_{k-2}, \ldots) & = & \int_B \pi_\theta(X_{k-1}, dx) \\ Y_k & = & f(X_k) \end{array} \right. \tag{9.1.7}$$

where $\pi_\theta(X, dx)$ is the transition probability of the Markov chain $(X_k)_k$ and where f is a nonlinear function. The problem is to detect changes in the parameter θ of the transition probability π_θ. This problem statement and the corresponding solution presented in section 8.4 are of interest for solving the vibration monitoring problem, which we discuss in section 9.3.

In this chapter, as in chapter 8, we use the first method of generating changes, which we described in section 8.1. We refer the reader to chapter 8 for a discussion of the effect of the first and second methods on the design of the change detection algorithm for several models.

Finally, the parameter θ, defined as a *matrix* in (9.1.2), (9.1.4), and (9.1.6), is sometimes more conveniently treated as the ℓ-dimensional *vector* Θ obtained by stacking the columns of θ on top of each other, which we note as

$$\Theta = \mathrm{col}(\theta) \tag{9.1.8}$$

However, we continue to use the notation θ when this distinction is not necessary.

9.1.2 Three Basic Detection Tools

We now briefly summarize the key concepts to be used for solving these change detection problems, namely the likelihood ratio, local approach, and non-likelihood-based algorithms. Then we discuss the issues that are new in the case of multidimensional signals, namely parameterization and identifiability.

9.1.2.1 Likelihood Ratio

The log-likelihood ratio is a sufficient statistic. For the observations $\mathcal{Y}_1^k$, it can be written as

$$\begin{aligned} S_1^k &= \sum_{i=1}^{k} s_i \\ s_i &= \ln \frac{p_{\theta_1}(Y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(Y_i|\mathcal{Y}_1^{i-1})} \end{aligned} \tag{9.1.9}$$

where $p_\theta(Y_1|\mathcal{Y}_1^0) = p_\theta(Y_1)$, under the assumption of a change generated by the first method (8.2.2).

Because of the above-mentioned parameterization problem, in the multidimensional case, a qualitatively new situation arises when there exist several parameters that result in the same likelihood of a given sample of observations. Theoretically, this situation can arise for scalar signals, but it is much more crucial for multidimensional signals. Note, however, that a change between two such parameterizations (which is, of course, of no practical interest and not detectable with the Kullback information-based detectability definition) is not detected by the likelihood ratio.

Example 9.1.1 (ARMA case). *In the case of ARMA models, we use the following notation :*

$$A^T = \begin{pmatrix} I_r & -A_1 & \dots & -A_p \end{pmatrix} \tag{9.1.10}$$

$$B^T = \begin{pmatrix} B_1 & \dots & B_q \end{pmatrix} \tag{9.1.11}$$

for the sets of AR and MA parameters, and

$$(\check{\mathcal{Y}}_{k-p}^k)^T = \begin{pmatrix} Y_k^T & Y_{k-1}^T & \dots & Y_{k-p}^T \end{pmatrix} \tag{9.1.12}$$

$$(\check{\mathcal{E}}_{k-q}^{k-1})^T = \begin{pmatrix} \varepsilon_{k-1}^T & \varepsilon_{k-2}^T & \dots & \varepsilon_{k-q}^T \end{pmatrix} \tag{9.1.13}$$

for the sets of past observations and innovations in backward order. In this case, the conditional probability distribution of the observation Y_k is given by

$$\begin{aligned} p_\theta(Y_k|\mathcal{Y}_1^{k-1}) &= \frac{1}{\sqrt{(2\pi)^r(\det R)}} e^{-\frac{1}{2}(A^T\check{\mathcal{Y}}_{k-p}^k - B^T\check{\mathcal{E}}_{k-q}^{k-1})^T R^{-1}(A^T\check{\mathcal{Y}}_{k-p}^k - B^T\check{\mathcal{E}}_{k-q}^{k-1})} \\ &= \frac{1}{\sqrt{(2\pi)^r(\det R)}} e^{-\frac{1}{2}\varepsilon_k^T R^{-1}\varepsilon_k} \end{aligned} \tag{9.1.14}$$

Thus, the increment of the log-likelihood ratio is

$$s_k = \frac{1}{2}\ln\frac{\det R_0}{\det R_1} + \frac{1}{2}(\varepsilon_k^0)^T R_0^{-1}\varepsilon_k^0 - \frac{1}{2}(\varepsilon_k^1)^T R_1^{-1}\varepsilon_k^1 \tag{9.1.15}$$

which reduces to

$$s_k = \frac{1}{2}\left[(\varepsilon_k^0)^T R^{-1}\varepsilon_k^0 - (\varepsilon_k^1)^T R^{-1}\varepsilon_k^1\right] \tag{9.1.16}$$

when the input covariance matrix does not change.

9.1.2.2 Local Approach

This approach is based upon the efficient score :

$$Z_i^* = \left. \frac{\partial \ln p_\theta(Y_i|\mathcal{Y}_1^{i-1})}{\partial \theta} \right|_{\theta=\theta^*} \tag{9.1.17}$$

For the local approach, the key difficulty is that, in case of overparameterization, the Fisher information matrix, which is by definition the covariance matrix of the efficient score, can be degenerated, and then the central limit theorem (9.1.19) does not apply and the local asymptotic approach cannot be used. An example of computation of the efficient score in the present multidimensional case is given in subsection 9.2.2.

When the Fisher information matrix is invertible, the following central limit theorem states that a small change in the parameter θ is reflected into a change in the mean of the normalized cumulative sum of efficient scores :

$$S_{N,t}(\theta^*) = \frac{1}{\sqrt{N}} \sum_{i=1}^{[Nt]} Z_i^* \tag{9.1.18}$$

defined for $t \in [0,1]$, and where $[Nt]$ is the integer part of Nt. Let us recall the following result stated in chapter 8. When $N \to \infty$,

$$\begin{array}{lllll} \text{under} \quad p_{\theta^*} \ : & \mathbf{I}_N^{-\frac{1}{2}}(\theta^*) \ S_{N,t}(\theta^*) & \rightsquigarrow & (W_t)_{t\in[0,1]} \\ \text{under} \quad p_{\theta^*+\frac{\nu}{\sqrt{N}}\Upsilon} \ : & \mathbf{I}_N^{-\frac{1}{2}}(\theta^*) \ [S_{N,t}(\theta^*) - \nu \, \mathbf{I}_N(\theta^*) \, \Upsilon \, t] & \rightsquigarrow & (W_t)_{t\in[0,1]} \end{array} \tag{9.1.19}$$

where $\mathbf{I}(\theta^*)$ is the Fisher information matrix, and where $(W_t)_{t\in[0,1]}$ is an ℓ-dimensional normalized Brownian motion. As in chapter 8, this result means that for detecting small deviations with respect to a reference model parameter θ^*, one can work with the process $(Z_k^*)_k$ as if it was an independent Gaussian sequence, with mean zero before change and $\nu \, \mathbf{I}(\theta^*)\Upsilon$ after change, and with covariance matrix $\mathbf{I}(\theta^*)$.

9.1.2.3 Non-Likelihood-Based Algorithms

These change detection algorithms concern the nonlinear ARMA models described before. They are based upon parameter estimation algorithms, which can be written in the following way :

$$\theta_k = \theta_{k-1} + \Delta \, \mathcal{K}(\theta_{k-1}, \mathcal{Y}_1^k) \tag{9.1.20}$$

where $\mathcal{K}$ satisfies some regularity conditions in order that this algorithm converges toward the true value of the parameter θ. Recall that the notation $\mathcal{K}(\theta, \mathcal{Y}_1^k)$ stands for a measurable random vector where the dependence on $\mathcal{Y}_1^k$ is only through a function of the observations, which is a semi-Markov process.

As we explained in chapter 8, for solving change detection problems in those models, an alternative solution to the complex log-likelihood ratio consists of basing the decision only upon the information contained in the statistic :

$$\breve{Y}_k^* = \mathcal{K}(\theta^*, \mathcal{Y}_1^k) \tag{9.1.21}$$

and in the following cumulative sum :

$$\breve{S}_{N,t}(\theta^*) = \frac{1}{\sqrt{N}} \sum_{i=1}^{[Nt]} \breve{Y}_i^* \tag{9.1.22}$$

defined for $t \in [0,1]$. Let us recall the following central limit theorem stated in chapter 8 [Benveniste *et al.*, 1987, Benveniste *et al.*, 1990, Ladelli, 1990]. When $N \to \infty$,

$$\begin{array}{lllll} \text{under} \quad p_{\theta^*} \ : & \breve{\Sigma}_N^{-\frac{1}{2}}(\theta^*) \ \breve{S}_{N,t}(\theta^*) & \rightsquigarrow & (W_t)_{t\in[0,1]} \\ \text{under} \quad p_{\theta^*+\frac{\nu}{\sqrt{N}}\Upsilon} \ : & \breve{\Sigma}_N^{-\frac{1}{2}}(\theta^*) \ \left[\breve{S}_{N,t}(\theta^*) + \nu \ \dot{\kappa}(\theta^*) \ \Upsilon \ t\right] & \rightsquigarrow & (W_t)_{t\in[0,1]} \end{array} \tag{9.1.23}$$

where

$$\begin{aligned} \kappa(\theta,\theta^*) &= \lim_{k\to\infty} \mathbf{E}_\theta[\mathcal{K}(\theta^*, \mathcal{Y}_1^k)] \\ \dot{\kappa}(\theta^*) &= \frac{\partial}{\partial\theta}\kappa(\theta,\theta^\dagger)\Big|_{\theta=\theta^*,\theta^\dagger=\theta^*} \end{aligned} \tag{9.1.24}$$

and where the covariance matrix is given by

$$\breve{\Sigma}_N(\theta^*) = \sum_{i=-N}^{N} \mathrm{cov}_{\theta^*}(\breve{Y}_i^*, \breve{Y}_1^*) \tag{9.1.25}$$

As before, this result means that for detecting small deviations with respect to a reference model parameter θ^*, one can work with the process $(\breve{Y}_k^*)_k$ as if it were an independent Gaussian sequence, with mean zero before change and $-\nu \ \dot{\kappa}(\theta^*)\Upsilon$ after change, and with covariance matrix $\breve{\Sigma}(\theta^*)$.

9.1.2.4 New Multidimensional Issues

These three basic detection tools are the same as those introduced in chapter 8 for the detection of nonadditive changes in scalar signals. Processing multidimensional signals can be a nontrivial extension of the processing of scalar signals, because some difficult issues turn out to be greatly magnified in the multidimensional case. Let us thus emphasize the new issues that arise for the detection tools when applied to multidimensional signals.

The first issue is concerned with the key problems of parameterization and parameter identifiability of models for multivariable systems. One of these problems is related to the fact that the pair (H, F) in a state-space representation (9.1.5) inferred from data is defined up to a multiplication by a matrix of change in the basis of state coordinates. Another crucial problem in our general likelihood framework is the fact that two different parameterizations can lead to the same likelihood function. We investigate two particular points related to this question. The first is examined in subsection 9.2.2, where we compute the efficient score in a particular multidimensional model. The second is discussed in section 9.3, where the proposed non-likelihood-based algorithm is shown to be able to detect any change in the minimal representation of the state-space model. We do not discuss further these issues here. The reader is referred to the extensive literature on the subject, for example [Hannan and Deistler, 1988, Ljung, 1987, Caines, 1988, Söderström and Stoïca, 1989].

The second issue is concerned with the complexity of the likelihood ratio approach, which is mentioned in chapter 8 in the case of nonadditive changes in scalar signals, and is further increased in the case of multidimensional signals. Thus, interest in simplifying the general likelihood ratio approach is even greater than in the case of scalar signals.

9.1.3 Diagnosis

We complete our introduction to the tools used in this chapter with a discussion of the diagnosis problem, which is typical for both vector parameters and multidimensional signals. Let us note first that the diagnosis

problem in the case of nonadditive changes is even more difficult than in the case of additive changes discussed in chapter 7. Actually, the dynamics of the system (9.1.5) introduce a coupling effect between additive changes : A failure in only one actuator can be reflected in changes in several sensor signals. Therefore, different types of changes in the dynamics itself, such as changes in the eigenvalues and eigenvectors of the state transition matrix F, can be difficult to discriminate by processing the sensor signals.

In this chapter, we investigate the diagnosis problem in the framework of section 9.3 devoted to changes in the eigenstructure of a state-space model with nonstationary and nonmeasured state noise, or equivalently in the AR part of a multivariable ARMA model with a nonstationary MA part. But the solution we describe is also valid in the more general case of nonlinear ARMA models, as discussed in [Benveniste *et al.*, 1987]. We first consider the diagnosis in terms of the parameters of the (small) black-box ARMA model which is used for monitoring. For this problem, we could consider two types of solutions. The first would be the *minmax robust* approach, introduced in subsection 7.2.5 for *additive* changes. The main limitation of this approach is that its implementation requires the number of sensors to be greater than or equal to the sum of the dimensions of the changes to be discriminated. In the case of nonadditive changes, the situation is even worse because of the nonlinearities. Thus, this approach is of limited interest in the present case, where typically the number of frequencies or eigenvectors to be monitored is greater than the number of sensors. Therefore, we describe here only the second possible solution, based upon what we call a *sensitivity* technique which has the advantage of being less computationally complex than the previous solution, while keeping reasonable although suboptimal properties. This sensitivity approach turns out to be of key interest for solving the second diagnosis problem, which we investigate in section 9.3. This problem is concerned with the diagnosis of changes in the AR parameters - or equivalently in the observed components of the eigenstructure - in terms of changes in the mass and stiffness parameters M and K of the (huge) model of the underlying mechanical system (see chapter 11). This second diagnosis problem is the most relevant. Note that we do *not* require the existence of a bijective map between the two model parameters sets, namely AR parameters and coefficients of the mechanical system, nor the identifiability of the mechanical model.

9.2 AR/ARMA Models and the Likelihood Ratio

In this section, we investigate the design of nonadditive change detection algorithms for multidimensional signals in the case of AR and ARMA models, considering both simple and composite hypotheses. In view of the discussions in chapter 8, the main tools to be used in this chapter are the CUSUM, divergence, and GLR algorithms, and the algorithms that result from the use of the local approach in the case of an unknown parameter after change.

The differences in the algorithms, which result from the first two ways of generating nonadditive changes explained in chapter 8, are outlined in that chapter and are no longer addressed in this section. From now on, we concentrate on the first method of generating a nonadditive change, and thus on the formula (8.1.8) for the probability densities of a sample of observations, namely

$$p(\mathcal{Y}_1^k|k \geq t_0) = p_{\theta_0}(Y_1) \left[\prod_{i=2}^{t_0-1} p_{\theta_0}(Y_i|\mathcal{Y}_1^{i-1}) \right] \left[\prod_{i=t_0}^{k} p_{\theta_1}(Y_i|\mathcal{Y}_1^{i-1}) \right] \tag{9.2.1}$$

9.2.1 Simple Hypotheses

As we explained in the chapters 2, 7, and 8, in the case of known parameters before and after change, the relevant change detection algorithm is the CUSUM algorithm. Therefore, we recall this algorithm in the case of changes in the parameter vector of a multivariable AR or ARMA model. We also give the

multidimensional counterpart of the divergence algorithm, which has proven useful in the particular case of scalar AR models.

9.2.1.1 The CUSUM Algorithm

As shown in (8.3.1)-(8.3.2), the CUSUM algorithm can be written as

$$\begin{aligned} t_a &= \min\{k : g_k \geq h\} \\ g_k &= \left(S^k_{k-N_k+1}\right)^+ = (g_{k-1} + s_k)^+ \qquad (9.2.2) \\ S^k_j &= \sum_{i=j}^{k} \ln \frac{p_{\theta_1}(Y_i|\mathcal{Y}_1^{i-1})}{p_{\theta_0}(Y_i|\mathcal{Y}_1^{i-1})} \\ &= \sum_{i=j}^{k} s_i \qquad (9.2.3) \end{aligned}$$

where

$$s_i = \frac{1}{2}\ln\frac{\det R_0}{\det R_1} + \frac{1}{2}(\varepsilon_i^0)^T R_0^{-1}\varepsilon_i^0 - \frac{1}{2}(\varepsilon_i^1)^T R_1^{-1}\varepsilon_i^1 \qquad (9.2.4)$$

and where $N_k = N_{k-1}\mathbf{1}_{\{g_{k-1}>0\}} + 1$, namely where N_k is the number of observations since the last vanishing of g_k, and under the assumption of a change generated by the first method.

The unconditional expectations of s_i before and after change are, respectively,

$$\mathbf{E}_{\theta_0}(s_i) = \frac{1}{2} + \frac{1}{2}\ln\frac{\det R_0}{\det R_1} - \frac{1}{2}\frac{\det R_0}{\det R_1}\left[1 + \frac{1}{2\pi}\int_{-\pi}^{\pi} \|A_0^{-1}(e^{i\omega})A_1(e^{i\omega})\|^2 d\omega\right] \qquad (9.2.5)$$

$$\mathbf{E}_{\theta_1}(s_i) = -\frac{1}{2} - \frac{1}{2}\ln\frac{\det R_1}{\det R_0} + \frac{1}{2}\frac{\det R_1}{\det R_0}\left[1 + \frac{1}{2\pi}\int_{-\pi}^{\pi} \|A_1^{-1}(e^{i\omega})A_0(e^{i\omega})\|^2 d\omega\right] \qquad (9.2.6)$$

where

$$A_l(z) = I_r - \sum_{i=1}^{p} A_i^l z^{-i} \qquad (9.2.7)$$

for $l = 0, 1$.

9.2.1.2 The Divergence Algorithm

The divergence algorithm, which was originally introduced in the AR case with this motivation of symmetry, is based upon the decision function (8.3.13) :

$$\begin{aligned} t_a &= \min\{k : \tilde{g}_k \geq h\} \\ \tilde{g}_k &= \left(\tilde{S}^k_{k-\tilde{N}_k+1}\right)^+ = (\tilde{g}_{k-1} + \tilde{s}_k)^+ \qquad (9.2.8) \\ \tilde{S}^k_j &= \sum_{i=j}^{k} \tilde{s}_i \end{aligned}$$

where

$$\begin{aligned} \tilde{s}_i &= \ln\frac{p_{\theta_1}(Y_i|\mathcal{Y}_{i-p}^{i-1})}{p_{\theta_0}(Y_i|\mathcal{Y}_{i-p}^{i-1})} - \mathbf{E}_{\theta_0}\left[\ln\frac{p_{\theta_1}(Y_i|\mathcal{Y}_{i-p}^{i-1})}{p_{\theta_0}(Y_i|\mathcal{Y}_{i-p}^{i-1})}\middle|\mathcal{Y}_{i-p}^{i-1}\right] - \nu \\ &= s_i - \frac{1}{2}\ln\frac{\det R_0}{\det R_1} + \frac{1}{2} - \frac{1}{2}\mathcal{I}(A_0^T\check{\mathcal{Y}}_{i-p}^i, A_1^T\check{\mathcal{Y}}_{i-p}^i) - \nu \qquad (9.2.9) \end{aligned}$$

and where

$$\mathcal{I}(\alpha,\beta) = \int \frac{1}{\sqrt{2\pi}(\det R_0)}\, e^{-\frac{1}{2}(y-\alpha)^T R_0^{-1}(y-\alpha)}\ (y-\beta)^T R_1^{-1}(y-\beta)\, dy \tag{9.2.10}$$

can be shown to be

$$\mathcal{I}(\alpha,\beta) = \mathrm{tr}(R_1^{-1}R_0) + (\beta-\alpha)^T R_1^{-1}(\beta-\alpha) \tag{9.2.11}$$

after straightforward but long computations. Finally,

$$\tilde{s}_i = -(\varepsilon_i^0)^T R_1^{-1}\varepsilon_i^0 + \frac{1}{2}(\varepsilon_i^0)^T\left(R_0^{-1}+R_1^{-1}\right)\varepsilon_i^0 + \frac{1}{2}\,\mathrm{tr}(R_1^{-1}R_0) - \frac{r}{2} - \nu \tag{9.2.12}$$

Again, the main difference with respect to the CUSUM algorithm (9.2.4) lies in the function of the two residuals, which is monitored.

In formula (9.2.8), $\tilde{N}_k = \tilde{N}_{k-1}\mathbf{1}_{\{\tilde{g}_{k-1}>0\}} + 1$, namely $\tilde{N}_k$ is the number of observations since the last vanishing of $\tilde{g}_k$. Note that here we again assume the first method of generating changes. Note also that the increment of the cumulative sum is the same here as in (8.2.18), except that we have subtracted the mean value of the increment of the CUSUM algorithm before change, and also a constant quantity ν chosen such that $\mathbf{E}_{\theta_0}(\tilde{s}_i) < 0$ and $\mathbf{E}_{\theta_1}(\tilde{s}_i) > 0$. In practice, ν is thought of as being a kind of minimum magnitude of spectral change to be detected. The choice of this constant in practice was discussed in section 8.6 and is addressed again in chapter 10.

9.2.2 Composite Hypotheses

The problem of detecting nonadditive changes in the case of composite hypotheses for a multidimensional ARMA model is a particular case of the general problem addressed in chapter 8 for conditional distributions. But, because of the above-mentioned parameterization problems, the actual computation of the formulas given there is quite complex, and it is thus of interest to outline, in two particular examples, how they can be effectively achieved. Therefore, we now investigate the two following questions : first, the design of a linear CUSUM algorithm for detecting a change in the covariance matrix, and second, the design of the linear CUSUM algorithm for the detection of a change in the parameter θ of an AR(1) process.

9.2.2.1 The Linear CUSUM Algorithm for a Change in the Covariance Matrix

We consider an ARMA process where the *only* unknown parameters are the covariance matrix R of the input excitation. It results from (9.1.14) that the density of such an ARMA model has an exponential structure with respect to the elements of the matrix $\check{R} = R^{-1}$. Let us compute the *matrix* of the efficient score :

$$Z_k^* = \left.\frac{\partial \ln p_{\check{R}}(Y_k|\mathcal{Y}_1^{k-1})}{\partial \check{R}}\right|_{\check{R}=\check{R}^*} = \frac{1}{2}R^* - \frac{1}{2}\,\varepsilon_k^*(\varepsilon_k^*)^T \tag{9.2.13}$$

The increment of the decision function of the linear CUSUM algorithm is thus

$$s_k = \frac{1}{2}\left\{\mathrm{tr}(R^*\,\Upsilon) - \mathrm{tr}\left[\varepsilon_k^*(\varepsilon_k^*)^T\,\Upsilon\right]\right\} \tag{9.2.14}$$

where Υ is here the "unit" $r \times r$ *matrix* of direction of changes in $\check{R}$, such that

$$\sum_{i,j=1}^{r} |\Upsilon_{i,j}|^2 = 1 \tag{9.2.15}$$

9.2.2.2 The Linear CUSUM Algorithm for a Change in an AR(1) Process

We now investigate the problem of detecting a change in the matrix parameter $\theta = (A, \check{R})$, where A is the matrix autoregressive coefficient. In this case, the efficient score with respect to $\check{R}$ is the same as before. The efficient score with respect to A is

$$\left.\frac{\partial \ln p_{A,\check{R}}(Y_k|\mathcal{Y}_1^{k-1})}{\partial A}\right|_{A=A^*,\check{R}=\check{R}^*} = \check{R}^* \, \varepsilon_k^* \, Y_{k-1}^T \tag{9.2.16}$$

The increment of the decision function of the local linear CUSUM algorithm is thus

$$s_k = \frac{1}{2}\left\{\mathrm{tr}(R^* \, \Upsilon_{\check{R}}) - \mathrm{tr}\left[\varepsilon_k^*(\varepsilon_k^*)^T \, \Upsilon_{\check{R}}\right]\right\} + \mathrm{tr}(\check{R}^* \, \varepsilon_k^* \, Y_{k-1}^T \, \Upsilon_A) \tag{9.2.17}$$

where Υ_A and $\Upsilon_{\check{R}}$ are the "unit" $r \times r$ matrices of direction of changes in A and $\check{R}$, respectively.

Let us compare this increment with the corresponding increment in the scalar AR(1) model, namely with

$$s_k = \frac{1}{2}\left[\sigma^{*2} \, \Upsilon_{\sigma^{-2}} - (\varepsilon_k^*)^2 \, \Upsilon_{\sigma^{-2}}\right] + (\sigma^*)^{-2} \, \varepsilon_k^* \, y_{k-1} \, \Upsilon_a \tag{9.2.18}$$

It is obvious that there exists a quite natural relation between these two expressions.

9.3 Detection and Diagnosis of Changes in the Eigenstructure

In this section, we discuss the solution to the vibration monitoring problem, which we described in example 1.2.5 in chapter 1. As far as possible, we introduce the solution to this problem while making an abstraction of the underlying application and keeping only the generic problem. Therefore, we consider the problem of detecting and diagnosing changes in the eigenstructure of the state transition matrix of a state-space model having nonstationary state noise, or equivalently in the AR part of a multivariable ARMA model having a nonstationary and unknown MA part. The formal reasons for which the vibration monitoring problem for mechanical systems can be stated in this manner are given in section 11.1. In this section, we first investigate the detection issue, and then give a possible solution to two diagnosis problems.

Before proceeding, let us comment upon the choice of the model set to be used for this type of monitoring problem. Even though the pieces of information that serve as a guideline and reference for monitoring are in the frequency domain, it is possible and often preferable to make decisions in the parametric domain, namely the space of multivariable AR coefficients, for identification as well as for both detection and diagnosis. This allows us, for example, to detect changes in the vibrating characteristics even when these changes do *not* affect the eigenfrequencies and thus affect only the geometry of the eigenvectors. Of course, this could not be achieved using power spectral densities, for example.

9.3.1 Instrumental Statistics and Detection

As we mentioned in chapter 8, because of the tight coupling between the AR and MA parts - the Fisher information matrix is not block diagonal and the efficient score (4.1.106) does depend upon the moving average part - the log-likelihood ratio approach, with or without the local approach, is not feasible for detecting changes in the AR part of a multivariable ARMA model. Thus, we use instead the non-likelihood approach, described in chapter 8 and section 9.1, for designing the detection algorithm. More precisely we use the multidimensional counterpart of example 8.4.6, which provides us with a detection algorithm associated with the instrumental variables (IV) identification method. Let us explain this now.

9.3.1.1 On-line Detection

The IV algorithm is aimed at the identification of the AR part of an ARMA(p,q) model. As shown in section 11.1, in the present case of vibration monitoring, we have $q = p-1$, basically because we assume no noise on the observation equation. Thus, we summarize the IV identification method for this case in the following formulas :

$$\begin{aligned} \underline{\theta}_k &= \underline{\theta}_{k-1} + \frac{1}{k}\,\Sigma_k^{-1}\,\check{\mathcal{Y}}_{k-2p+1}^{k-p}\,e_k^T(\underline{\theta}_{k-1}) \\ \Sigma_k &= \Sigma_{k-1} + \frac{1}{k}\,[\check{\mathcal{Y}}_{k-2p+1}^{k-p}(\check{\mathcal{Y}}_{k-p}^{k-1})^T - \Sigma_{k-1}] \\ e_k(\underline{\theta}) &= Y_k - \underline{\theta}^T\,\check{\mathcal{Y}}_{k-p}^{k-1} \end{aligned} \tag{9.3.1}$$

where

$$\underline{\theta}^T = \begin{pmatrix} A_1 & \dots & A_p \end{pmatrix} \tag{9.3.2}$$

This identification algorithm has been proven to be efficient, namely to provide us with *consistent* estimates of the AR matrix parameters $(A_i)_{1\le i\le p}$, even in the present nonstationary situation of time-varying MA coefficients [Benveniste and Fuchs, 1985].

Following the last example of section 8.4, let us thus consider what we call the *instrumental statistics* :

$$\check{Y}_k^* = \check{\mathcal{Y}}_{k-2p+1}^{k-p}\,e_k^T(\underline{\theta}^*) \tag{9.3.3}$$

where θ^* is the nominal (assumed) value of θ. Following the notation we introduced at the end of subsection 9.1.1, we rewrite the *matrix* $\check{Y}_k^*$ into the *vector* $\check{\mathbf{Y}}_k^*$ defined as

$$\begin{aligned} \check{\mathbf{Y}}_k^* &= \mathrm{col}(\check{Y}_k^*) \\ &= e_k(\theta^*) \otimes \check{\mathcal{Y}}_{k-2p+1}^{k-p} \end{aligned} \tag{9.3.4}$$

where $\otimes$ denotes the Kronecker product of two matrices [Söderström and Stoïca, 1989]. Note that

$$e_k(\underline{\theta}) = \begin{pmatrix} -\check{\underline{\theta}}^T & I_r \end{pmatrix}\,\mathcal{Y}_{k-p}^k \tag{9.3.5}$$

where $\check{\underline{\theta}}^T = \begin{pmatrix} A_p & \dots & A_1 \end{pmatrix}$. Let $\check{\Sigma}(\underline{\theta}^*)$ and $\dot{\kappa}(\underline{\theta}^*)$ be the asymptotic covariance and derivative of the mean value of $\check{\mathbf{Y}}_k^*$. These quantities are computed in the sequel.

As we stated in chapter 8, the initial change detection problem on the $(Y_k)_k$, namely the problem of testing between the hypotheses

$$\mathbf{H}_0 = \{\underline{\Theta} = \underline{\Theta}^*\} \quad \text{and} \quad \mathbf{H}_1 = \left\{\underline{\Theta} = \underline{\Theta}^* + \frac{\nu}{\sqrt{N}}\Upsilon\right\} \tag{9.3.6}$$

(where $\underline{\Theta}^* = \mathrm{col}(\underline{\theta}^*)$) is transformed into a change in the mean of the process $(\check{Y}_k^*)_k$, which has to be considered as if it was an independent Gaussian sequence, with mean zero before change and $-\nu\,\dot{\kappa}(\underline{\theta}^*)\Upsilon$ after change, and with covariance matrix $\check{\Sigma}(\underline{\theta}^*)$, with $\dot{\kappa}$ and $\check{\Sigma}$ as in (8.4.40) and (8.4.39). The corresponding solution to this problem is any one of the solutions of the basic problem of chapter 7, which were described in subsection 7.2.1, according to the amount of available *a priori* information about the change vector Υ and magnitude ν.

Now, assuming that no *a priori* information is available about ν and Υ, which is often the case in practice, we use the GLR detection algorithm presented in subsection 7.2.1. This results in the following χ^2

test :

$$\begin{aligned} g_k &= \max_{1\le j\le k} \frac{k-j+1}{2}\,(\chi_j^k)^2 \\ (\chi_j^k)^2 &= \frac{1}{(k-j+1)^2}\left(\sum_{i=j}^{k}\breve{\mathbf{Y}}_{\mathbf{i}}^*\right)^T \breve{\Sigma}^{-1}\dot{\kappa}(\dot{\kappa}^T\breve{\Sigma}^{-1}\dot{\kappa})^{-1}\dot{\kappa}^T\breve{\Sigma}^{-1}\left(\sum_{i=j}^{k}\breve{\mathbf{Y}}_{\mathbf{i}}^*\right) \end{aligned} \tag{9.3.7}$$

This gives the *on-line* change detection algorithm.

9.3.1.2 Off-line Detection

We now consider the corresponding hypotheses testing problem. In other words, we consider the problem of deciding whether or not a fixed size sample of data corresponds to a nominal (assumed) value θ^* of the parameter. We also call this problem *off-line detection* - often called model validation in the engineering literature - but it should be clear that it is different from the off-line change detection problem stated in chapter 2, where we test for the presence of a change inside the fixed size sample.

The reasons we consider this off-line detection problem are the following. First, the off-line point of view corresponds to a relevant situation for the vibration monitoring problem, where the fatigue and cracks to be detected have a time constant very much greater than the sampling period, and greater than the mean duration of inspection during which measurements are recorded. Therefore, an off-line processing of Shewhart's type, with fixed size samples of data as described in chapter 2 is definitely adequate. Second, the off-line derivation is useful for investigating the detectability and the diagnosis issues. Actually, as we explained in subsection 7.2.5, we do not solve the very complex on-line diagnosis problem in this book, but only give possible off-line solutions.

It results from (9.3.7) that the function of a sample of observations of size N to be computed for off-line detection is

$$\bar{\mathbf{Y}}_{\mathbf{N}}^* = \frac{1}{\sqrt{N}}\sum_{k=1}^{N}\breve{\mathbf{Y}}_{\mathbf{k}}^* \tag{9.3.8}$$

where $\breve{\mathbf{Y}}_{\mathbf{k}}^*$ is defined in (9.3.4). Using (9.3.5), straightforward computations lead to the following expression :

$$\bar{\mathbf{Y}}_{\mathbf{N}}^* = \sqrt{N}\left(\mathcal{H}_{p+1,p}^T \otimes I_r\right)\mathrm{col}\begin{pmatrix} -\breve{\underline{\theta}} \\ I_r \end{pmatrix} \tag{9.3.9}$$

where the empirical *Hankel matrix* $\mathcal{H}_{p,q}$ is given by

$$\mathcal{H}_{p,q} = \begin{pmatrix} R_0 & R_1 & \dots & R_p & \dots & R_{q-1} \\ R_1 & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ R_{p-1} & \dots & \dots & \dots & \dots & R_{q+p-2} \end{pmatrix} \tag{9.3.10}$$

and the empirical covariance matrices R_l are computed from the sampled measurements by

$$R_l = \frac{1}{N}\sum_{k=1}^{N-l} Y_{k+l}Y_k^T \tag{9.3.11}$$

From expression (9.3.4), it is obvious that the mean value of $\bar{\mathbf{Y}}_N^*$ is zero before change, because e_k is a MA process which is uncorrelated with $\breve{\mathcal{Y}}_{k-2p+1}^{k-p}$. From (9.3.9), we deduce that, under the hypothesis $\underline{\Theta}^* + \frac{\nu}{\sqrt{N}}\Upsilon$,

the mean of $\bar{\mathbf{Y}}_N^*$ is equal to

$$\mathbf{E}_{\Theta^*+\frac{\nu}{\sqrt{N}}\Upsilon}(\bar{\mathbf{Y}}_N^*) = \nu\ (\mathcal{H}_{p,p}^T \otimes I_r)\ \Upsilon = -\nu\ \dot{\kappa}(\underline{\theta}^*)\ \Upsilon \tag{9.3.12}$$

On the other hand, from (9.3.4) again and the fact that $\left(\check{\mathcal{Y}}_{k-2p+1}^{k-p}, e_k\right)$ is independent of $\left(\check{\mathcal{Y}}_{l-2p+1}^{l-p}, e_l\right)$ for $|k-l| \geq p$, we deduce that the covariance matrix of $\bar{\mathbf{Y}}_N^*$ is

$$\check{\Sigma}_N(\underline{\theta}^*) = \frac{1}{N}\sum_{k=1}^{N}\sum_{i=-p+1}^{p-1} \mathbf{E}_{\theta^*}\left[\check{\mathcal{Y}}_{k-2p+1}^{k-p}(\check{\mathcal{Y}}_{k-i-2p+1}^{k-i-p})^T \otimes e_k e_{k-i}^T\right] \tag{9.3.13}$$

It is proven in [Moustakides and Benveniste, 1986] that the following estimate

$$\hat{\Sigma}_N(\underline{\theta}^*) = \frac{1}{N}\sum_{k=1}^{N}\sum_{i=-p+1}^{p-1} e_k\ e_{k-i}^T \otimes \check{\mathcal{Y}}_{k-2p+1}^{k-p}\ (\check{\mathcal{Y}}_{k-i-2p+1}^{k-i-p})^T \tag{9.3.14}$$

is consistent even in the present nonstationary situation. A clever way of computing this estimate is given in [Devauchelle-Gach, 1991] and turns out to be reliable from the numerical point of view.

In summary, the off-line decision function is

$$\chi_N^2 = \bar{\mathbf{Y}}_N^T\ \hat{\Sigma}_N^{-1}\ (\mathcal{H}_{p,p}^T \otimes I_r)\ \left[(\mathcal{H}_{p,p}^T \otimes I_r)^T \hat{\Sigma}_N^{-1} (\mathcal{H}_{p,p}^T \otimes I_r)\right]^{-1}\ (\mathcal{H}_{p,p}^T \otimes I_r)^T\ \hat{\Sigma}_N^{-1}\ \bar{\mathbf{Y}}_N \tag{9.3.15}$$

which is asymptotically distributed as a χ^2 random variable with $2mr$ degrees of freedom (where $n = 2m$ is the state dimension). This χ^2 global test has mean zero when no change occurs and, under the hypothesis of a small change, has a noncentrality parameter equal to

$$\lambda = \nu^2\ \Upsilon^T\ (\mathcal{H}_{p,p}^T \otimes I_r)^T\ \hat{\Sigma}_N^{-1}\ (\mathcal{H}_{p,p}^T \otimes I_r)\ \Upsilon \tag{9.3.16}$$

We rewrite this test as

$$\chi_N^2 = \bar{\mathbf{Y}}_N^T\ \Sigma_N^{-1}\ D(D^T\Sigma^{-1}D)^{-1}D^T\ \Sigma_N^{-1}\ \bar{\mathbf{Y}}_N \tag{9.3.17}$$

where

$$D = \mathcal{H}_{p,p}^T \otimes I_r \tag{9.3.18}$$

We investigate the corresponding detectability issue in section 9.4.

It can be shown that this χ^2-test (9.3.17) is nothing but the test that results from the minmax approach for detecting a change in the AR parameters, while considering the MA parameters as *nuisance* ones [Rougée *et al.*, 1987].

Let us add one comment on the use of the statistical properties of this algorithm for optimal sensor location. Since the power of a χ^2 test is an increasing function of its noncentrality parameter, the noncentrality parameter (9.3.16) of the instrumental test is used as a quantitative measure of the quality of a given sensor location in [Basseville *et al.*, 1987b]. The optimal sensor location problem, which is of crucial interest in some applications, is then solved with the aid of an exhaustive search for maximizing this criterion. This criterion can be used in two different ways [Devauchelle-Gach, 1991] : For a given change to be detected and diagnosed, find the relevant sensor locations; for a given sensor location, find the most detectable and diagnosable changes.

We now address the two diagnosis problems mentioned in section 9.1, using what we call a *sensitivity* technique.

9.3.2 Sensitivity Approach to Diagnosis

Let us first outline the reason we do not address the diagnosis problem using the statistical decoupling approach described in section 7.2. Actually, as we show in that section, the minmax robust approach to statistical diagnosis assumes implicitly that the number of sensors is at least as large as the sum of the dimensions of the changes among which we want to discriminate. It turns out that for the present case of monitoring the eigenstructure of a dynamical system, this condition is scarcely fulfilled : The typical situation is to have only a few number of sensors at our disposal for monitoring a system with significantly more modes. This is the main motivation for the sensitivity approach to diagnosis which we describe now.

First, let us note that sensitivity techniques are classically used in the field of mechanical engineering. What we call a sensitivity method for monitoring can be generally described as follows. The global χ^2 test that we derive in the previous subsection can be written in the following manner :

$$\tilde{\chi}\left[\Theta_0, \Upsilon, (Y_k)_{1\leq k\leq N}\right] \tag{9.3.19}$$

and measures how likely the new record of observations $(Y_k)_{1\leq k\leq N}$ corresponds to the reference model $\Theta_0 + \frac{\Upsilon}{\sqrt{N}}$ rather than to Θ_0. (Here Υ denotes what we called $\nu\Upsilon$) before. Now, for a given Θ_0, the list of changes among which we are interested in discriminating can be characterized by a list of subsets Γ for the change vector Υ. Since we are mainly interested in *small* changes - typically less than 1% in eigenfrequencies and a few percent in the eigenvectors or model shapes in vibration monitoring - we can restrict these subsets Γ to be *linear subspaces*, without loss of generality [Moustakides and Benveniste, 1986, Basseville *et al*., 1987a, Benveniste *et al*., 1987, Benveniste *et al*., 1990]. Following the GLR approach, for each such subspace Γ, the corresponding test is

$$\chi\left[\Theta_0, \Gamma, (Y_k)_{1\leq k\leq N}\right] = \sup_{\Upsilon\in\Gamma} \tilde{\chi}\left[\Theta_0, \Upsilon, (Y_k)_{1\leq k\leq N}\right] \tag{9.3.20}$$

and measures how likely the new record $(Y_k)_{1\leq k\leq N}$ corresponds to the change characterized by Γ. We call this test a *sensitivity test*.

Now assume that our model Θ (9.1.8) can be defined in terms of another parameterization φ by

$$\Theta = f(\varphi) \tag{9.3.21}$$

where f is a locally smooth function. Let ϕ be a subspace spanned by some subset of coordinates of φ of interest in view of diagnosis purposes. Then a relevant choice for the subspace Γ is

$$\Gamma = f'(\varphi^*) \cdot \phi \tag{9.3.22}$$

where φ^* is the value of the parameter φ corresponding to the nominal model Θ^*, and where f' denotes the Jacobian. This provides us with a systematic procedure for monitoring selected components in the φ parameter space.

This general procedure can be used for solving the two diagnosis problems, in terms of the eigen characteristics and in terms of the underlying physical characteristics of the monitored system, in the following manner. From now until the end of this section, we refer to the models described in the vibration monitoring application example of subsection 11.1.4.

9.3.2.1 Parametric or Frequency Diagnosis

In this case, the parameterization φ is made of the eigenfrequencies and observed components of the eigenvectors. They are related to θ - matrix of the AR parameters - through (11.1.27). The only thing that has to

be done is a convenient differentiation of this relation, in order to obtain the Jacobian $\mathcal{J}$ [Basseville *et al.*, 1987a]. When using the new parameterization φ, the mean of the instrumental statistic is

$$D = (\mathcal{H}_{p,p}^T \otimes I_r)\, \mathcal{J} \tag{9.3.23}$$

and the sensitivity test is given by (9.3.17) with this choice of D .

9.3.2.2 Mechanical Diagnosis

In this case, the parameterization φ is equivalent to the physical model of the monitored system. In the case of vibration monitoring, this model is given by the mass, damping, and stiffness matrices (M, C, K) or any equivalent parameterization. However two new difficulties occur in this case. First, usually the physical model is *not* identifiable, which means that f is not invertible and φ^* is not available; and second, the dimension of φ is much larger than that of θ, which means that the subspaces of changes are listed in the φ parameter space and not the θ parameter space. The solution proposed in [Moustakides *et al.*, 1988, Benveniste *et al.*, 1987, Devauchelle-Gach, 1991, Basseville *et al.*, 1993] proceeds as follows. Instead of φ^*, we take a possibly rough approximation of it, for example, the model provided by the designer. Then we compute the image, by the Jacobian computed at this approximate value, of the subspaces in the φ parameter space in order to compute the subspaces in the θ parameter space. Next we cluster these points with the aid of a metric tightly related to the metric of the χ^2 test (9.3.17). The centers of gravity of the resulting classes then give the synthetic Jacobians $\mathcal{J}$ to be used in (9.3.23) and (9.3.17) for computing the sensitivity tests

Experimental results about these two types of diagnosis are reported in chapter 11.

9.4 Detectability

In this section, we investigate the detectability of nonadditive changes in multidimensional signals in the case of changes in the eigenstructure of a state-space model or, equivalently, in the AR part of an ARMA model, which we investigated in section 9.3. As in chapter 7, we consider both the statistical and geometrical points of view.

Remembering the discussion in section 6.3, we can investigate the detectability issue using either an intrinsic information-based point of view with the Kullback information between the distributions before and after change, or a detection-based point of view with the power of the detection test that is used. In the present case of changes in the AR part of an ARMA process, we can start either from the expression of the Kullback information $\mathbf{K}(\theta_1, \theta_0)$ given in (4.1.108) or from the χ^2 decision function that we derived in (9.3.15). This χ^2 variable is centered when no change occurs, and has noncentrality parameter λ given in (9.3.16) when a small change occurs. The use of the detection-based detectability definition leads us to investigate the detectability in terms of the strict positivity of this noncentrality parameter. Now, remembering the approximation (4.1.48) of the Kullback information and the limit (7.3.37), we get that, asymptotically when the sample size goes to infinity, the Kullback information between $\chi^2(nr)$ and $\chi'^2(nr, \lambda)$ goes to half the noncentrality parameter $\frac{\lambda}{2}$. Therefore, we deduce that in the present case the detection-based detectability definition is asymptotically equivalent to the information-based detectability definition.

We also mentioned in section 6.3 that a change can usefully be considered detectable if the mean value of the sufficient statistics - here the instrumental statistics - is *different* when considered before and after the change (see (6.3.5)). This means here that the mean value $(\mathcal{H}_{p,p}^T \otimes I_r)\Upsilon$ after change should be nonzero for this change to be detectable. It turns out that, as in chapter 7, *this statistical detectability definition is equivalent to a pure geometric definition*. Let us show this now.

It can be shown [Benveniste and Fuchs, 1985] that under observability and other weak conditions upon the nominal system (H^*, F^*), the empirical Hankel matrix can be factorized as

$$\mathcal{H}_{p,p}(N) = \mathcal{O}_p(H^*, F^*)\, \mathcal{C}_p(F^*, G_N^*) + \alpha(N) \tag{9.4.1}$$

where $\mathcal{O}_p(H^*, F^*)$ and $\mathcal{C}_p(F^*, G_N^*)$ are the observability and controllability matrices, the latter being assumed uniformly of full row rank n, where G_N is the empirical cross-covariance between the state X and the observation Y, and where $\alpha(N)$ converges to zero in distribution when N goes to infinity.

It results from this factorization that the only changes Υ that are not detectable with the aid of our instrumental statistics are those for which

$$\Upsilon^T \mathcal{O}_p(H^*, F^*) = 0 \tag{9.4.2}$$

These changes are precisely those that are orthogonal to the range of the observability matrix of the system, as the parity checks in subsection 7.4.2. These changes do *not* correspond to any change in the minimal state-space representation of the system [Basseville *et al.*, 1987a], which means that for detecting changes in a minimal representation (H^*, F^*) of the observed signals $(Y_k)_k$ with the aid of the instrumental statistics (9.3.9), it is not necessary to use a minimal ARMA representation of $(Y_k)_k$. This geometrical condition is equivalent to the above-mentioned statistical detectability condition.

9.5 Properties of the Algorithms for Nonadditive Changes

In this section, we describe the available results concerning the properties of the nonadditive change detection algorithms described in *both* chapters 8 and 9. We mainly concentrate on the CUSUM and divergence algorithms designed for simple hypotheses, and on the local linear CUSUM algorithm. All these algorithms have a linear decision function. Recall that, in the general case of conditional distributions investigated in section 8.2, these algorithms can be put into the following general framework

$$t_a = \min\{k \geq 1 : g_k \geq h\} \tag{9.5.1}$$
$$g_k = (g_{k-1} + s_k)^+ \tag{9.5.2}$$
$$g_0 = z \tag{9.5.3}$$

We thus investigate the properties of the algorithms in a manner that depends upon the possible independence of the process of cumulative sum increments $(s_k)_k$.

9.5.1 Independent Increments

We thus first assume that the increment

$$s_k = \ln \frac{p_{\theta_1}(Y_k | \mathcal{Y}_1^{k-1})}{p_{\theta_0}(Y_k | \mathcal{Y}_1^{k-1})} \tag{9.5.4}$$

results in an independent sequence. This property can be achieved under either nonlocal or local assumptions. Let us discuss these two assumptions now.

Nonlocal assumptions In this case, we can use all the theoretical results concerning the optimality and properties of the change detection algorithms that we described in chapter 5. The only thing we need for this purpose is the probability density of the increment s_k parameterized in terms of θ_0 and θ_1.

Example 9.5.1 (CUSUM algorithm for AR case). *In this case, the CUSUM increment is*

$$s_k = \frac{1}{2} \ln \frac{\sigma_0^2}{\sigma_1^2} + \frac{(\varepsilon_k^0)^2}{2\sigma_0^2} - \frac{(\varepsilon_k^1)^2}{2\sigma_1^2} \tag{9.5.5}$$

It should be clear that the only change in the parameter θ (8.1.2) of an AR model for which the increment s_k is independent is the case of a change in the input variance σ^2, for which the two residuals are equal. In the other cases, s_k is a function of two *different residuals and is* not *independent.*

Local assumptions For nonadditive changes, obtaining the density of the increment s_k and/or the ARL function is a complex problem. It is thus of interest to consider particular assumptions that help simplify these derivations. We now show that local assumptions result in such simplifications. The key reason for this is that in local situations we can use the results of section 7.3 for additive changes in the Gaussian case. We investigate the CUSUM and the local linear CUSUM algorithms.

It results from subsection 5.2.2 that the ARL function of the CUSUM algorithm depends upon the single root of the following equation :

$$\mathbf{E}_\theta(e^{-\omega_0 s_k}) = 1 \tag{9.5.6}$$

Let us assume that conditions (4.3.55) and (4.3.56) hold. Define θ^* such that

$$\mathbf{E}_{\theta^*}(s_k) = 0 \tag{9.5.7}$$

It was proven in [Wald, 1947] that the left side of (9.5.6) can be approximated with the aid of the following Taylor expansion :

$$\mathbf{E}_\theta(e^{-\omega_0 s_k}) = 1 - \omega_0 \, \mathbf{E}_\theta(s_k) + \frac{\omega_0^2}{2} \, \mathbf{E}_\theta(s_k^2) - \frac{\omega_0^3}{6} \, \mathbf{E}_\theta(s_k^3 \, e^{-u \, \omega_0 s_k}) \tag{9.5.8}$$

where $0 \leq u \leq 1$ and for θ in a small neighborhood of θ^*. Equations (9.5.6) and (9.5.8) result in

$$-\omega_0 \, \mathbf{E}_\theta(s_k) + \frac{\omega_0^2}{2} \, \mathbf{E}_\theta(s_k^2) - \frac{\omega_0^3}{6} \, \mathbf{E}_\theta(s_k^3 \, e^{-u \, \omega_0 s_k}) = 0 \tag{9.5.9}$$

Let us discuss the solution of this equation in ω_0. First, the third term on the left side is negligible because it is bounded, and thus of the order of $o(\omega_0)$. Therefore, an approximate solution of (9.5.9) can be written as

$$\omega_0 \approx \frac{2 \, \mathbf{E}_\theta(s_k)}{\mathbf{E}_\theta(s_k^2)} \tag{9.5.10}$$

By continuity of the second moment of s_k, we get

$$\omega_0 \approx \frac{2 \, \mathbf{E}_\theta(s_k)}{\mathbf{E}_{\theta^*}(s_k^2)} \tag{9.5.11}$$

Therefore, for local assumptions, we can rewrite equation (5.2.44) for Wald's approximation of the ARL function in the following manner :

$$\begin{aligned} \hat{L}_0(\theta) &= \frac{1}{\mathbf{E}_\theta(s_k)} \left(h + \frac{e^{-\omega_0 h} - 1}{\omega_0} \right) \quad \text{for} \quad \theta \neq \theta^* \\ \hat{L}_0(\theta^*) &= \frac{h^2}{\mathbf{E}_{\theta^*}(s_k^2)} \end{aligned} \tag{9.5.12}$$

and, for an exponential family of distributions, the Siegmund's approximation can be written as

$$\begin{aligned}\tilde{L}_0(\theta) &= \frac{1}{\mathbf{E}_\theta(s_k)}\left(h+\varrho_+-\varrho_-+\frac{e^{-\omega_0(h+\varrho_+-\varrho_-)}-1}{\omega_0}\right) \text{ for } \theta\neq\theta^* \\ \tilde{L}_0(\theta^*) &= \frac{(h+\varrho_+-\varrho_-)^2}{\mathbf{E}_{\theta^*}(s_k^2)}\end{aligned} \tag{9.5.13}$$

The comparison between (5.5.7) and (9.5.12)-(9.5.11) shows that, in some sense, we get the ARL function of the CUSUM algorithm through a Gaussian approximation of the law of s_k. The computation of the first two moments of the increment s_k is as follows (see subsection 4.1.2) :

$$\begin{aligned}\mathbf{E}_{\theta_0}(s_k) &= -\tfrac{1}{2}(\theta_1-\theta_0)^T\ \mathbf{I}(\theta^*)\ (\theta_1-\theta_0) \\ \mathbf{E}_{\theta_1}(s_k) &= +\tfrac{1}{2}(\theta_1-\theta_0)^T\ \mathbf{I}(\theta^*)\ (\theta_1-\theta_0) \\ \mathbf{E}_{\theta^*}(s_k^2) &= \quad (\theta_1-\theta_0)^T\ \mathbf{I}(\theta^*)\ (\theta_1-\theta_0)\end{aligned} \tag{9.5.14}$$

where $\theta^* = \frac{1}{2}(\theta_0+\theta_1)$. Therefore, for *known* parameter values θ_0 and θ_1, we get $\omega_0(\theta_0) = -1$ and $\omega_0(\theta_1) = +1$, and the computation of the ARL function at these points is complete.

On the other hand, it is of interest to investigate the ARL function for actual parameter values different from assumed parameter values. For this reason, let us continue our discussion for the local linear CUSUM algorithm. We first recall the expression of the increment of the decision function of this algorithm :

$$s_k = \Upsilon^T\ Z_k(\theta^*) = \Upsilon^T\ Z_k^* \tag{9.5.15}$$

As we explained in subsection 4.1.2, under local assumptions, the expectation of this increment can be written as

$$\mathbf{E}_\theta(s_k) = \Upsilon^T\ \mathbf{I}(\theta^*)\ (\theta-\theta^*) \tag{9.5.16}$$

and its variance as

$$\mathbf{E}_{\theta^*}(s_k^2) = \Upsilon^T\mathbf{I}(\theta^*)\Upsilon \tag{9.5.17}$$

Therefore, we get

$$\omega_0(\theta) \approx \nu\ \frac{2\Upsilon^T\ \mathbf{I}(\theta^*)\ \tilde{\Upsilon}}{\Upsilon^T\ \mathbf{I}(\theta^*)\ \Upsilon} \tag{9.5.18}$$

where $\theta-\theta^* = \nu\ \tilde{\Upsilon}$ and $\|\tilde{\Upsilon}\| = 1$. From this we deduce the ARL function for the local linear CUSUM algorithm using (9.5.12)-(9.5.13).

Note that all the results obtained in section 7.3 are also valid in the present case for the local linear CUSUM algorithm as local approximations. Another approximation of the ARL function of the CUSUM and local linear CUSUM algorithm can be obtained by inserting the above values of mean and variance of the increment in the solution of the Fredholm integral equation for Gaussian case given in the example 5.2.1.

Finally, as we discussed in subsection 4.3.2, the convenient local characteristic of a statistical test for testing

$$\mathbf{H}_0 = \{\theta\leq\theta^*\} \text{ against } \mathbf{H}_1 = \{\theta>\theta^*\} \tag{9.5.19}$$

is the derivative of the power function $\frac{\partial\beta(\theta^*)}{\partial\theta^*}$. For change detection algorithms in the case of *local* hypotheses, the convenient local characteristic is $\dot{L}_0(\theta^*) = \frac{\partial L_0(\theta^*)}{\partial\theta^*}$. But for *linear* hypotheses, the change direction

Υ is assumed to be known, and, as we explained in section 7.3, the following expression of the ARL function is more convenient :

$$\hat{L}_0(b) = \frac{2b\hat{L}_0(0)^{\frac{1}{2}} + e^{-2b\hat{L}_0(0)^{\frac{1}{2}}} - 1}{2b^2} \tag{9.5.20}$$

where

$$b = \nu \frac{\tilde{\Upsilon}^T \mathbf{I}(\theta^*) \Upsilon}{[\Upsilon^T \mathbf{I}(\theta^*) \Upsilon]^{\frac{1}{2}}} \tag{9.5.21}$$

Let us assume that the change direction is known and the actual and assumed change directions are equal, $\Upsilon = \tilde{\Upsilon}$. In this situation, the convenient characteristic is the derivative of this ARL function with respect to ν at $\nu = 0$. Direct computations result in

$$\dot{L}(0) = \lim_{\nu \to 0} \frac{\partial L_0(\nu)}{\partial \nu} = -\frac{2}{3} \hat{L}^{\frac{3}{2}}(0) \left[\Upsilon^T \mathbf{I}(\theta^*) \Upsilon\right]^{\frac{1}{2}} \tag{9.5.22}$$

9.5.2 Dependent Increments

Recall that the case of changes in the AR parameters of an AR model leads to examples of decision function with dependent increments. In this subsection, we first report some optimality results for the CUSUM algorithm in the general case of conditional densities with known parameters. Then, in some particular cases, we give an approximation of the ARL function based upon an approximation of the decision function by a Brownian motion. It is important to note that the quality of this approximation is not known, although it was investigated in [R.Johnson and Bagshaw, 1974, Bagshaw and R.Johnson, 1975a]. Finally, we show that from a *local* point of view the CUSUM and divergence algorithms have the same properties in the AR case.

Optimality of the CUSUM algorithm The properties of the CUSUM algorithm are investigated in [Bansal and Papantoni-Kazakos, 1986], in the general case of conditional densities and for simple hypotheses about a change generated by the *second* method. This paper contains an extension to the case of dependent processes of the results in [Lorden, 1971] which we reported in chapter 5. Let us recall, in the dependent case, the derivation of the CUSUM algorithm based upon open-ended tests :

$$t_a = \min_{j=1,2,\ldots} \{T_j\} \tag{9.5.23}$$

$$T_j = \min \{k \geq j \ : \ln \breve{\Lambda}_j^k \geq h\} \tag{9.5.24}$$

$$\breve{\Lambda}_j^k = \frac{p_{\theta_1}(\mathcal{Y}_j^k)}{p_{\theta_0}(\mathcal{Y}_j^k)} \tag{9.5.25}$$

Following [Bansal and Papantoni-Kazakos, 1986], we consider the class of extended stopping times :

$$K_T = \{T^* \ : \mathbf{E}_{\theta_0}(T^*) \geq \bar{T}\} \tag{9.5.26}$$

and we define

$$\eta^*(\bar{T}) = \inf_{T^* \in K_T} \bar{\mathbf{E}}_{\theta_1}(T^*) \tag{9.5.27}$$

where

$$\bar{\mathbf{E}}_{\theta_1}(T^*) = \sup_{k \geq 1} \operatorname{ess\,sup} \mathbf{E}_k[(T^* - k + 1)^+ | \mathcal{Y}_1^{k-1}] \tag{9.5.28}$$

and where $\mathbf{E}_k$ is the expectation under the distribution of the observations when the change time is k. The two main results in [Bansal and Papantoni-Kazakos, 1986] are the following. When $\bar{T}$ goes to infinity, we have first

$$\eta^*(\bar{T}) \sim \frac{\ln(2\,\bar{T})}{\mathbf{E}_{\theta_1}(\mathbf{K}_{1,0})} \tag{9.5.29}$$

where

$$\mathbf{K}_{1,0} = \lim_{k\to\infty} \frac{\ln \breve{\Lambda}_1^k}{k} \tag{9.5.30}$$

and second

$$\bar{\tau}^* \sim \frac{\ln(2\,\bar{T})}{\mathbf{E}_{\theta_1}(\mathbf{K}_{1,0})} \tag{9.5.31}$$

The first result provides us with a lower bound for the worst mean delay for detection in the class K_T. The second results states that the CUSUM algorithm (9.5.23) reaches this lower bound asymptotically. Note that here we recover the dependent counterpart of Lorden's result in (5.2.10).

Approximation of the ARL function In the case of dependent increments, we approximate the decision function by a Brownian motion as we did in subsection 8.1.3. Furthermore, it should be clear from the decision rule (9.5.2) that the relevant boundaries for the limit Brownian motion are lower reflecting and upper absorbing boundaries, respectively. The expectation of such an exit time $T_{0,h}$ was given in subsection 3.1.4, and thus provides us with an approximation of the ARL function :

$$\mathbf{E}_\theta(T_{0,h}) = \frac{1}{\mu}\left(h + \frac{e^{-\frac{2\mu}{\sigma^2}h} - 1}{\frac{2\mu}{\sigma^2}}\right) \tag{9.5.32}$$

where μ and σ^2 are the drift and diffusion coefficients of the Brownian motion. It results from the central limit theorem (8.1.39) that the formulas for the drift and diffusion coefficients are given in (9.5.14) for the CUSUM and in (9.5.16)-(9.5.17) for the local linear CUSUM algorithm. Note that this approximation is the same as the Wald's approximation (9.5.12) given above.

Comparison between the CUSUM and divergence algorithms Now we show that, from a local point of view, the CUSUM and divergence algorithms have the same properties. More precisely, we show that, up to second-order terms, the decision functions of the two algorithms have the *same* mean values before and after change, respectively. For simplicity of notation, we focus on the case of a scalar signal.

First we recall that the CUSUM algorithm is based upon the likelihood ratio increment (9.5.4) and that, up to an additive constant which does not change the following reasoning, the divergence algorithm is based upon the increment :

$$\tilde{s}_i = s_i - \mathbf{E}_{\theta_0}(s_i|\mathcal{Y}_1^{i-1}) \tag{9.5.33}$$

$$= s_i - \bar{s}_i \tag{9.5.34}$$

Now we show that $\bar{s}_i$ *is almost surely equal to zero up to second-order terms under both distributions* p_{θ_0} and $p_{\theta_0+\frac{\nu}{\sqrt{N}}\Upsilon}$. First we note that it is sufficient to show this property under p_{θ_0} only, since the sets of measure zero are the same under both distributions p_{θ_0} and $p_{\theta_0+\frac{\nu}{\sqrt{N}}\Upsilon}$.

Let us consider the following first-order approximations of s_i and $\bar{s}_i$:

$$s_i \approx \frac{\nu}{\sqrt{N}}\Upsilon^T Z_i \tag{9.5.35}$$

$$\bar{s}_i \approx \frac{\nu}{\sqrt{N}}\Upsilon^T \bar{Z}_i \tag{9.5.36}$$

where Z_i is the efficient score computed at θ_0 :

$$Z_i = \frac{\partial}{\partial \theta} \ln p_\theta(y_i|\mathcal{Y}_1^{i-1})\Big|_{\theta=\theta_0} \tag{9.5.37}$$

Recall that, as we showed in section 4.1, the efficient score Z_i satisfies

$$\mathbf{E}_{\theta_0}(Z_i|\mathcal{Y}_1^{i-1}) = 0 \tag{9.5.38}$$

$$\mathbf{E}_{\theta_0+\frac{\nu}{\sqrt{N}}\Upsilon}(Z_i|\mathcal{Y}_1^{i-1}) \approx \frac{\nu}{\sqrt{N}}\,\mathbf{I}(\theta_0)\,\Upsilon \tag{9.5.39}$$

Let us now compute the conditional expectation of $\tilde{s}_i$ before change :

$$\begin{aligned}\mathbf{E}_{\theta_0}(\tilde{s}_i|\mathcal{Y}_1^{i-1}) &= \mathbf{E}_{\theta_0}(s_i|\mathcal{Y}_1^{i-1}) - \mathbf{E}_{\theta_0}(\bar{s}_i|\mathcal{Y}_1^{i-1}) \\ &= \mathbf{E}_{\theta_0}(s_i|\mathcal{Y}_1^{i-1}) - \bar{s}_i \\ &\approx \frac{\nu}{\sqrt{N}}\Upsilon^T\left[\mathbf{E}_{\theta_0}(Z_i|\mathcal{Y}_1^{i-1}) - \bar{Z}_i\right]\end{aligned} \tag{9.5.40}$$

The left side of the last relation is equal to zero by definition of $\tilde{s}_i$. The first term of the right side is zero by definition of the efficient score. Therefore, we get that, up to first-order terms, $\bar{Z}_i$ is almost surely equal to zero under p_{θ_0}, and also under $p_{\theta_0+\frac{\nu}{\sqrt{N}}\Upsilon}$, as stated before. Thus, up to second-order terms, $\bar{s}_i$ is almost surely equal to zero under both p_{θ_0} and $p_{\theta_0+\frac{\nu}{\sqrt{N}}\Upsilon}$, and $\tilde{s}_i$ and s_i have the same mean values before and after change, up to second-order terms again. The CUSUM and divergence algorithms are thus not distinguishable when considered from a local point of view.

9.6 Notes and References

Section 9.2

The on-line detection of nonadditive changes in multidimensional AR/ARMA models is investigated in [R.Jones *et al.*, 1970, Nikiforov, 1980, Nikiforov, 1983, Vorobeichikov and Konev, 1988].

Section 9.3

The use of the local approach together with non-likelihood based statistics was introduced in [Basseville *et al.*, 1986, Basseville *et al.*, 1987a] for solving the vibration monitoring problem. The noncentrality parameter (9.3.16) of the instrumental test is used as a quantitative measure of the quality of a given sensor location in [Basseville *et al.*, 1987b].

The use of the local approach together with non-likelihood-based statistics was extended in [Benveniste *et al.*, 1987, Benveniste *et al.*, 1990] for designing change detection algorithms associated with any adaptive parametric identification algorithms. An extension of the method, allowing model reduction and biased identification, is reported in [Zhang, 1991, Zhang *et al.*, 1994], together with an example concerned with the monitoring of the combustion chambers of a gas turbine.

Section 9.5

The optimality of the CUSUM algorithm for the general case of a conditional density of a dependent process was investigated first in [Bansal and Papantoni-Kazakos, 1986]. The ARL function for the independent case

is investigated in [Van Dobben De Bruyn, 1968, Lorden, 1971, Reynolds, 1975, Khan, 1978, Nikiforov, 1980, Nikiforov, 1983, Siegmund, 1985b].

In [Ladelli, 1990], the non-likelihood-based algorithm designed with the aid of the central limit theorem (8.4.17)-(9.1.23) is shown to be optimal among all algorithms based upon the non-likelihood statistics (9.1.21).

9.7 Summary

AR/ARMA Models and Likelihood Ratios

Increment of the log-likelihood ratio

$$s_k = \frac{1}{2}\ln\frac{\det R_0}{\det R_1} + \frac{1}{2}(\varepsilon_k^0)^T R_0^{-1}\varepsilon_k^0 - \frac{1}{2}(\varepsilon_k^1)^T R_1^{-1}\varepsilon_k^1$$

Efficient score for an AR(1) process

$$Z_k^* = \begin{pmatrix} \frac{1}{2}\check{R}^* - \frac{1}{2}\varepsilon_k^*(\varepsilon_k^*)^T \\ \\ \check{R}^*\varepsilon_k^* Y_{k-1}^T \end{pmatrix}$$

Detection of Changes in the Eigenstructure

Instrumental statistic

$$\begin{aligned}
\bar{\mathbf{Y}}_{\mathbf{N}} &= \sqrt{N}\left(\mathcal{H}_{p+1,p}^T \otimes I_r\right)\operatorname{col}\begin{pmatrix} -\check{\underline{\theta}} \\ I_r \end{pmatrix} \\
\check{\underline{\theta}}^T &= \begin{pmatrix} A_p & \dots & A_1 \end{pmatrix} \\
\hat{\Sigma}_N(\underline{\theta}^*) &= \frac{1}{N}\sum_{k=1}^{N}\sum_{i=-p+1}^{p-1} \check{\mathcal{Y}}_{k-2p+1}^{k-p}(\check{\mathcal{Y}}_{k-i-2p+1}^{k-i-p})^T \otimes e_k e_{k-i}^T \\
(\chi_j^k)^2 &= \bar{\mathbf{Y}}_N^T\,\hat{\Sigma}_N^{-1}\,(\mathcal{H}_{p,p}^T \otimes I_r)\,\left[(\mathcal{H}_{p,p}^T \otimes I_r)^T\hat{\Sigma}_N^{-1}(\mathcal{H}_{p,p}^T \otimes I_r)\right]^{-1}\,(\mathcal{H}_{p,p}^T \otimes I_r)^T\,\hat{\Sigma}_N^{-1}\,\bar{\mathbf{Y}}_N \\
\lambda &= \nu^2\,\Upsilon^T\,(\mathcal{H}_{p,p}^T \otimes I_r)^T\,\hat{\Sigma}_N^{-1}\,(\mathcal{H}_{p,p}^T \otimes I_r)\,\Upsilon
\end{aligned}$$

Detectability

For the instrumental statistic

$$\Upsilon^T\mathcal{O}_p(H^*, F^*) \neq 0$$

Properties of the CUSUM and Linear CUSUM Algorithms

Independent increments

$$\dot{L}(0) = \left.\frac{\partial L_0(\nu)}{\partial\nu}\right|_{\nu=0} = -\frac{2}{3}\,\hat{L}^{\frac{3}{2}}(0)\left[\Upsilon^T\,\mathbf{I}(\theta^*)\,\Upsilon\right]^{\frac{1}{2}}$$

Dependent increments

$$\bar{\tau}^* \sim \frac{\ln \bar{T} + \ln 2}{\mathbf{E}_{\theta_1}(\mathbf{K}_{1,0})}$$

Moreover, the CUSUM and divergence algorithms are not distinguishable when considered from a local point of view.

Part III

Tuning and Applications

10

Implementation and Tuning

In this chapter, we investigate the important issues of implementing and tuning change detection algorithms. In chapters 2, 7, 8, and 9, we described how to design algorithms. In chapters 5, and also 7 and 9, we described how to investigate analytically and numerically the properties of detection algorithms. Now, we first discuss how to select a convenient algorithm. Then we consider change detection algorithms with a *given structure*, and we investigate how to choose what we call the *tuning values*. Actually, in each change detection algorithm, there exist free parameters that must be chosen before using the algorithm on real data. These values are typically threshold for the decision function, window sizes, weights, nominal values of parameter θ_0 before change, expected values of parameter θ_1 after change, or equivalently minimum magnitude of change.

The main **goals** of this chapter are the following. We first propose a *unified methodology* for solving the above problem using the analytical and numerical results for computing the ARL function for those algorithms for which such results do exist. When the ARL function is not available, analytically or numerically, we apply the same methodology, using the detectability criterion introduced in chapter 6 as a weak performance index. Note that in this case no tuning of the threshold can be achieved. Second, we investigate the critical problem of *robustness* with respect to several issues. We of course discuss the robustness of change detection algorithms with respect to deviations between the values of the parameters θ_0 and θ_1 that are chosen for tuning and the true values. Next, we discuss the robustness with respect to *nuisance parameters*. Model parameters can often be classified into two groups: informative parameters and nuisance parameters. The first group contains the parameters that are to be monitored. The second group contains the parameters that are of no interest as far as change detection is concerned, but that can have a nonnegligible influence on the performance of the change detection algorithm. For example, in the case of additive changes, the noise variance is not a parameter of interest in itself, but sometimes has a critical influence when it is either underestimated or time-varying, as we discussed in example 7.2.3. The last type of robustness that we address deals with the issue of errors in assumptions concerning the model itself, such as unmodelled correlations or non-Gaussian noises.

The **tools** we use for achieving these goals are the following. Several aspects of tuning of detection algorithms have strong connections with the theory of pattern recognition. We show that the problem of tuning the free values (reference point, minimum magnitude of change, size of sliding window, forgetting coefficient, and so on, but not threshold) of an algorithm is equivalent to the choice of the free parameters of a discriminant function. Consequently, optimization procedures can be used for this tuning. On the other hand, for simple hypotheses before and after change, the tuning of the threshold is achieved with the aid of analytical results concerning the ARL function.

In this chapter, we first describe the general methodology we propose for this tuning. Then, in the subsequent sections, we apply this methodology to several change detection algorithms introduced in chapter 2 for the scalar parameter case and in chapters 7, 8, and 9 for the more complex cases.

10.1 General Methodology

This methodology should be thought of more as a collection of heuristic ideas obtained from our experience than as a closed theory on this topic. Generally speaking, the tuning of change detection algorithms for a particular application can be made of the following steps :

1. preliminary investigation of the problem;

2. choice of a relevant change detection algorithm;

3. tuning the parameters of this algorithm;

4. checking the robustness of this algorithm on real data.

From a practical point of view, there exist different possible paths through these steps for solving a given problem. In ideal situations, this path is forward from step 1 to step 4. But often several backward paths are of interest. More precisely, checking for robustness at step 4 often leads us to change the tuning values in step 3, to choose another algorithm in step 2, or even to reconsider the investigation of the problem at step 1.

Let us describe these steps more precisely.

10.1.1 Investigating the Problem

An initial investigation of the problem should include the following :

- an informal problem statement, namely where the detection problem is!
- choice of relevant measurements on the considered system, including possible adequate quality indexes : temperature, pressure, concentration, speed, acceleration, ...;
- choice of convenient models for these signals;
- choice of the changes of interest in these models;
- qualitative evaluation of the possible influence of nuisance parameters on the changes of interest;
- estimation of the time constants of the changes with respect to the dynamics of the system;
- evaluation of admissible rate of false alarms and delay for detection, again relative to the dynamics of the system and the sampling frequency.

Note that the second and third issues are not specific to change detection. They arise in identification problems, and we refer the reader to the corresponding literature for some methodology about choice of models and other identification issues [Box and Jenkins, 1970, Ljung, 1987, Söderström and Stoïca, 1989, Benveniste *et al.*, 1990].

On the other hand, the choice of changes of interest can be achieved in different ways according to the complexity of the parameters that are subject to change. This choice can be made through elementary processing of numerous sets of real data that are typical of situations before and after change, when such data are available of course. An additional tool for achieving this choice consists of using insights about the underlying physical model when available too. Moreover, it is of key interest to capture all possible information about the shape of the sets of parameters Θ_0 and Θ_1 before and after change, respectively. This information can be obtained from physical models, technological charts and standards, and is useful

for selecting the type or amount of available *a priori* information about the changes, such as the seven cases discussed in subsection 7.2.1. It is important at this step to outline possible nuisance parameters in the considered model, because this helps us investigate the detectability of the relevant changes and the robustness of the algorithm with respect to these nuisance parameters.

By time constants of the changes, we refer to two types of quantities. First, we refer to the empirical frequencies of occurrence of the different changes, which have to be viewed as relative with respect to other time constants in the underlying system, namely the dynamics of the system and of possible perturbations, and the sampling frequency of the measurements. It is often of interest to take these empirical frequencies into account *together* with the admissible delay for detection. The consequence of this on the choice of algorithms is discussed next. Second, we also refer to the dynamics of the change, or equivalently to the durations of the dynamic profiles of the transient responses of the considered system to the different changes, which also must be viewed as relative with respect to the above-mentioned dynamics of the system itself. Finally, the admissible rate of false alarms and delay for detection help in selecting relevant algorithms.

10.1.2 Choosing the Algorithm

Generally speaking, the issues that can help in selecting relevant change detection algorithms are the following :

- detailed problem statements, such as detection versus estimation of the change;
- time constants of the changes, as defined before;
- admissible values of the criteria (rate of false alarms, delay for detection);
- available *a priori* information about Θ_0 and Θ_1;
- possible influence of the nuisance parameters;
- implementation problems.

Let us discuss how these issues influence the choice of an algorithm. Requirements concerning the estimation of the change time and magnitude automatically exclude the algorithms that have no estimation ability. Such requirements typically arise for recognition oriented signal processing, signal onset detection, and so on.

The main goal of the estimation of the empirical frequencies of the possible changes is aimed at the choice between Bayesian and non-Bayesian algorithms. However, a Bayesian change detection algorithm is relevant only when the *a priori* distribution of the change time, and not only its mean, is available.

Next, the knowledge of admissible values of the rate of false alarms and mean delay for detection can be used as bounds for the properties that the chosen algorithm must have. For this purpose, these known values should be compared to the ARL function for optimal algorithms. The solution to this problem shows whether the considered problem statement is realistic or not, and how far this problem statement is from the optimal solution.

The available *a priori* information about Θ_0 and Θ_1 helps in choosing the relevant algorithm simply because we use precisely this information for classifying the algorithms, as in subsection 7.2.1. Moreover, when there exist, for one given level of *a priori* information, several possible algorithms, it is necessary to take into account the sensitivity of these algorithms with respect to this information in order to choose the most convenient one.

On the other hand, the algorithms described in this book are known to have different sensitivity with respect to nuisance parameters, as we discussed in chapter 7 and investigate in the rest of this chapter. Therefore, any information concerning the interaction between the parameters of interest and the nuisance parameters, considered together with these sensitivities, can help in choosing the convenient algorithm.

Finally, the constraints arising from the available computing facilities should be taken into account as a tradeoff between the complexity and the efficiency of the algorithms. This issue was discussed in chapters 2, 5, 7, 8, and 9.

10.1.3 Tuning the Parameters

Tuning is the most important of the steps we listed at the beginning of this section. The reason for this is that it can be investigated basically with more technical than philosophical issues, as opposed to the two previous steps. The key tool for this investigation is the ARL function, or the detectability index when the ARL function does not exist. For this reason, our methodology for tuning the parameters of a change detection algorithm distinguishes between the following situations :

- scalar parameter case, algorithms for which the ARL function exists;
- known vector parameter case, algorithms for which the ARL function exists;
- vector parameter case, linear decision functions for which the ARL function exists;
- vector parameter case, linear decision functions for which the ARL function does not exist;
- vector parameter case, quadratic decision functions for which the ARL function exists;
- vector parameter case, quadratic decision functions for which the ARL function does not exist.

Generally speaking, the problem of tuning the parameters of the algorithm mainly reduces to the numerical solution of an implicit equation and/or to the solution of an optimization problem. The details underlying this general statement are explained for the above different situations in the subsequent sections. Note that this general structure does not apply to such parameters as window sizes, for example, but only to thresholds, weights, and assumed values of parameters.

10.1.4 Robustness Issues

We have selected an algorithm and chosen its tuning parameters. This particular algorithm implicitly contains simplifications with respect to the complexity of the data that we are ready to analyze. The problem is now to measure the robustness of the algorithm, tuned as it is, with respect to these simplifications and to the possible nuisance parameters. In most cases, no analytical solution to this problem exists, and the only way to investigate it is to try the algorithm on numerous data sets. However, we discuss this robustness issue for some particular cases in the following sections.

10.2 Scalar Case

In this section, we begin our discussion of tuning change detection algorithms by investigating the case of a scalar parameter in an independent sequence. The corresponding algorithms were described in chapter 2 and their properties were investigated in chapter 5 mainly in terms of the ARL function. Recall that this function provides us with both the mean time between false alarms and the mean delay for detection.

10.2.1 Main Idea

We consider change detection algorithms for which there exists an analytical expression of the ARL function L or a numerical method for computing this function, and distinguish between the cases of simple and composite hypotheses.

10.2.1.1 Simple Hypotheses

In the case of simple hypotheses, the parameters θ_0 and θ_1 are assumed to be known. Because of our definition of criteria, on-line change detection algorithms are characterized by the mean time between false alarms $\bar{T}$ and the mean delay for detection $\bar{\tau}$. We assume that $\bar{T} = L(\theta_0)$ and $\bar{\tau} = L(\theta_1)$. Note here that for some algorithms, this relation concerning $\bar{\tau}$ is not straightforward, because of the random behavior of the initial value of the decision function at the change time. We refer to chapter 5 for more thorough investigations of this point.

From now on, we make use of the following notation :

$$\Delta = \{h, \lambda, N, \alpha, \ldots\} \tag{10.2.1}$$

for the vector of the tuning parameters : thresholds, sample size, forgetting coefficient, and so on. For fixed values of θ_0 and θ_1, $\bar{T}$ and $\bar{\tau}$ are functions of these tuning parameters :

$$\begin{aligned} \bar{T} &= \bar{T}(\Delta) \\ \bar{\tau} &= \bar{\tau}(\Delta) \end{aligned} \tag{10.2.2}$$

Three possibilities exist :

- fix $\bar{T} = T^*$, compute the tuning values as $\Delta^* = \bar{T}^{-1}(T^*)$, and then compute the delay $\bar{\tau}^* = \bar{\tau}(\Delta^*)$;
- fix $\bar{\tau} = \tau^*$, compute the tuning values as $\Delta^* = \bar{\tau}^{-1}(\tau^*)$, and then compute the mean time between false alarms $\bar{T}^* = \bar{T}(\Delta^*)$;
- choose a penalty function $\omega = \omega(\bar{T}, \bar{\tau})$, for example, $\omega = \gamma_1 \bar{\tau} + \gamma_2 / \bar{T}$, compute the tuning values with the aid of an optimization algorithm :

  $$\Delta^* = \arg \inf_{\Delta} \ \omega(\bar{T}, \bar{\tau}) \tag{10.2.3}$$

  and then compute the criteria using $\bar{T}^* = \bar{T}(\Delta^*)$ and $\bar{\tau}^* = \bar{\tau}(\Delta^*)$.

Even in the simple scalar case of independent increments, these inversion and minimization problems are not easy to solve.

10.2.1.2 Composite Hypotheses

In previous chapters, we distinguished simple and composite hypotheses for the design of change detection algorithms. Now we take the point of view of tuning the parameters of a given algorithm, and investigate how to choose the free parameters Δ of this algorithm, whether it is optimal or not.

In the case of composite hypotheses, we must use a slightly different criterion because $\bar{T}$ and $\bar{\tau}$ are also functions of θ :

$$\begin{aligned} \bar{T} &= \bar{T}(\theta, \Delta) \\ \bar{\tau} &= \bar{\tau}(\theta, \Delta) \end{aligned} \tag{10.2.4}$$

In this case, two possibilities exist :

- choose two weighting functions ω_T and ω_τ to define two weighted criteria :

  $$\begin{aligned} \bar{T}_\omega &= \int_{\theta \in \Theta_0} \omega_T(\theta) \, \bar{T}(\theta, \Delta) \, d\theta \\ \bar{\tau}_\omega &= \int_{\theta \in \Theta_1} \omega_\tau(\theta) \, \bar{\tau}(\theta, \Delta) \, d\theta \end{aligned} \tag{10.2.5}$$

  which can then be used in the same way as $\bar{T}, \bar{\tau}$ in the case of simple hypotheses;

- compute the least favorable criteria with respect to the *a priori* information :

$$\begin{aligned} T &= \inf_{\theta \in \Theta_0} \bar{T}(\theta, \Delta) \\ \tau &= \sup_{\theta \in \Theta_1} \bar{\tau}(\theta, \Delta) \end{aligned} \tag{10.2.6}$$

  which again can be used as before.

10.2.2 Examples of Algorithms

We now add some comments on the application of this main idea to the algorithms of chapter 2, and then of chapters 7 and 8.

10.2.2.1 Elementary Algorithms

We begin with Shewhart's charts. Assuming that the mean μ_0 before change and the variance σ^2 are known, the tuning parameters of this algorithm are the sample size N and the control limit κ. The tuning of these parameters is discussed in [Page, 1954c] where tables are given according to two possible optimizations. The first consists of, for fixed $\bar{T}$ and change magnitude $\nu = \mu_1 - \mu_0$, finding the optimal values of N and κ that minimize $\bar{\tau}$. The second uses a fixed $\bar{\tau}$ and maximizes $\bar{T}$. More recent investigations, taking into account the serial correlation in the observations, can be found in [Vasilopoulos and Stamboulis, 1978].

Under the same assumptions as before, the tuning of the GMA algorithm concerns the forgetting coefficient α and the threshold h. Many tables can be found in [Robinson and Ho, 1978], for the one-sided and two-sided versions of the GMA algorithm, and for either fixed $\bar{T}$ or fixed $\bar{\tau}$. Other tables are found in [Crowder, 1987].

The tuning parameters Δ of the FMA algorithm are the window size N, the weighting coefficients $\gamma_0, \ldots, \gamma_{N-1}$, and the threshold h. The ARL function of the FMA algorithm, for known values of Δ, is computed in [Laï, 1974, Böhm and Hackl, 1990]. But, to our knowledge, there does not exist an optimization procedure for choosing Δ.

10.2.2.2 CUSUM-type Algorithms

We consider first the case of the CUSUM algorithm corresponding to simple hypotheses. In this case, the only tuning parameter is the threshold h. Using formulas for the ARL function, we can compute $\bar{\tau}$ and $\bar{T}$ and use them for choosing a relevant h, as we explained before. Tables and nomograms for the ARL function can be found in [Van Dobben De Bruyn, 1968, Goel and Wu, 1971]. In the case of unknown change magnitude, the tuning of $\Delta = (\nu, h)$ can be achieved by using one of the above-mentioned methods for the case of composite hypotheses.

For the CUSUM algorithm detecting a change in the mean, typical nuisance parameters are the input variance and the correlations. The problem of robustness of one- and two-sided CUSUM algorithms with respect to the unknown variance is investigated in [Bagshaw and R.Johnson, 1975b]. The robustness with respect to correlations is investigated in [Goldsmith and Whitfield, 1961, Kemp, 1961, R.Johnson and Bagshaw, 1974, Bagshaw and R.Johnson, 1975a, Nikiforov, 1983]. In [Bagshaw and R.Johnson, 1975b], the Wald's approximation (5.5.7) is used for investigating the effect of the unknown variance on the estimates $\bar{\tau}$ and $\bar{T}$. It turns out that the robustness of the CUSUM algorithm with respect to the variance actually depends upon the change magnitude : We recover here the standard signal-to-noise ratio issue. Moreover, the two-sided CUSUM is less robust than the one-sided one with respect to the unknown variance. Finally, if the variance is underestimated, that is if the actual variance is greater than the assumed variance, then $\bar{\tau}$ and

$\bar{T}$ are overestimated, that is the actual criteria are less than the assumed criteria. The converse statements hold true when the variance is overestimated. On the other hand, the effect of the presence of correlations among the observed data can be summarized as follows. In [Goldsmith and Whitfield, 1961, Kemp, 1961], simulations show that the CUSUM algorithm is robust provided that the correlations remain reasonable. In the case of an AR(1) and a MA(1) process, the use of a Brownian motion approximation for computing the ARL function leads to the same conclusion [R.Johnson and Bagshaw, 1974, Bagshaw and R.Johnson, 1975a]. The computation of the Fisher information matrix, with respect to the mean, variance, and autoregressive coefficients (which is diagonal in the present cases of model of order 1) leads to similar conclusions [Nikiforov, 1983].

The robustness of the CUSUM algorithm designed for the independent Gaussian case with respect to the higher order moments of the distribution is investigated in [Bissell, 1969] where nomograms are given. Finally, the tuning quantities of the weighted CUSUM algorithm are the weighting function $F(\theta)$ and the threshold h. No procedure exists for choosing the weighting function. From the practical point of view, this function should obviously reflect all the available *a priori* information concerning the change detection problem to be solved. Again the choice of h is achieved by using the formulas given in subsection 5.2.3.

10.2.2.3 GLR Algorithm

Again assuming θ_0 to be known, the tuning parameter of the GLR algorithm is $\Delta = (\underline{\theta}, \bar{\theta}, h)$. In section 5.3, we gave asymptotic formulas for computing $\bar{\tau}$ and $\bar{T}$ as functions of Δ, which again can be used following the main idea explained before.

10.2.2.4 More Complex Cases for which the ARL Function Exists

As discussed in chapters 7 and 9, there exist many processes with models much more complex than the simple case discussed in chapter 2, but for which the ARL function of the CUSUM algorithm can be computed. In these cases, the tuning of the parameters is *exactly* as before and we can use all the solutions described in the present section.

10.3 Vector Case with Linear Decision Function

We now discuss the changes in multidimensional parameters detected with a decision function that is linear in the parameters. In other words, we discuss the tuning of the linear CUSUM algorithm. We start with the case of additive changes discussed in section 7.2.1, and consider a generalization to the algorithms resulting from local approximations in the case of nonadditive changes.

10.3.1 Additive Changes

Here we refer to linear CUSUM algorithms of the form

$$\begin{aligned} t_a &= \min\{k : g_k \geq h\} && (10.3.1)\\ g_k &= (g_{k-1} + s_k)^+ && (10.3.2)\\ s_k &= \Upsilon^T \Sigma^{-1} (Y_k - \theta^*) && (10.3.3) \end{aligned}$$

where we assume that $(s_k)_k$ is an i.i.d. Gaussian sequence, with mean and variance :

$$\begin{aligned} \mathbf{E}(s) &= \Upsilon^T \Sigma^{-1} (\theta - \theta^*) \\ \mathbf{E}(s^2) &= \Upsilon^T \Sigma^{-1} \Upsilon \end{aligned} \qquad (10.3.4)$$

In this case, the ARL function is

$$\hat{L}(b) = \frac{2b\hat{L}^{\frac{1}{2}}(0) + e^{-2b\hat{L}^{\frac{1}{2}}(0)} - 1}{2b^2} \tag{10.3.5}$$

where

$$b = \nu \, \frac{\Upsilon^T \Sigma^{-1} \tilde{\Upsilon}}{(\Upsilon^T \Sigma^{-1} \Upsilon)^{\frac{1}{2}}} \tag{10.3.6}$$

Recall that $\nu \, \tilde{\Upsilon} = \theta - \theta^*$ is the actual value of the change vector, and that Υ is the (unit) assumed value of the change direction. The tuning parameter is then $\Delta = (\theta^*, \Upsilon, h)$. There are several possibilities for tuning such algorithms, but they all rely upon the same idea of choosing a discriminant surface between Θ_0 and Θ_1, as depicted in figure 7.2. Two possible solutions exist for this tuning. The first consists of choosing an optimal vector Υ for a fixed reference point θ^*. The second consists of a joint optimal choice of Υ and θ^* by using methods of mathematical programming. Let us describe this now.

10.3.1.1 Optimal Choice of Υ

We fix θ^* and discuss how to choose Υ. As we explained in section 7.3, the ARL function of the linear CUSUM algorithm depends upon the ratio :

$$f(\Upsilon, \tilde{\Upsilon}) = \frac{\Upsilon^T \Sigma^{-1} \tilde{\Upsilon}}{(\Upsilon^T \Sigma^{-1} \Upsilon)^{\frac{1}{2}}} \tag{10.3.7}$$

Let us consider the unit sphere $\mathcal{S}$ centered at θ^* generated by the other extremity of the vector Υ. We assume that there exists a weighting function $p(\tilde{\Upsilon})$ on the surface of this unit sphere, such that $\int_{\mathcal{S}} p(\tilde{\Upsilon}) \; d\mathcal{S} = 1$. The optimal tuning problem can be stated as the following optimization problem :

$$\Upsilon^* = \arg\sup_{\Upsilon} \int_{\mathcal{S}} \frac{\Upsilon^T \Sigma^{-1} \tilde{\Upsilon}}{(\Upsilon^T \Sigma^{-1} \Upsilon)^{\frac{1}{2}}} \; p(\tilde{\Upsilon}) \; d\mathcal{S} \tag{10.3.8}$$

On the other hand, if such a weighting function does not exist, but if there exists a region $\mathcal{S}_1$ on the surface of the sphere, characterizing the range of the possible extremities of $\tilde{\Upsilon}$, then the following minmax approach can be used :

$$\Upsilon^* = \arg\sup_{\Upsilon} \inf_{\tilde{\Upsilon} \in \mathcal{S}_1} \frac{\Upsilon^T \Sigma^{-1} \tilde{\Upsilon}}{(\Upsilon^T \Sigma^{-1} \Upsilon)^{\frac{1}{2}}} \tag{10.3.9}$$

Example 10.3.1 (Two-dimensional case). *We now consider the two-dimensional case and assume that $\Sigma = I_2$. The ratio (10.3.7) can be written as*

$$f(\Upsilon, \tilde{\Upsilon}) = \Upsilon^T \tilde{\Upsilon} \tag{10.3.10}$$

and is discussed while making use of figure 10.1. Assume that there exists an a priori known weighting function $p(\beta)$ in the sector (β_1, β_2). In this case, solution (10.3.8) is

$$\alpha^* = \arg\sup_{\alpha} \left[\int_{\beta_1}^{\beta_2} (\cos\alpha \cos\beta + \sin\alpha \sin\beta) \; p(\beta) \; d\beta \right] \tag{10.3.11}$$

When no such distribution is known, the minmax approach (10.3.9) can be written as

$$\alpha^* = \arg\sup_{\alpha} \inf_{\beta \in [\beta_1, \beta_2]} (\cos\alpha \cos\beta + \sin\alpha \sin\beta) \tag{10.3.12}$$

$$= \arg\sup_{\alpha} \inf_{\beta \in [\beta_1, \beta_2]} \cos(\alpha - \beta) \tag{10.3.13}$$

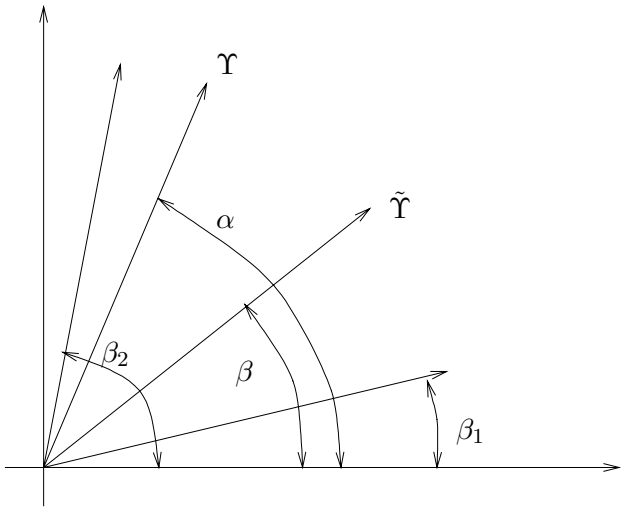

Figure 10.1 Two-dimensional tuning : Υ is the assumed change direction and $\tilde{\Upsilon}$ is the actual change direction.

Because $|\alpha - \beta| \leq \pi$, we have

$$\sup_{\alpha} \inf_{\beta \in [\beta_1, \beta_2]} \cos(\alpha - \beta) = \cos\left(\inf_{\alpha} \sup_{\beta \in [\beta_1, \beta_2]} |\alpha - \beta| \right) \tag{10.3.14}$$

Finally, we get

$$\alpha^* = \frac{\beta_1 + \beta_2}{2} \tag{10.3.15}$$

10.3.1.2 Minmax Choice of (θ^*,Υ)

We now discuss the joint optimization of the reference point and the change direction.

Main idea Let us assume that sets Θ_0 and Θ_1 before and after change are two convex nonintersecting sets that can be separated by a linear discriminant function as in figure 7.2. The minmax approach to the optimization of both θ^* and Υ consists of the following two steps :

- choose two least favorable points $\theta_0^* \in \Theta_0$ and $\theta_1^* \in \Theta_1$, namely the closest possible points with respect to the Kullback information, as depicted in figure 10.2; then optimize the ARL function (10.3.5) in these least favorable points;
- then, as is obvious from (10.3.5), at these two points, the optimal choices of the tuning parameters (θ^*, Υ) and of the threshold h can be achieved *separately*. We thus can tune (θ^*, Υ) by optimizing (10.3.5) again, and then compute the threshold exactly as in the case of simple hypotheses that we discussed before.

Minmax tuning as a mathematical programming problem Let us first define what we mean by minmax tuning. We consider again the set K_T of the tuning parameters Δ of a CUSUM change detection algorithm $g(\Delta)$ with alarm time $t_a(\Delta)$:

$$K_T = \left\{ \Delta \ : \ \inf_{\theta \in \Theta_0} \mathbf{E}_\theta[t_a(\Delta)] \geq \bar{T} \right\} \tag{10.3.16}$$

and we define minmax tuning as a search for the algorithm parameters that achieve the minimum value of the worst mean delay in this class :

$$\bar{\tau} = \inf_{\Delta \in K_T} \sup_{\theta_1 \in \Theta_1} \bar{\tau}^*(\theta_1) \tag{10.3.17}$$

Now the expectation of the increment of the decision function is

$$\mathbf{E}_\theta(s) = \Upsilon^T \Sigma^{-1}(\theta - \theta^*) \tag{10.3.18}$$

Using the decomposition $\Sigma^{-1} = (R^{-1})^T R^{-1}$ and the transformed parameter $\bar{\theta} = R^{-1}\theta$, we get

$$\mathbf{E}_{\bar{\theta}}(s) = \bar{\Upsilon}^T(\bar{\theta} - \bar{\theta}^*) = \bar{\Upsilon}^T\bar{\theta} - \Upsilon_0 \tag{10.3.19}$$

where $\bar{\Upsilon} = R^{-1}\Upsilon$ and $\Upsilon_0 = \bar{\Upsilon}^T\bar{\theta}^*$. In other words, we recover here the linear structure. From this it is obvious that the general covariance has no additional characteristic feature with respect to the unit covariance. From now on, we thus consider, without loss of generality, the unit covariance matrix, and we do not keep the distinction between $\bar{\theta}$ and θ. Moreover, because of the optimization to be done, we no longer assume that Υ is a unit vector, which has no consequence because, as is obvious from (10.3.5)-(10.3.7), the ARL function does not depend upon the length of Υ.

Let us now discuss the discriminant surface between Θ_0 and Θ_1. We recall that the expectation of the increment of the decision function satisfies

$$-\mathbf{E}_{\theta_0^*}(s) = \mathbf{E}_{\theta_1^*}(s) \tag{10.3.20}$$

where θ_0^* and θ_1^* are the least favorable points. As we explained when introducing our detectability definition in chapter 6, a conveniently tuned algorithm should be such that

$$\begin{aligned} \sup_{\theta \in \Theta_0} (\Upsilon^T\theta - \Upsilon_0) &\leq -\mu \\ \inf_{\theta \in \Theta_1} (\Upsilon^T\theta - \Upsilon_0) &\geq +\mu \end{aligned} \tag{10.3.21}$$

where μ is any positive constant. On the other hand, optimization of the ARL function can be achieved by searching for the minimum value of the following quadratic form :

$$\Upsilon^T\Upsilon \tag{10.3.22}$$

Therefore, we find that the minmax tuning is equivalent to the following mathematical programming problem :

$$(\Upsilon, \Upsilon_0) = \arg\inf_{\Upsilon} \Upsilon^T\Upsilon \tag{10.3.23}$$

under the constraints

$$\begin{aligned} \sup_{\theta \in \Theta_0} (\Upsilon^T\theta - \Upsilon_0) &\leq -\mu \\ \inf_{\theta \in \Theta_1} (\Upsilon^T\theta - \Upsilon_0) &\geq +\mu \end{aligned} \tag{10.3.24}$$

In general, this problem is quite complex. Let us consider the case where Θ_0 and Θ_1 are convex nonintersecting polyhedra, depicted in figure 10.2. In this situation, the constraints (10.3.24) can be rewritten as

$$\begin{aligned} (\Upsilon^T\theta_i - \Upsilon_0) &\leq -\mu \quad \text{for} \quad 1 \leq i \leq n \\ (\Upsilon^T\theta_j - \Upsilon_0) &\geq +\mu \quad \text{for} \quad 1 \leq j \leq m \end{aligned} \tag{10.3.25}$$

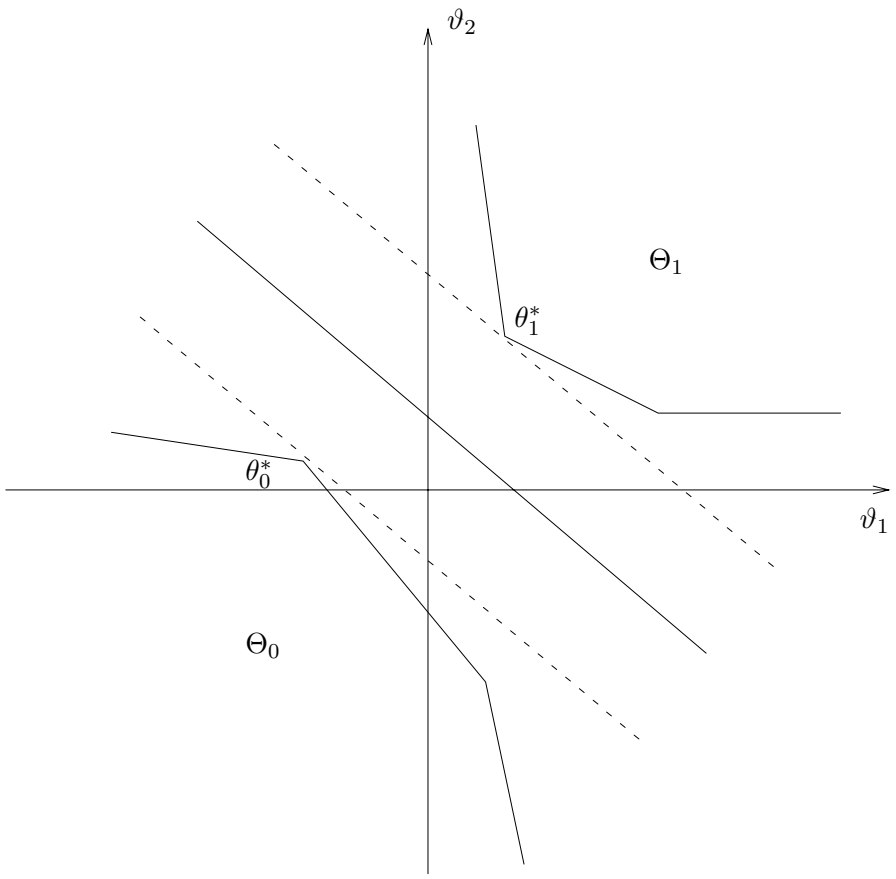

Figure 10.2 Minmax tuning.

where θ_i and θ_j are any vertices of the polyhedral regions Θ_0 and Θ_1, respectively. Therefore, the constraints can be written as

$$\begin{aligned} \Upsilon^T \breve{\Theta}_0 - \Upsilon_0 \mathbb{1}_n^T &\leq -\mu \mathbb{1}_n^T \\ \Upsilon^T \breve{\Theta}_1 - \Upsilon_0 \mathbb{1}_m^T &\geq +\mu \mathbb{1}_m^T \end{aligned} \tag{10.3.26}$$

where $\breve{\Theta}_l$ (l=0,1) is the matrix, the columns of which are made of all the vertices of Θ_l. In other words, we can assume that the least favorable points θ_0^* and θ_1^* are equidistant from the discriminant line (see figure 10.2). The mathematical programming problem (10.3.23)-(10.3.24) can thus be rewritten as

$$(\Upsilon, \Upsilon_0) = \arg \inf_{\Upsilon} \Upsilon^T \Upsilon \tag{10.3.27}$$

under the constraints

$$\begin{aligned} \Upsilon^T \breve{\Theta}_0 - \Upsilon_0 \mathbb{1}_n^T &\leq -\mu \mathbb{1}_n^T \\ \Upsilon^T \breve{\Theta}_1 - \Upsilon_0 \mathbb{1}_m^T &\geq +\mu \mathbb{1}_m^T \end{aligned} \tag{10.3.28}$$

Note that the problem of (10.3.27)-(10.3.28) is now a *quadratic* programming problem for which standard solutions exist.

Geometrical interpretation We continue to assume that $\Sigma = I_\ell$. The ratio (10.3.7) can be rewritten as

$$f(\Upsilon, \theta - \theta^*) = \frac{\Upsilon^T(\theta - \theta^*)}{(\Upsilon^T \Upsilon)^{\frac{1}{2}}} \tag{10.3.29}$$

Its absolute value is thus nothing but

$$|f(\Upsilon, \theta - \theta^*)| = \frac{|\Upsilon^T \theta - \Upsilon_0|}{\|\Upsilon\|} \tag{10.3.30}$$

which is the Euclidean distance between point θ and the discriminant surface. The minmax tuning problem is thus the problem of designing a discriminant surface to maximize the equal distance between this surface and the closest vertices θ_0^* and θ_1^* (see figure 10.2 again). Sometimes it can be useful to modify this tuning by choosing the least favorable points inside the polyhedra. This can be done by using the values μ_0 and μ_1 in the constraints

$$\begin{aligned} \Upsilon^T \breve{\Theta}_0 - \Upsilon_0 \mathbb{1}_n^T &\leq \mu_0 \mathbb{1}_n^T \\ \Upsilon^T \breve{\Theta}_1 - \Upsilon_0 \mathbb{1}_m^T &\geq \mu_1 \mathbb{1}_m^T \end{aligned} \tag{10.3.31}$$

where $\mu_0 < 0 < \mu_1$.

10.3.2 Nonadditive Changes and the Local Case

In this case, the linear local CUSUM decision function has the following increment :

$$s_k = \Upsilon^T Z_k^* \tag{10.3.32}$$

where Z_k^* is the efficient score. As we explained in section 4.1, the mean and variance of the increment s_k can be approximated as

$$\begin{aligned} \mathbf{E}(s) &\approx \Upsilon^T \mathbf{I}(\theta^*)\ (\theta - \theta^*) \\ &\approx \Upsilon^T \mathbf{I}(\theta^*)\ \theta - \Upsilon_0 \qquad (10.3.33) \\ \mathbf{E}(s^2) &\approx \Upsilon^T \mathbf{I}(\theta^*)\ \Upsilon \qquad (10.3.34) \end{aligned}$$

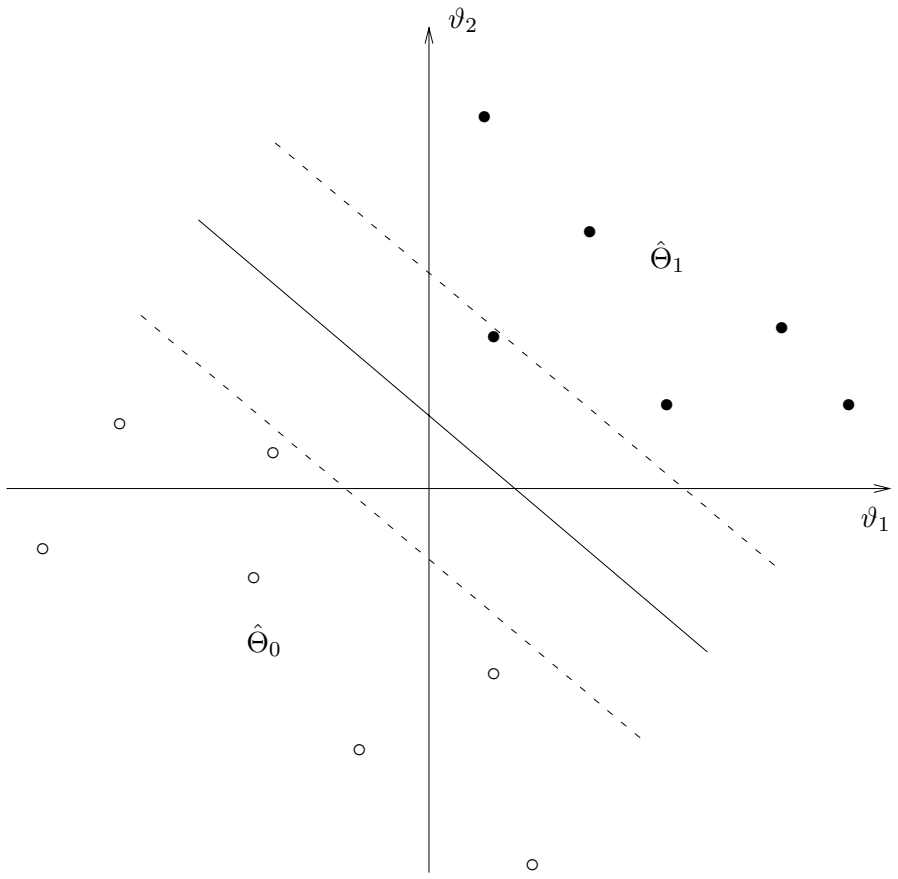

Figure 10.3 Tuning with empirical information.

where $\Upsilon_0 = \Upsilon^T \mathbf{I}(\theta^*)\ \theta^*$. The results of the previous subsection can thus be applied in this nonadditive local situation, replacing Σ^{-1} by $\mathbf{I}(\theta^*)$.

10.3.3 Tuning and Detectability

We now discuss two different but connected topics. The first concerns the situation where parameter sets Θ_0 and Θ_1 are given only in terms of collections of empirical values. The second is aimed at establishing a link between the tuning and robust detectability issues.

10.3.3.1 Tuning with Empirical Information

We assume that, instead of the theoretical knowledge of Θ_0 and Θ_1, two sets of empirical points in $\hat{\Theta}_0$ and $\hat{\Theta}_1$ are available, as depicted in figure 10.3. The problem is then to tune the algorithms by using this empirical information. The first question is whether it is possible to classify these two sets of points using a *linear* discriminant function. In the theory of linear mathematical programming [Gass, 1958], there exists a particular algorithm for answering this question. If the answer is positive, the second problem is the minmax tuning of the change detection algorithm by using these empirical points. The same philosophy as before can be used, replacing the vertices of the sets Θ_0 and Θ_1 by the empirical points in $\hat{\Theta}_0$ and $\hat{\Theta}_1$. It should be clear that some of the empirical points do *not* play any role in constraints (10.3.25) (see figure 10.3 again).

10.3.3.2 Tuning and Robust Detectability

We now discuss the relationships between tuning and detectability. We first show how tuning can influence detectability, and then conversely how a detectability criterion can be used for tuning the algorithm.

As we explained in subsection 7.2.6, a change from $\tilde{\theta}_0$ to $\tilde{\theta}_1$ is said to be detectable by a statistics s tuned with the aid of the parameters θ_0 and θ_1, if

$$\mathbf{E}_{\tilde{\theta}_0}(s) < 0 < \mathbf{E}_{\tilde{\theta}_1}(s) \tag{10.3.35}$$

Let us show that this condition is automatically fulfilled if there exists a linear discriminant function between the two sets Θ_0 and Θ_1, or $\hat{\Theta}_0$ and $\hat{\Theta}_1$. We consider again the Gaussian case with unit covariance matrix. Then the equation of the linear discriminant surface associated with the linear CUSUM algorithm is

$$\mathbf{E}_\theta(s) = \Upsilon^T\theta - \Upsilon_0 = 0 \tag{10.3.36}$$

It results from constraints (10.3.28) that when this linear discriminant function exists, then inequalities (10.3.35) are satisfied.

Let us now discuss the use of a detectability criterion for tuning an algorithm. As we explained before, the tuning process comprises two steps : the tuning of the reference value θ^* and the change direction Υ, and the tuning of the threshold h. When the ARL function cannot be computed, we can use the detectability criterion as a weak performance index for tuning θ^* and Υ, exactly as we do with the ARL function. In this case of course, the threshold should be chosen empirically.

10.4 Vector Case with Quadratic Decision Function

We now discuss the case of changes in multidimensional parameters detected with a decision function that is quadratic. In other words, we discuss the tuning of the χ^2-CUSUM and GLR algorithms, when θ_0 is known and the change magnitude b is also known. We start from the case of additive changes discussed in the subsection 7.2.1, and consider a generalization to the algorithms resulting from local approximations in the case of nonadditive changes.

10.4.1 Additive Changes

We now investigate the case of a quadratic alternative hypothesis :

$$\theta(k) = \begin{cases} \theta_0 & \text{when } k < t_0 \\ \theta \ : (\theta-\theta_0)^T\Sigma^{-1}(\theta-\theta_0) = b^2 & \text{when } k \geq t_0 \end{cases} \tag{10.4.1}$$

as depicted in figure 10.4. In this case, the decision function is

$$t_a = \min\{k \ : g_k \geq h\} \tag{10.4.2}$$

$$g_k = \max_{1\leq j\leq k}(k-j+1)\left(b\,\chi_j^k - \frac{b^2}{2}\right) \tag{10.4.3}$$

where

$$(\chi_j^k)^2 = (\bar{Y}_j^k - \theta_0)^T\Sigma^{-1}(\bar{Y}_j^k - \theta_0) \tag{10.4.4}$$

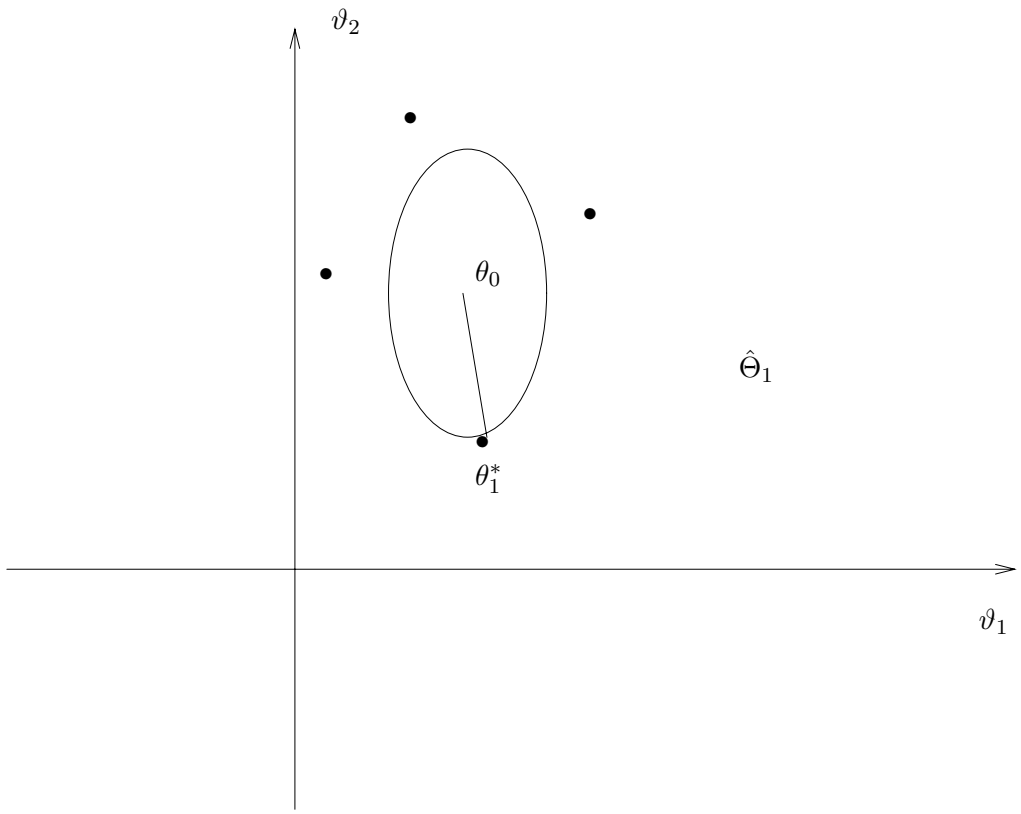

Figure 10.4 Tuning a quadratic algorithm.

The two tuning parameters are b and h. Now, as explained in section 7.3, the mean time between false alarms and the mean delay for detection are asymptotically related through

$$\bar{\tau}^* \sim \frac{\ln \bar{T}}{\mathbf{K}(\theta, \theta_0)} = \frac{2 \ln \bar{T}}{b^2} \tag{10.4.5}$$

where

$$\mathbf{K}(\theta, \theta_0) = \frac{1}{2}(\theta - \theta_0)^T \Sigma^{-1} (\theta - \theta_0) \tag{10.4.6}$$

in the Gaussian case. In this situation, the minmax approach consists of choosing the least favorable point θ_1^* with minimum Kullback distance $\mathbf{K}(\theta_1^*, \theta_0)$, and tuning b accordingly :

$$\theta_1^* = \arg \inf_{\theta \in \Theta_1} \mathbf{K}(\theta, \theta_0) \tag{10.4.7}$$

$$b^2 = 2\, \mathbf{K}(\theta_1^*, \theta_0) \tag{10.4.8}$$

Then, for this b, the quantities $\bar{\tau}$ and $\bar{T}$ can be computed and optimized with respect to the threshold, as in the case of simple hypotheses again.

10.4.2 Nonadditive Changes and the Local Case

In this case, the structure of the decision function is the same as before, but the computation of the χ^2 statistics is based upon the efficient score :

$$(\chi_j^k)^2 = (\bar{Z}_j^k)^T\, \mathbf{I}^{-1}(\theta_0)\, \bar{Z}_j^k \tag{10.4.9}$$

The results obtained in the Gaussian case can be used as a crude approximation in the present local case, as we explained in section 9.5. For this purpose, we again use detectability in terms of the Kullback information as a performance index for tuning, as in (10.4.7), where we use the following approximation of this information :

$$\mathbf{K}(\theta_1, \theta_0) \approx \frac{1}{2}(\theta_1 - \theta_0)^T\, \mathbf{I}(\theta_0)\, (\theta_1 - \theta_0) \tag{10.4.10}$$

10.5 Notes and References

Section 10.1

This section implicitly contains methodological comments extracted from many earlier investigations. To our knowledge, a complete and less philosophical methodology for tuning does not exist.

Section 10.2

The tuning of change detection algorithms in the scalar case is investigated in [Page, 1954c, Van Dobben De Bruyn, 1968, Phillips, 1969, Goel and Wu, 1971, Robinson and Ho, 1978, Montgomery, 1980, Montgomery, 1985, Wetherill and Brown, 1991].

Section 10.3

The tuning of the linear CUSUM algorithm in the vector case is investigated in [Nikiforov, 1983, Nikiforov and Tikhonov, 1986].

The main textbooks devoted to optimization problems are [Gass, 1958, Shapiro, 1979, Fletcher, 1980, MacCormick, 1983, Luenberger, 1984, Polyak, 1987, Pardalos and Rosen, 1987].

11

Applications

In this chapter, we describe applications where typical change detection problems occur. We briefly introduced some of these examples in chapter 1, and the corresponding models are given in the appendix to this chapter.

The main **goals** of this chapter are as follows. First, we discuss several applications of the algorithms introduced in chapter 2 and chapters 7 to 9 for processing real signals. These applications are :

1. fault detection in navigation systems;
2. onset detection in seismic signal processing;
3. automatic segmentation of continuous speech signals;
4. in situ detection of changes in the vibrating characteristics of mechanical systems.

Through these examples, we want to emphasize several issues. We first exhibit the main abilities of the change detection algorithms described in this book. Next we also show how the available theoretical results of the properties of the algorithms can be used in actual situations of signal processing. More precisely, we both show that these theoretical properties can be exhibited in practice, and how they can be used for tuning the design parameters of the algorithms. Finally, we try to extract from these examples methodological points that can be of help in *other* application domains.

The second goal of this chapter is to describe several other possible areas of application of change detection algorithms. Some change detection algorithms have been applied in some of these application domains, but we do not show experimental results. Other problems have not been yet solved with change detection methods, at least to our knowledge, but we think it interesting to indicate how they could be. In doing so, we hope to help the reader interested in a particular area to find a path for reading the book. For this purpose, we discuss the following applications :

1. statistical quality control;
2. biomedical signal processing;
3. fault detection in chemical processes.

We thus subdivide this chapter into two main sections corresponding to our two goals.

11.1 Examples of the Use of Some Algorithms

In this section, we discuss several actual applications of change detection algorithms, and we show experimental results obtained from real data. We describe what can be obtained in practice when using change detection algorithms and how to use the available theoretical results about these algorithms.

11.1.1 Fault Detection in Navigation Systems

In this example, we follow [Newbold and Ho, 1968, Sturza, 1988, Varavva *et al.*, 1988, Kireichikov *et al.*, 1990, Nikiforov *et al.*, 1991, Nikiforov *et al.*, 1993]. A navigation system is a typical equipment for planes, boats and other mobiles. Conventional navigation systems use some measurement sources or sensors. For example, an inertial navigation system has two types of sensors: laser gyros and accelerometers. Using these sensors information and the motion equations, the estimation of the useful signal (the geodesic coordinates and the velocities of the plane, etc.) can be achieved.

In view of safety and accuracy requirements, redundant fault tolerant navigation systems are used. *Fault detection and isolation of faulty sensors* are among the main problems for the design of these navigation systems. We concentrate now on the *fault detection* problem which can be stated as a statistical change detection problem. The criterion to be used is fast detection and few false alarms. Fast detection is definitely necessary because, between the fault onset time and the fault detection time, we use abnormal measurements in the navigation equations, which is obviously highly non desirable. On the other hand, false alarms result in lower accuracy of the estimate because some correct information is not used. The optimal solution is again a tradeoff between these two contradictory requirements.

11.1.1.1 Models of Interest

For this purpose, two models are of interest. For inertial navigation systems, the state-space model A is the most useful. For strapdown reference units and for global navigation sets (GPS), the regression model B is adequate.

- **Model A** : This model can be represented in the following linear discrete time state-space form :

$$\begin{array}{rcl} X_{k+1} & = & F(k+1,k)X_k + W_k + \Upsilon(k,t_0) \\ Y_k & = & H(k)X_k + U_k + V_k \end{array} \tag{11.1.1}$$

  where $X_k \in \mathbf{R}^n$ is the state vector containing the physical errors, $Y_k \in \mathbf{R}^r$ is the INS measurement, $U_k \in \mathbf{R}^r$ is the useful signal (geodesic coordinates, velocities, ...), and $W_k \in \mathbf{R}^n$ and $V_k \in \mathbf{R}^r$ are nonstationary zero mean Gaussian white noises having covariance matrices $Q(k) \geq 0$ and $R(k) > 0$, respectively. The initial state X_0 is a Gaussian zero mean vector with covariance matrix $P_0 > 0$. The change vector $\Upsilon(k,t_0)$ is $\Upsilon(k,t_0) = 0$ for $k < t_0$ and $\Upsilon(k,t_0) \neq 0$ for $k \geq t_0$. The matrices F, H, Q, R, and P_0 are known, and the change time t_0 and change vector Υ are unknown.

- **Model B** : This model can be represented in the following linear form :

$$Y_k = HX_k + V_k + \Upsilon(k,t_0) \tag{11.1.2}$$

  In this case X_k is the unknown input useful signal and H is a constant full rank matrix. We assume that there exists measurement redundancy, namely that $r > n$. The covariance matrix R is scalar: $R = \sigma^2 I_r$, where I_r is the identity matrix of size r.

11.1.1.2 Example of Model A

Modern commercial airplanes are usually equipped with a triplicate strapdown inertial navigation system (INS). This system is made of two types of sensors : laser gyros and accelerometers. The detection of soft drifting-type faults in one of these sensor types is of interest. It is known [Huddle, 1983] that INS errors models can be reduced to the following model.

Let $Y_k(i)$ denote the output of the INS$_i$ ($i = 1, 2, 3$) and assume that only *one INS* can fail simultaneously. In this case the difference $\Delta Y_k^{ij} = Y_k(i) - Y_k(j)$ can be written in the following manner :

$$\left\{ \begin{array}{rcl} \Delta X_{k+1} & = & F(k+1,k)\Delta X_k + W_k + \Upsilon(k,t_0) \\ \Delta Y_k^{ij} & = & H(k)\Delta X_k + V_k \end{array} \right. \tag{11.1.3}$$

where $\Delta X_k = X_k(i) - X_k(j)$, $(W_k)_{k\geq 0}$ and $(V_k)_{k\geq 1}$ have covariance matrices $2Q(k)$ and $2R(k)$ respectively. $\Upsilon(k,t_0)$ is the vector of bias in one sensor error :

$$\Upsilon(k,t_0) = \left\{ \begin{array}{lll} 0 & \text{if} & k < t_0 \\ \Upsilon & \text{if} & k \geq t_0 \end{array} \right. , \quad \Upsilon = (0,\ldots,0,\nu,0,\ldots,0)^T$$

Therefore, for fault detection we have to compute three differences Y^{12}, Y^{13}, Y^{23} and three Kalman filter innovations ϵ^{12}, ϵ^{13}, ϵ^{23} by using system (11.1.3), and then we have to detect changes in each of the innovation sequences $(\epsilon_k^{12})_{k\geq 1}$, $(\epsilon_k^{13})_{k\geq 1}$, $(\epsilon_k^{23})_{k\geq 1}$ by using change detection algorithms.

Let us now concentrate on a simple but representative example, and compare three change detection algorithms. We consider the following state-space system ($n = 2, r = 1$) which is a simplified model of the inertial system heading gyro error:

$$\begin{array}{lll} X = \left(\begin{array}{c} x_1 \\ x_2 \end{array} \right) & F = \left(\begin{array}{cc} 1 & \delta \\ 0 & 1 - \frac{\delta}{T_g} \end{array} \right) & H = \left(\begin{array}{cc} 1 & 0 \end{array} \right) \\ R = (\sigma_V^2) & Q = \left(\begin{array}{cc} 0 & 0 \\ 0 & \sigma_W^2 \end{array} \right) & \Upsilon(k,t_0) = \left(\begin{array}{c} \nu(k,t_0) \\ 0 \end{array} \right) \end{array} \tag{11.1.4}$$

where δ is the sampling period, T_g is the gyro error time constant, and $\delta \ll T_g$. This type of model is discussed in [Newbold and Ho, 1968].

Let us compare the GLR and CUSUM algorithms together with a specific detection algorithm, based upon the Kalman filter state estimate, which was introduced in [Kerr, 1982]. For this comparison, we assume that the Kalman gain is a constant value; in other words, we assume that the steady state is reached. In order to avoid a *dynamic profile* in the innovation $(\varepsilon_k)_k$ and to *simplify* our comparison, we have to assume that the gyro fault can be modeled as the dynamic profile $\nu(k,t_0)$ such that its signature on the Kalman filter innovation is a *step*, as in the following equation :

$$\mathcal{L}(\varepsilon_k) = \left\{ \begin{array}{ll} \mathcal{N}(0,1) & \text{for } k < t_0 \\ \mathcal{N}(\nu,1) & \text{for } k \geq t_0 \end{array} \right.$$

where $|\nu|$ is the jump magnitude.

Algorithms We investigate three detection algorithms based on this model.

The *first* is the two-sided CUSUM algorithm :

$$\begin{array}{rcl} t_a & = & \min\{k : (g_k^+ \geq h) \cup (g_k^- \geq h)\} \\ g_k^+ & = & \left(g_{k-1}^+ + \varepsilon_k - \dfrac{|\nu|}{2} \right)^+ \\ g_k^- & = & \left(g_{k-1}^- - \varepsilon_k - \dfrac{|\nu|}{2} \right)^+ \end{array} \tag{11.1.5}$$

The *second* is defined in the following manner [Kerr, 1982]. The Kalman filter estimate can be computed with the aid of the recursive equation

$$\hat{X}_{k|k} = F\hat{X}_{k-1|k-1} + K_k\,\varepsilon_k$$

where K_k is the Kalman gain. In our case, it turns out that it is relevant to use the second component $(\hat{x}_2)_{k|k}$:

$$(\hat{x}_2)_{k|k} = (1-\alpha)\ (\hat{x}_2)_{k-1|k-1} + k_2\ \varepsilon_k \tag{11.1.6}$$

where $\alpha = \frac{\delta}{T_g}$ and k_2 is the second component of the Kalman gain K_k. Therefore the second stopping time is :

$$\begin{aligned} t_a &= \min\{k : |(\hat{x}_2)_{k|k}| \geq h_1\} \\ (\hat{x}_2)_{k|k} &= (1-\alpha)\ (\hat{x}_2)_{k-1|k-1} + k_2\ \varepsilon_k \end{aligned} \tag{11.1.7}$$

Note that, up to a change in the scale of ε, this algorithm is nothing but the geometric moving average algorithm (GMA).

The *third* algorithm is the GLR algorithm :

$$\begin{aligned} t_a &= \min\{k : g_k \geq h_2\} \\ g_k &= \max_{1\leq j\leq k} \frac{1}{k-j+1} \left(\sum_{i=j}^{k} \varepsilon_i\right)^2 \end{aligned} \tag{11.1.8}$$

Criteria for Comparison As we explained more formally in section 4.4, the relevant criteria for performance evaluation of change detection algorithms are the mean time between false alarms $\bar{T}$ (4.4.1) and the mean delay for detection $\bar{\tau}$ (4.4.2). Usually algorithms are evaluated by comparing their mean delay for detection for a *given* mean time between false alarms. Recall that there exist several definitions of the mean delay for detection, one of which is the worst mean delay $\bar{\tau}^*$ (4.4.3).

For proving that the CUSUM algorithm is better than the GMA, it is relevant to show that the *worst* mean delay $\bar{\tau}^*_{\text{CUSUM}}$ is less than the worst mean delay $\bar{\tau}^*_{\text{GMA}}$. In fact, for a wide range of values of α, we show that $\bar{\tau}^*_{\text{CUSUM}}$ is less than the mean (and not worst) delay $\bar{\tau}_{\text{GMA}}$, which is a stronger property. For the comparison between the CUSUM and GLR algorithms, we use the worst mean delay for both algorithms, basically because there does not exist a uniformly better algorithm for all possible change magnitudes.

Comparison Between CUSUM and GMA Let us first compare the two-sided CUSUM and the GMA algorithms. For this comparison, we assume that the change magnitude $|\nu|$ is known, but not the sign of ν. For the CUSUM algorithm, it is possible to compute $\bar{\tau}^*$ and $\bar{T}$ by using the numerical solution of the Fredholm integral equation, but for this comparison it is easier to use bounds for these quantities. Because we want to show an advantage of the two-sided CUSUM algorithm, we need to know the upper bound for $\bar{\tau}^*_{\text{CUSUM}}$ and the lower bound for $\bar{T}_{\text{CUSUM}}$ given in the subsection 5.5.1, which we write as

$$\bar{\tau}^*_{\text{CUSUM}} \leq \frac{2\ h}{|\nu|} + \frac{2\ \varphi\left(\frac{|\nu|}{2}\right)}{|\nu|\ \phi\left(\frac{|\nu|}{2}\right)} + 1 \tag{11.1.9}$$

$$2\ \bar{T}_{\text{CUSUM}} \geq 2\ \frac{e^{|\nu|h} - 1 - |\nu|h}{|\nu|^2} - \frac{2\ \varphi\left(-\frac{|\nu|}{2}\right)}{|\nu|\ \phi\left(-\frac{|\nu|}{2}\right)} + 1 \tag{11.1.10}$$

The properties of the GMA algorithm are as follows. It is well known that for INS systems where the gyro error time constant T_g is large, the relevant value of the constant α in (11.1.6) is close to 0. The computation of $\bar{T}$ and $\bar{\tau}$ for GMA is done in [Robinson and Ho, 1978] for values of α greater than or equal

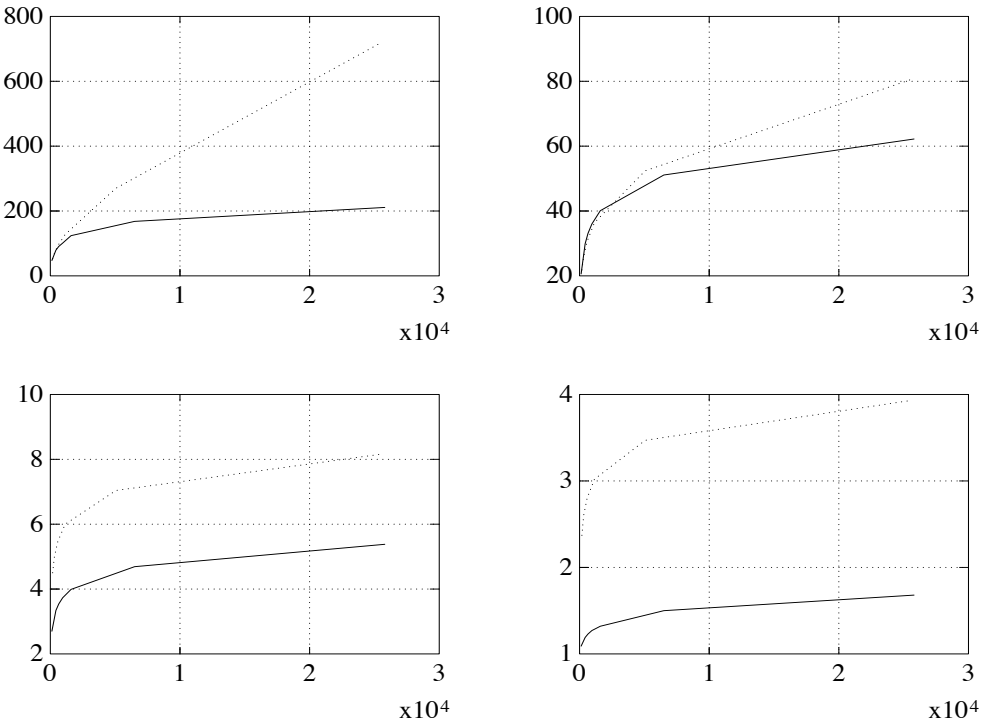

Figure 11.1 Comparison between the CUSUM (solid lines) and GMA (dotted lines) algorithms, $\alpha = 0.05$. Delays $\bar{\tau}^*_{\text{CUSUM}}(\bar{T})$ and $\bar{\tau}_{\text{GMA}}(\bar{T})$ as functions of the mean time between false alarms, for $|\nu| = 0.25$ (upper left); $|\nu| = 0.5$ (upper right); $|\nu| = 2$ (lower left); $|\nu| = 4$ (lower right) .

to 0.05. In this paper, the mean (and not worst) delay is computed by assuming a stationary distribution for the decision function $(\hat{x}_2)_{t_0-1|t_0-1}$ just before the change time t_0. To extend these results to lower values of α, let us consider the limit case $\alpha = 0$. In this case, the filter equation (11.1.6) is in fact a cumulative sum for which we can use the formula giving the bounds for the ASN in sequential analysis discussed in subsection 4.3.2. For the comparison with the two-sided CUSUM algorithm, we need to know the lower bound for $\bar{\tau}^*_{\text{GMA}}$ and the upper bound for $\bar{T}_{\text{GMA}}$. It results from (4.3.74)-(4.3.75) that

$$\begin{aligned}
\bar{\tau}^*_{\text{GMA}} &\geq \max_{0\leq\epsilon\leq h_1} \left\{ \frac{h_1+\epsilon}{|\nu|} - \left[\frac{2h_1}{|\nu|} + \frac{\varphi(|\nu|)}{|\nu|\ \phi(-|\nu|)} - 1 \right] \bar{Q}(|\nu|) \right\} \\
\text{where } \bar{Q}(|\nu|) &= \frac{\phi(-|\nu|)\ e^{-2(h_1+\epsilon)|\nu|} - \phi(|\nu|)}{\phi(-|\nu|)\ e^{-2(h_1+\epsilon)|\nu|} - \phi(|\nu|)\ e^{2(h_1-\epsilon)|\nu|}} \\
\bar{T}_{\text{GMA}} &\leq h_1^2 + 1 + \frac{4h_1}{\sqrt{2\pi}}.
\end{aligned}$$

The results of this comparison are presented in figure 11.1 for $\alpha = 0.05$ and in figure 11.2 for $\alpha = 0,$ where the functions $\bar{\tau}(\bar{T})$ and $\bar{\tau}^*(\bar{T})$ are depicted for $|\nu| = 0.25, 0.5, 1, 2,$ and 4. They show that the two-sided CUSUM is more efficient in all cases except when $\alpha = 0.05, |\nu| = 0.25; 0.5$ and $\bar{T} < 10^3$. In these latter cases, the efficiency of the two-sided CUSUM and the GMA algorithms is approximately the same.

Comparison Between CUSUM and GLR In the previous comparison, we compared the GMA and CUSUM algorithms in the situation where the assumed and actual values of the change magnitudes are

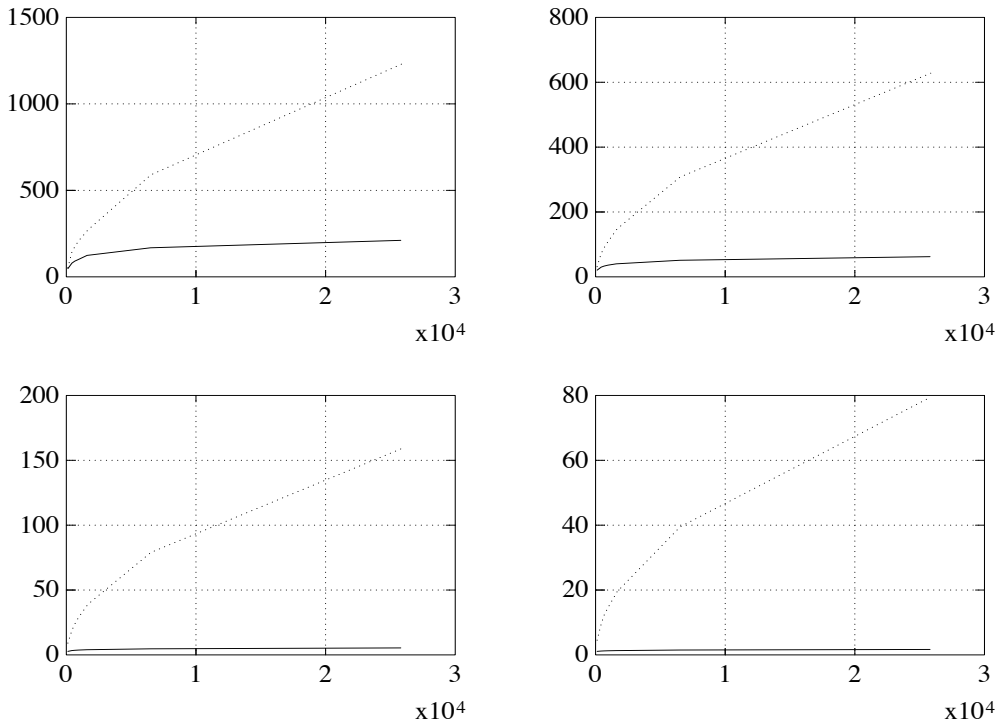

Figure 11.2 Comparison between the CUSUM (solid lines) and GMA (dotted lines) algorithms, $\alpha = 0$. Delays $\bar{\tau}^*(\bar{T})$ as functions of the mean time between false alarms, for $|\nu| = 0.25$ (upper left); $|\nu| = 0.5$ (upper right); $|\nu| = 2$ (lower left); $|\nu| = 4$ (lower right) .

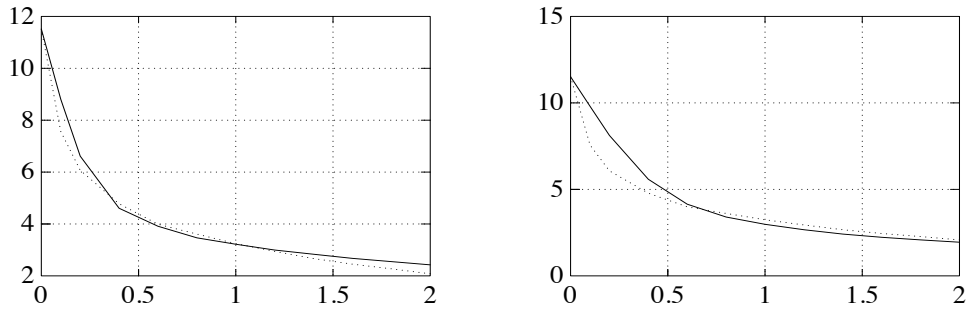

Figure 11.3 Comparison between the CUSUM (solid lines) and GLR (dotted lines) algorithms; logarithm of the ARL function $\ln \bar{\tau}^*(\tilde{\nu})$ as a function of the actual change magnitude $\tilde{\nu}$, for $\bar{T} = 10^5$. The CUSUM algorithm uses the assumed change magnitudes $\nu = 0.5$ (left) and $\nu = 1$ (right).

the same. Now, let us compare the one-sided GLR and two one-sided CUSUM algorithms designed with different *assumed* change magnitudes $\nu = 0.5$ and $\nu = 1$. We make this comparison for different values of the *actual* and *a priori* unknown change magnitude $\tilde{\nu}$, for a given value of the mean time between false alarms $\bar{T}$.

The result of this investigation can be seen in figure 11.3. For the CUSUM algorithm, the worst mean delay $\bar{\tau}^*_{\text{CUSUM}}$ and the mean time between false alarms $\bar{T}_{\text{CUSUM}}$ satisfy the Fredholm integral equation. This equation can be solved by the numerical method described in subsection 5.5.1. The function $\bar{\tau}^*_{\text{GLR}}(\bar{T}_{\text{GLR}})$ for the one-sided GLR can be estimated with the aid of the asymptotic formulas (5.3.18). The mean time between false alarms is chosen to be $\bar{T}_{\text{GLR}} = \bar{T}_{\text{CUSUM}} = 10^5$. As can be seen in figure 11.3, the CUSUM is slightly more efficient than the GLR around the optimal change magnitude, namely when $|\nu - \tilde{\nu}| < 0.4 \div 0.8$, and is less efficient in the converse case. This can be viewed as a lack of robustness of the CUSUM algorithm. But, as discussed in chapter 2, the GLR algorithm can be approximated by two CUSUM algorithms. To reach some tradeoff between complexity and efficiency of the algorithms, when the range of the possible change magnitude is wide, it is useful to use this approximation.

11.1.1.3 Example of Model B

Two typical applications are now discussed : skewed axis strapdown inertial reference units (SIRU), which contain several sensors (such as six single degrees of freedom gyros and six accelerometers), and global satellite navigation sets. Fault detection in these systems is of primary interest for reliability reasons [Sturza, 1988, Jeerage, 1990].

The measurement model of a SIRU can be represented as a regression model (11.1.2), where the physical quantities X should be considered as a nuisance parameter. This model basically represents the redundancy that exists in the SIRU, because the six sensor axes ($r = 6$) are symmetrically distributed with respect to the main axis of a cone, and because the physical parameters (accelerations, etc.) are three dimensional ($n = 3$).

The measurement model of a global navigation set can also be represented as a regression model (11.1.2) [Sturza, 1988]. Conventional global navigation sets require measurements from four satellites to estimate the three spatial coordinates and time ($n = 4$). Because for 18-satellite global navigation sets, five or more satellites ($r \geq 5$) are visible 99.3% of the time, it is possible to provide integrity monitoring by using these redundant measurements.

In [Kireichikov *et al*., 1990, Nikiforov *et al*., 1991, Nikiforov *et al*., 1993], it has been shown that the SIRU fault detection problem, stated as an additive change detection problem in a regression model, can then be solved with the aid of the χ^2-CUSUM and the GLR algorithms corresponding to a known change magnitude but an unknown change direction, as explained in subsection 7.2.2. In [Sturza, 1988] the detection problem is solved with the aid of the minmax approach described in subsection 4.2.8 applied to fixed size samples of measurements. We call this algorithm a χ^2-Shewhart chart because it is based upon a quadratic form of the residuals.

Let us thus now compare the χ^2-Shewhart chart and the χ^2-CUSUM algorithm. Both these algorithms are based upon the transformation from the observations Y_k to the residuals e_k of the LS algorithm :

$$\begin{aligned} e_k &= Y_k - H\hat{X}_k \\ \hat{X}_k &= \arg\min_X \|Y - HX\|^2 \end{aligned} \tag{11.1.11}$$

as explained in subsection 7.2.2. Therefore, we can replace the original problem by the problem of detecting a change in the noncentrality parameter b^2 of a χ^2 distribution with $r - n$ degrees of freedom.

Criteria for Comparison As we explained in the case of the previous model, the relevant criteria for comparison are the mean time between false alarms $\bar{T}$ and the mean delay for detection $\bar{\tau}$. We consider here the mean delay $\bar{\tau}$, under the assumption that the change occurs at the first sample point, and the worst mean delay $\bar{\tau}^*$. See section 4.4 for more formal definitions of these delays. Note that for the χ^2-Shewhart we use both delays, but for the χ^2-CUSUM algorithm we use only $\bar{\tau}^*$.

Comparison Between χ^2-Shewhart and χ^2-CUSUM The results of this comparison are summarized in tables 11.1 and 11.2. Recall that a Shewhart chart (5.1.2) has two tuning parameters : the sample size N and the threshold λ; and its properties are given by

$$\bar{T}(N, \lambda) = \frac{N}{1 - \mathbf{P}[\chi^2(r-n) < \lambda]} \tag{11.1.12}$$

$$\bar{\tau}(N, \lambda) = \frac{N}{1 - \mathbf{P}[\chi'^2(r-n, b^2) < \lambda]} \tag{11.1.13}$$

where $\chi'^2(r-n, b^2)$ is a χ^2 distributed random variable with $r-n$ degrees of freedom and noncentrality parameter b^2. The tuning of the χ^2-CUSUM algorithm depends only upon the threshold. Therefore, assuming $b^2 = 1$ and using several values of the number of degrees of freedom $r - n$, we compare the two algorithms in two different ways.

First, we use a *given* sample size $N = 10$ as in [Sturza, 1988] for navigation systems integrity monitoring. We fix a mean time between false alarms $\bar{T}$, deduce λ and compute $\bar{\tau}$ for χ^2-Shewhart chart. For the χ^2-CUSUM algorithm, we use the asymptotic formula (7.3.56) for computing $\bar{\tau}^*$ for this value of $\bar{T}$. Note that the asymptotic properties of the χ^2-CUSUM algorithm do not depend upon the number of degrees of freedom. These results are in table 11.1. It is obvious that even the worst mean delay $\bar{\tau}^*$ of the χ^2-CUSUM algorithm is significantly lower than the mean delay $\bar{\tau}$ of the χ^2-Shewhart chart for all the values of $\bar{T}$. If we were to compare the two algorithms using the single worst mean delay $\bar{\tau}^*$ for both of them, the advantage of the χ^2-CUSUM over the χ^2-Shewhart would be even greater.

Second, we fix a mean time between false alarms $\bar{T}$, and deduce the optimal values of N and λ by minimizing the above expression of $\bar{\tau}$ for χ^2-Shewhart chart. The corresponding results are shown in table 11.2. In this table, we also add the values of the optimal sample size N_{opt} in the case $r - n = 4$. Furthermore, we also add in the third column of this table the worst mean delay $\bar{\tau}^*$ for the χ^2-Shewhart chart for the case

Table 11.1 Comparison between the χ^2-Shewhart chart with nonoptimal sample size ($N = 10$) and the χ^2-CUSUM algorithm.

$\bar{T}$	χ^2-Shewhart		χ^2-CUSUM
	$\bar{\tau}_{\text{Shew}}$	$\bar{\tau}_{\text{Shew}}$	$\bar{\tau}^*_{\text{CUSUM}}$
	$r - n = 3$	$r - n = 4$	$r - n \geq 1$
10^2	11.8	12.3	9.2
10^3	18.5	20.6	13.8
10^4	35.9	42.7	18.4
10^5	82.4	103.9	23.0
10^6	214.4	284.3	27.6
10^7	612.7	845.2	32.2
10^8	1854.0	2482.0	36.8

Table 11.2 Comparison between the χ^2-Shewhart chart with optimal sample size and the χ^2-CUSUM algorithm.

$\bar{T}$	χ^2-Shewhart					χ^2-CUSUM	
						(7.3.56)	Fredholm
	$r - n = 1$		$r - n = 4$		$r - n = 10$	$r - n \geq 1$	$r - n = 1$
	$\bar{\tau}_{\text{Shew}}$	$\bar{\tau}^*_{\text{Shew}}$	$\bar{\tau}_{\text{Shew}}$	N_{opt}	$\bar{\tau}_{\text{Shew}}$	$\bar{\tau}^*_{\text{CUSUM}}$	$\bar{\tau}^*_{\text{CUSUM}}$
10^3	13.8	20.5	19.3	15	24.9	13.8	11.9
10^4	20.4	31.1	27.6	22	35.2	18.4	16.5
10^5	27.4	42.1	35.8	28	45.0	23.0	21.1
10^6	34.5	53.4	43.9	35	54.5	27.6	25.0
10^7	41.7	64.7	52.0	41	63.7	32.2	
10^8	49.6	76.1	59.8	44	72.8	36.8	

$r - n = 1$, and in column 8 the "exact" delay (Fredholm integral equations) for the χ^2-CUSUM algorithm for the case $r-n = 1$. This last additional column shows that, for the χ^2-CUSUM algorithm, the asymptotic value of the delay is relatively accurate, at least for $r - n = 1$. The comparative results can be summarized as follows. The χ^2-Shewhart with optimal sample size again has a greater delay than the χ^2-CUSUM, for all the values of $\bar{T}$. Comparing the columns corresponding to $r - n = 4$ in tables 11.1 and 11.2, we deduce that the sample size plays a key role in the performance of the χ^2-Shewhart chart. We also find that the optimal sample size should increase with the mean time between false alarms. Finally, it results from the column $r - n = 1$ of table 11.2 that the difference between the worst mean delay $\bar{\tau}^*$ and the mean delay $\bar{\tau}$ of the χ^2-Shewhart is significant.

11.1.2 Onset Detection in Seismic Signal Processing

As explained in chapter 1, the *in situ* estimation of the geographical coordinates and other parameters of earthquakes is often of crucial importance [Kushnir *et al.*, 1983, Morita and Hamaguchi, 1984, Nikiforov and Tikhonov, 1986, Pisarenko *et al.*, 1987, Nikiforov *et al.*, 1989, Mikhailova *et al.*, 1990, Tikhonov *et al.*, 1990]. We consider here the case where the available measurements are three-dimensional signals from one seismic station, as depicted in figure 11.4.

11.1.2.1 Physical Background

The physical framework of this problem is depicted in figure 11.5. The standard sensor equipment of a seismic station results in the availability of records of seismograms with three components, namely the east-west (EW), north-south (NS), and vertical (Z) components. When an earthquake arises, the sensors begin to record several types of seismic waves, the more important of which are the P-wave and the S-wave. Because the P-wave is polarized in the source-to-receiver direction, namely from the epicenter of the earthquake to the seismic station, it is possible to estimate the source-to-receiver azimuth α using the linear polarization of the P-wave in the direction of propagation of the seismic waves. It is known that the different waves have different speeds of propagation. Therefore, the source-to-receiver distance d can be approximately computed by using the following simple equation :

$$d = V_{S-P}\,(t_S - t_P) + d_0 \tag{11.1.14}$$

where V_{S-P} is the speed of the "artificial" wave $S - P$, t_S and t_P are the onset times of the S and P waves, respectively, and d_0 is a known constant. The values of V_{S-P} and d_0 depend upon the seismic properties of the considered region, and also upon the depth of the hypocenter, the mean value of which is assumed to be known. Therefore, if we know the delay $t_S - t_P$ between the onset times of the S-wave and the P-wave (see figure 11.4) and the source-to-receiver azimuth, it is possible to estimate the source-to-receiver distance and the geographical coordinates of the earthquake epicenter using this formula.

It results from this discussion that the estimation of the earthquake coordinates requires the estimation of the onset times of the P and S waves. As we explained in section 8.7 when discussing the off-line change time estimation issue, this problem can be split into three tasks :

- on-line detection and identification of the seismic waves;
- off-line estimation of the onset times of these waves;
- off-line estimation of the azimuth using correlation between components of P-wave segments.

We consider only the first two tasks.

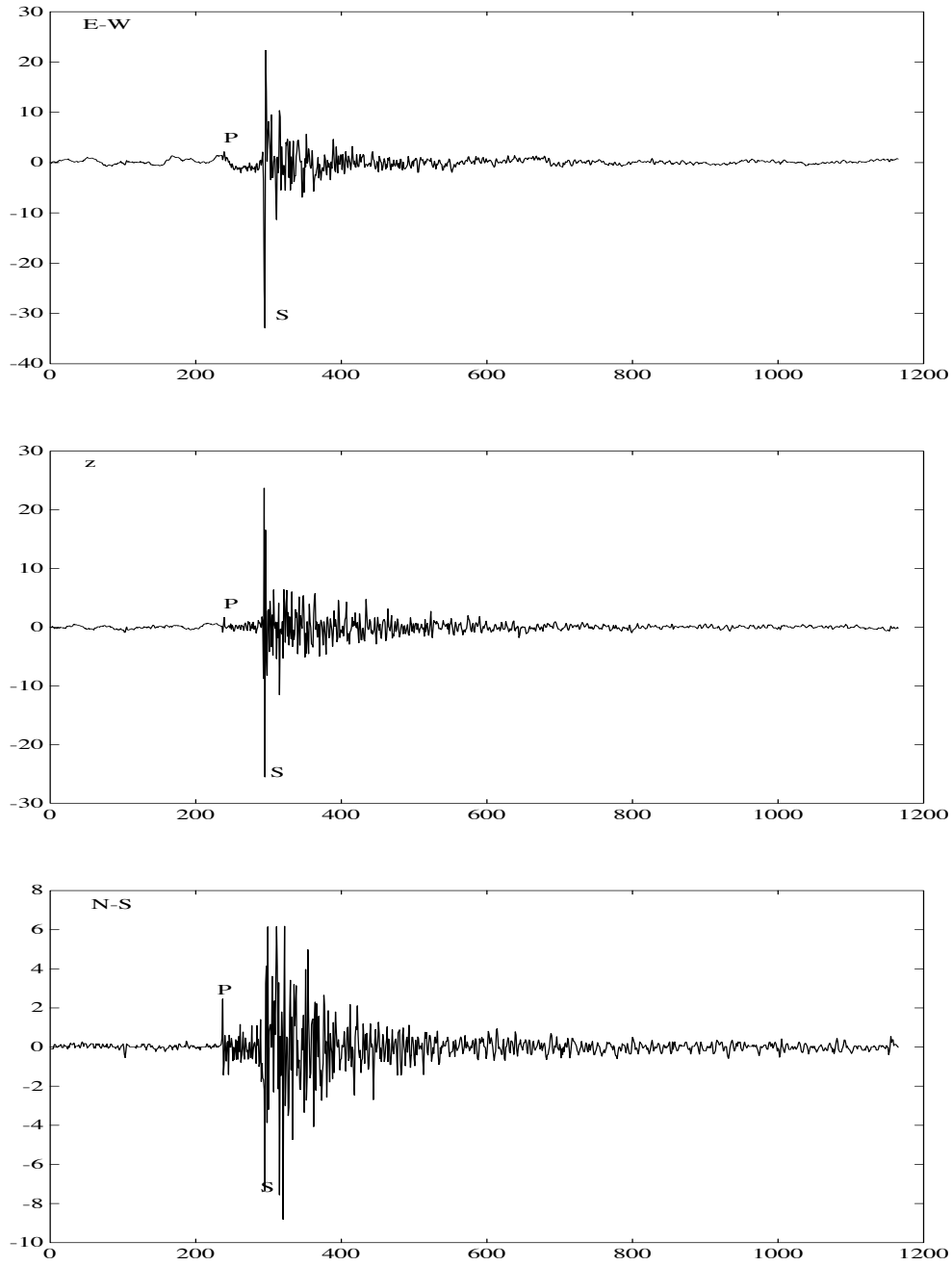

Figure 11.4 A three-dimensional seismogram (Courtesy of the Academy of Sciences of USSR, Far Eastern Scientific Center, Institute of Sea Geology and Geophysics).

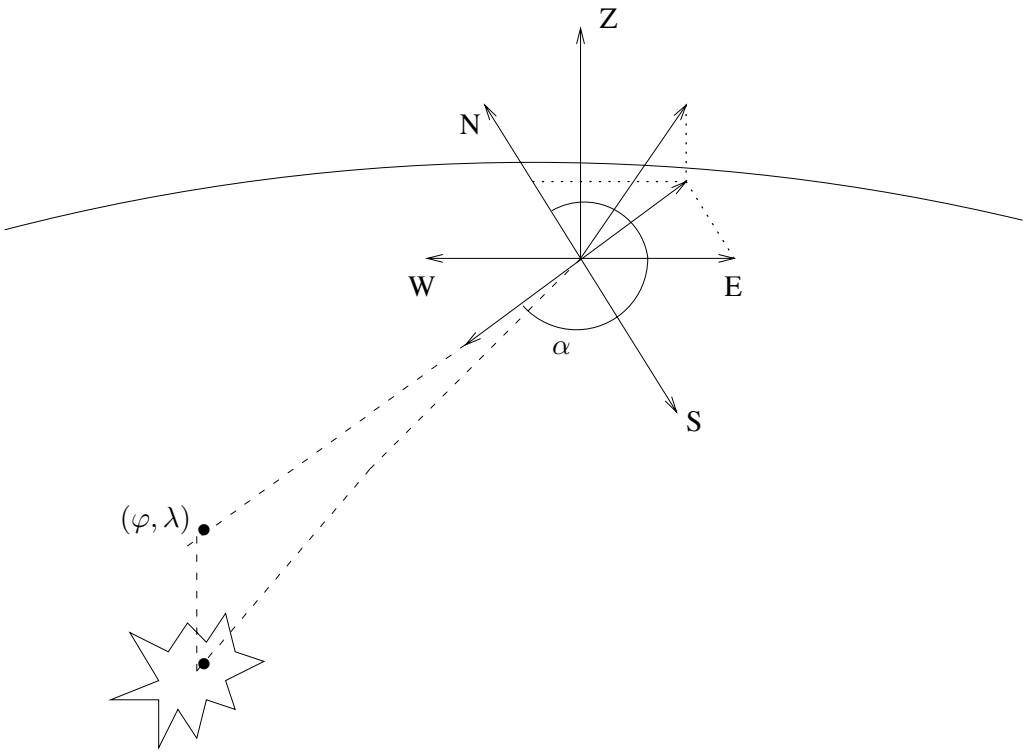

Figure 11.5 The physical background in seismic data processing.

11.1.2.2 Onset Time Detection and Estimation

The P-wave detection has to be achieved *very quickly with a small false alarms rate*. The main reason for this is to allow also S-wave detection in this on-line processing. The P-wave detection is a difficult problem, because the data contain many other nuisance signals coming from the environment of the seismic station, and discriminating between these events and a true P-wave is not easy, as is obvious in figure 11.4. The same situation holds for the S-wave, as can be seen in figures 11.6 and 11.7. The difficulty then is even greater, because of low signal-to-noise ratio and numerous nuisance signals between the P-wave and S-wave. The local and regional earthquakes shown in figures 11.6 and 11.7 raise the most difficult onset detection problems, basically because they correspond to the smallest source-to-receiver distances d ($d < 300$ km and $300 \leq d < 2000$ km, respectively). In these cases, the size of the source is no longer negligible with respect to d, and the difference between the onset times t_P and t_S is small, which makes the presence of the nuisance waves much more critical.

After P-wave and S-wave detection, the *off-line estimation of onset times* is done for both types of waves. We use some fixed size samples of the three-dimensional signals, centered at a rough estimate of the onset time provided by the detection algorithm, as explained in section 8.7.

Problem statement For the on-line detection and the off-line estimation of the P-wave onset times, two types of signal characteristics can be used : either the polarization or the spectral properties. Because the azimuth of the earthquake is unknown, the on-line detection is much more easily achieved when using the spectral properties. Therefore, the following on-line change detection problem turns out to be of interest for this purpose. For computational reasons, we consider parallel processing of the three components separately.

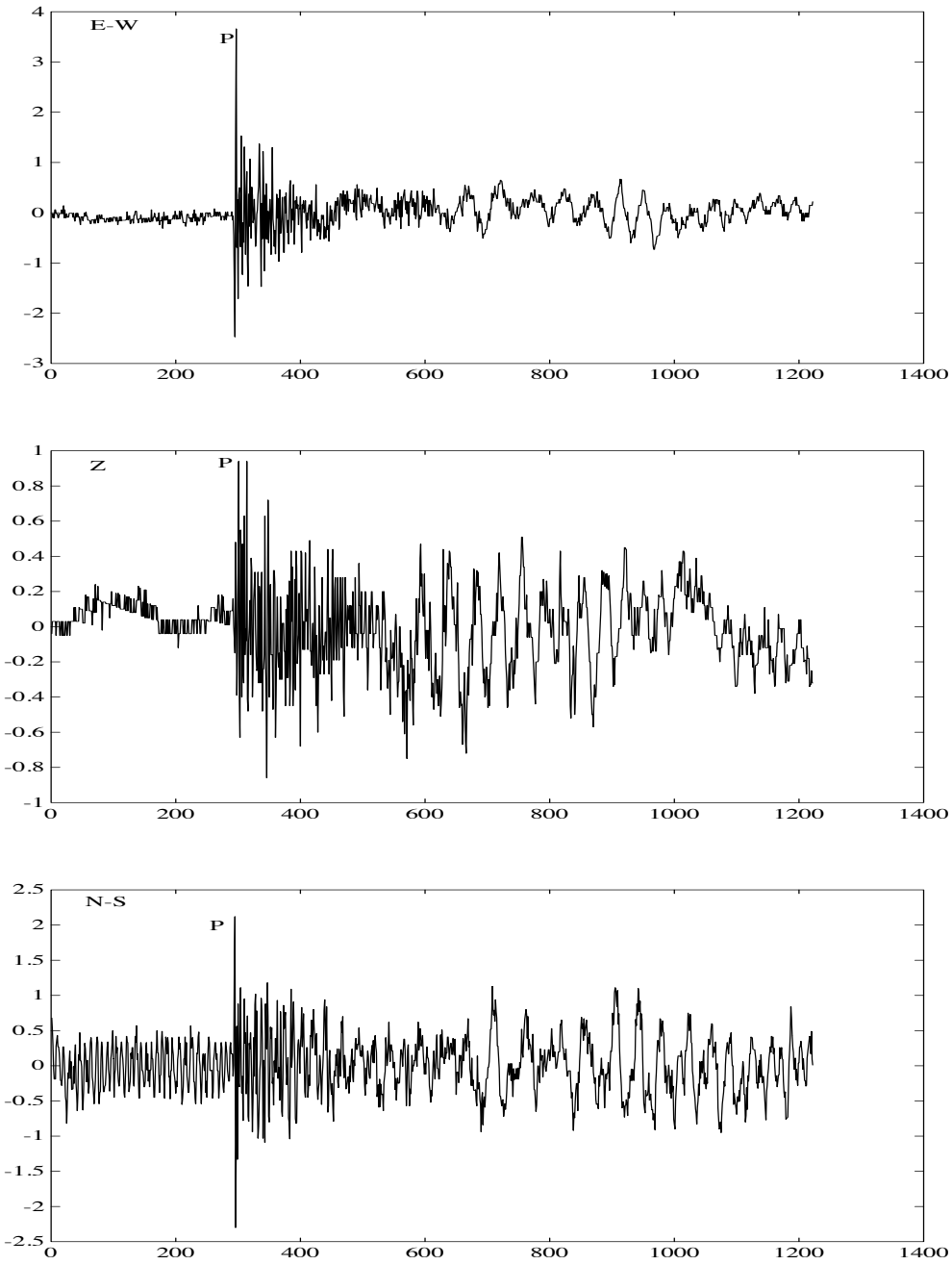

Figure 11.6 Seismogram of a local earthquake (Courtesy of the Academy of Sciences of USSR, Far Eastern Scientific Center, Institute of Sea Geology and Geophysics).

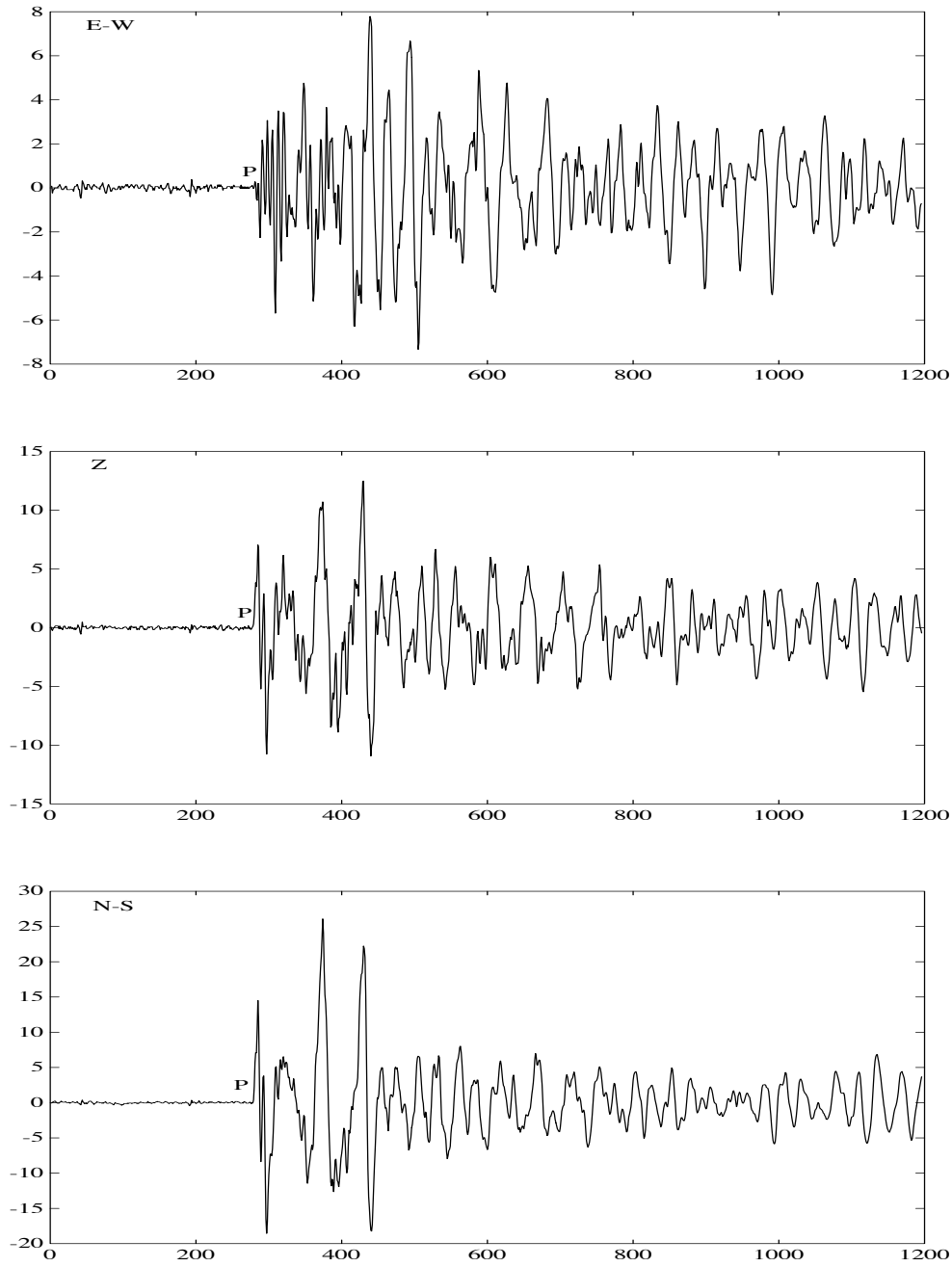

Figure 11.7 Seismogram of a regional earthquake (Courtesy of the Academy of Sciences of USSR, Far Eastern Scientific Center, Institute of Sea Geology and Geophysics).

As usual in this book, we assume that only one change has to be detected at a time, and we consider a scalar zero-mean signal $(y_k)_k$ described by the AR model :

$$y_k = \sum_{i=1}^{p} a_i^{(k)} y_{k-i} + v_k, \quad \mathrm{var}(v_k) = \sigma_k^2 \tag{11.1.15}$$

where

$$\begin{array}{llll} a_i^{(k)} = a_i^0 & \text{and} & \sigma_k^2 = \sigma_0^2 & \text{for} \quad k \leq t_0 - 1 \\ a_i^{(k)} = a_i^1 & \text{and} & \sigma_k^2 = \sigma_1^2 & \text{for} \quad k \geq t_0 \end{array}$$

and where $(v_k)_k$ is a white noise sequence. In other words, the AR coefficients and the variance of the excitation may change at an unknown onset time, and the problem of interest is the on-line detection of such a change, and the estimation of the change time t_0 and possibly of the AR parameters before and after change.

On-line P-wave detection Even in the nonobvious situations, such as the ones mentioned above and those depicted in figure 11.8, the main critical issue for on-line detection of the P-wave is the false alarms rate, and not the missed detection. The first reason for this situation is the presence of the seismic noise which is highly nonstationary, because it depends upon various environment conditions, such as the weather, the technical activity around the station, and the state of the sea. The second reason is related to the practical consequences of a false alarm, which is related both to the psychological aspects of man/machine interaction and to the increased risk of missed detection of the next wave. To minimize the false alarm rate, the following solution is of interest. First, the reference AR model is estimated inside fixed size sliding windows (not necessarily overlapping). Second, the local quadratic CUSUM or GLR algorithm for detecting changes in the AR coefficients and the input variance is used together with a *robust* tuning of its free parameters, namely high values of the minimum magnitude of change in terms of the Kullback information (see (8.2.46)) and of the threshold. Finally, a simple logic is used to merge the results of the three parallel processings, and to obtain only one change time estimate by using the median of the three individual estimates. The results of this processing are shown in figures 11.8 and 11.9.

Off-line P-wave onset time estimation As explained before, the estimated change time resulting from the on-line detection algorithm can be used as a starting point for an off-line change time estimation algorithm, usually called *a posteriori* processing. This processing consists of using, on each component, a fixed size data window centered at the previous estimate, inside which a scalar autoregressive model is identified. Then, because of the additivity of the likelihood function and assuming the independence of the three components, we compute the off-line joint likelihood of the change to estimate t_P. The most critical issue in this off-line processing is the presence, in the data window and *after* the onset time of the P-wave, of several waves that are intermediate between the P- and S-waves, which can have a signal-to-noise ratio greater than that of the P-wave. Therefore, to avoid a wrong onset time t_P estimation, we use a special frequency filtering of each of the three components before this off-line processing [Nikiforov *et al.*, 1989]. Another difficulty can result from the possible vanishing of the seismic signal at the end of the considered data window. For this reason, the scalar autoregressive models are identified with the aid of an additional constraint (lower bound) on the estimated input variance.

The results of this off-line processing are depicted in figure 11.10.

S-wave detection and estimation For the S-wave, the problem statement is the same as for the P-wave, except the polarization is different. Processing of the S-wave can be achieved as we explained

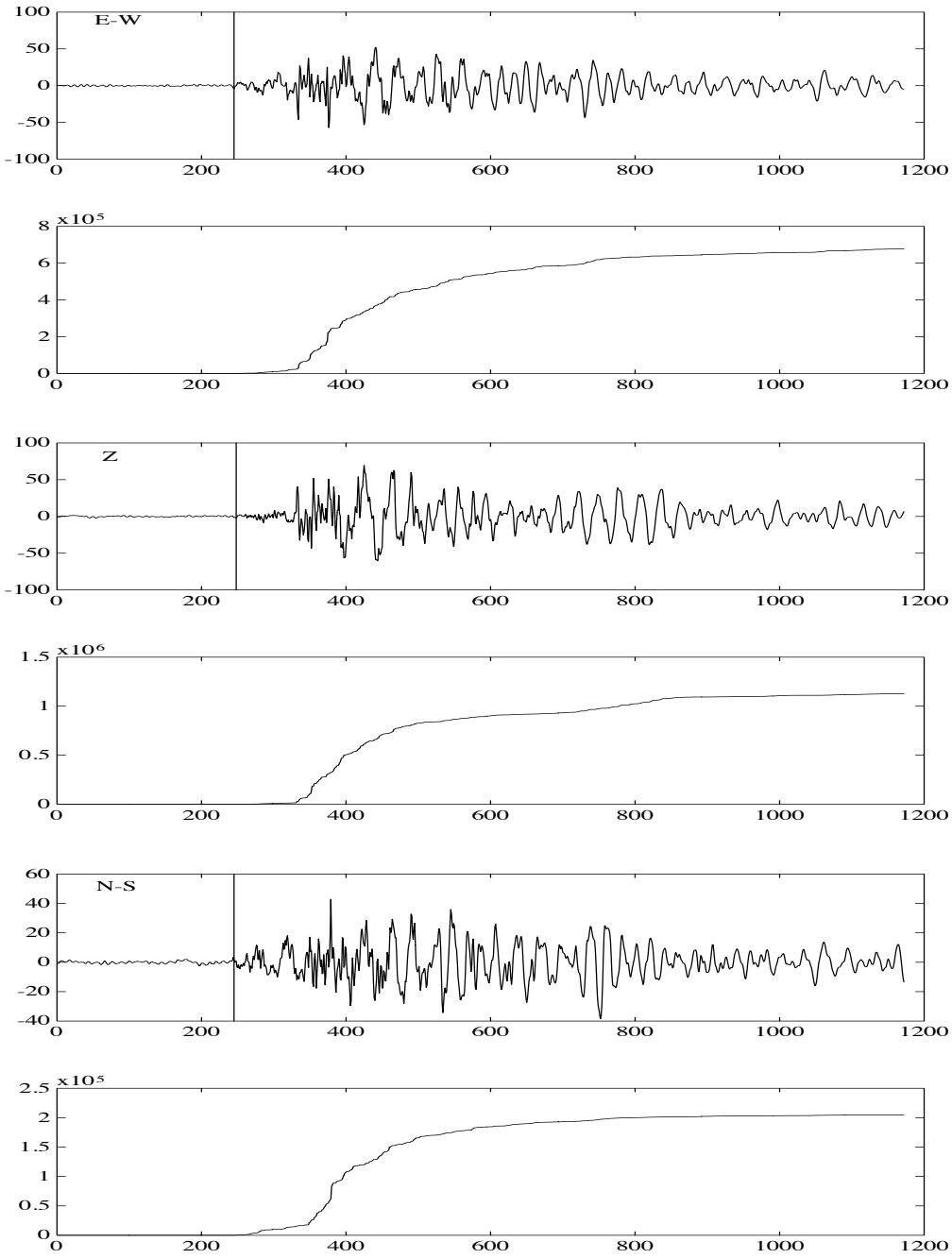

Figure 11.8 On-line P-wave detection : the three components of a seismogram and the corresponding decision functions. The detection times are indicated by the vertical lines (Courtesy of the Academy of Sciences of USSR, Far Eastern Scientific Center, Institute of Sea Geology and Geophysics).

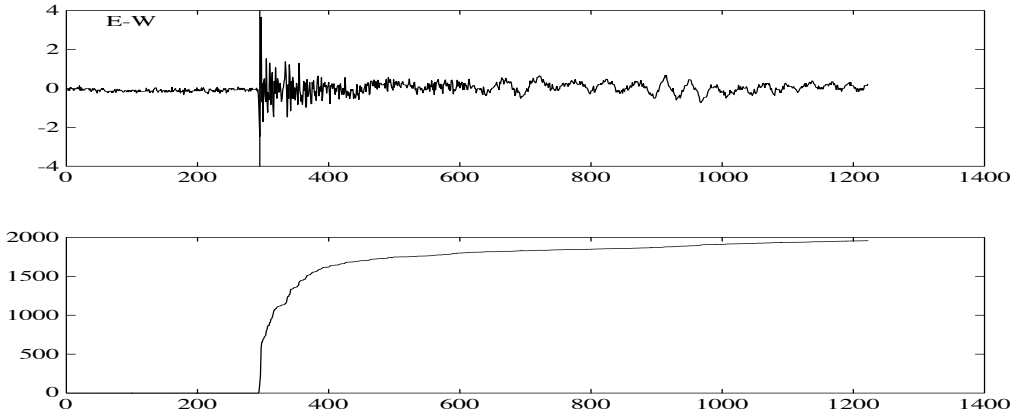

Figure 11.9 On-line P-wave detection for another seismogram (Courtesy of the Academy of Sciences of USSR, Far Eastern Scientific Center, Institute of Sea Geology and Geophysics).

before for the P-wave, but this task is much more difficult for the S-wave. The main reasons for this are the much lower signal-to-noise ratio that is often present in many records, and the high number of intermediate waves that can exist between the P-wave and the S-wave onset times. Many other algorithms, such as different filtering operations and/or additional logic resulting from the underlying physical background, are thus necessary in this case, but are not reported here. The interested reader is referred to [Nikiforov and Tikhonov, 1986, Nikiforov *et al.*, 1989].

11.1.2.3 Results of Real Implementation in Seismic Stations

These algorithms are implemented in the two seismic stations Yujno-Sakhalinsk and Petropavlosk-Khamchatsky in the far eastern part of Russia, where the signals shown in this book have been recorded. One of the tasks of these stations is the detection of underwater earthquakes in the northern part of the Pacific Ocean, in order to predict the occurrence of tsunami. The distance between the stations and the epicenters of these earthquakes typically ranges between 200 and 2000 km. The estimation of the coordinates and the magnitude of the earthquakes should be achieved within the first ten minutes. Two years of real processing in these stations leads to the following conclusions.

On-line Detection First, the above-mentioned on-line P-wave detection algorithm turns out to be robust with respect to the structure and the order of the "exact" model of the seismic noise before the onset time. Actually, it results from several investigations that the seismic noise can be conveniently approximated by an ARMA or an AR model of order 10 to 12. But long real processing shows that, for P-wave detection, the χ^2-CUSUM algorithm corresponding to an AR model of order 2 provides us with convenient results. Moreover, it results from this real processing that this algorithm is also robust with respect to the *a priori* information concerning the parameter vector θ_0 before change : usually records of only 5 min of three-component seismograms are used for tuning the χ^2-CUSUM algorithm, which is run during periods of 8 to 24 h.

Second, the χ^2-CUSUM algorithm proves to be relatively not very reliable with respect to occasional large measurement errors. Such errors can arise from the influence of some nuisance electric field on the measurement system of the seismic station. The detection performances can be improved by using a median-type prefiltering of the seismic signals.

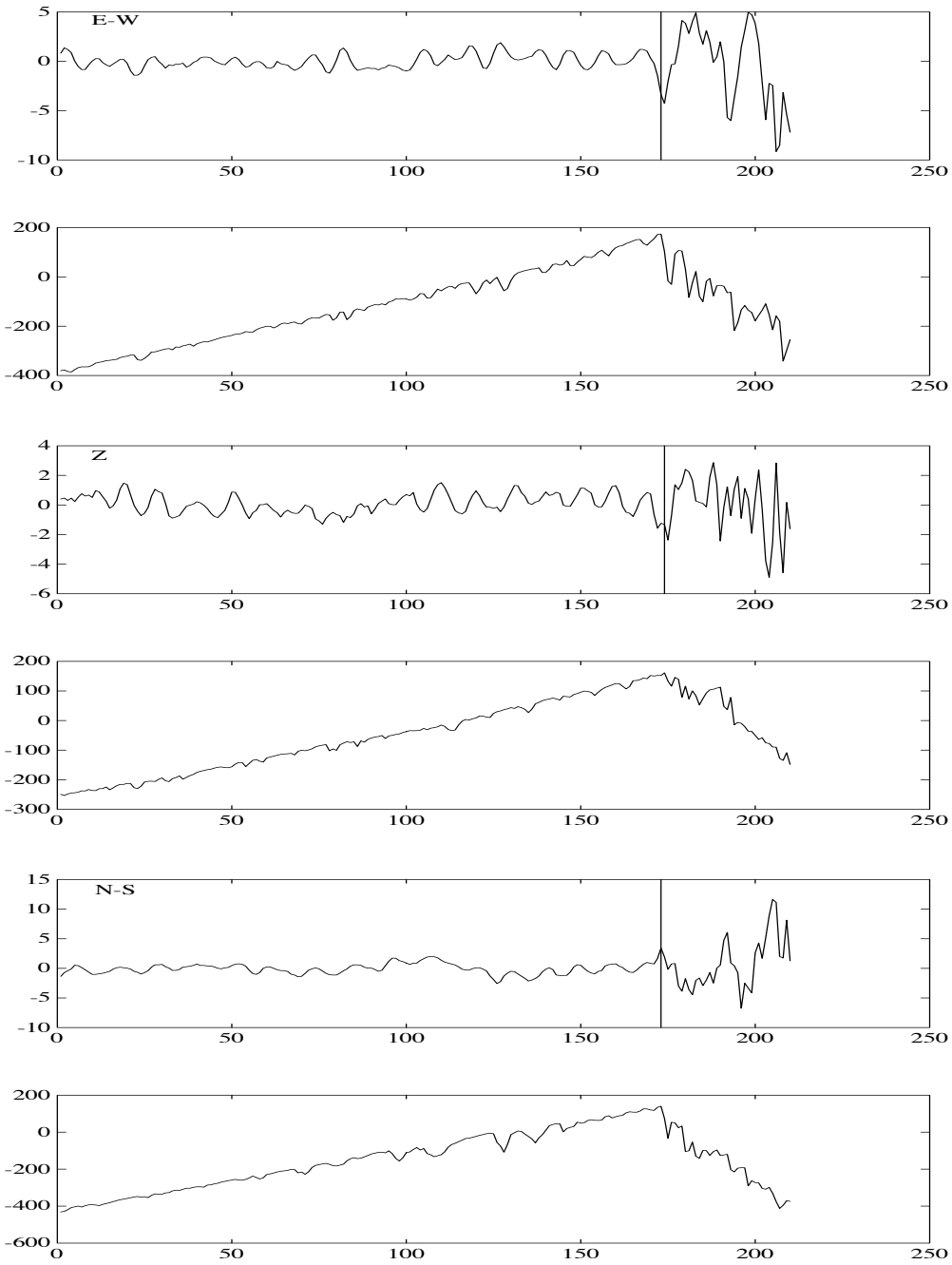

Figure 11.10 Off-line P-wave onset time estimation : data windows from the three components depicted in the figure 11.8, and the corresponding off-line decision functions. The estimated onset time is the abscissa of the maximum of the decision function, and is indicated by the vertical line on the signal (Courtesy of the Academy of Sciences of USSR, Far Eastern Scientific Center, Institute of Sea Geology and Geophysics).

Off-line Estimation The experimental results of the above off-line P- and S-wave onset time estimation algorithms show the same robustness with respect to the structure and the order of the exact models of the seismic noise and signal. But these off-line estimation algorithms are much too sensitive with respect to the size of the data window that is used to estimate the onset time.

Finally, let us add some general comments about real time implementation. First, the above-mentioned algorithms turn out to be reliable and to lead to approximately the same results as the *real time* man-made processing. Second, the additional advantage of this automatic processing is to save time for the operator and to increase the reliability of other operations that can be achieved, especially during the first minutes following an important seismic event.

11.1.3 Segmentation of Speech Signals

Now we describe the main features of the segmentation problem for continuous speech recognition, and the main results that can be obtained when using the divergence algorithm. All the results described in this subsection are due to Régine André-Obrecht. The contribution of Bernard Delyon for helpful discussions and drawing the figures is also gratefully acknowledged. All the signals shown in this subsection belong to two databases designed by the French National Agency for Telecommunications (CNET) for testing and evaluating speech recognition algorithms. The first database is made of ten phonetically balanced French sentences, sampled at 12.8 KHz. The second database - which is also used in the French CNRS National Research Group in Signal and Image Processing - is made of noisy speech signals recorded inside a car with sampling frequency 8 KHz and prefiltered versions of these signals with a high-pass filter with cutting frequency equal to 150 Hz. Both sets of signals are quantized with 16 bits.

11.1.3.1 Problem Statement

It has been recognized for several years that a continuous speech recognition system can usefully contain an analytic acoustic-phonetic processor as its first component - which is not necessarily the case for isolated words recognition systems. This processor takes the continuous speech signal as input and produces a string of phonetic units. When the parametric representation of speech is thus defined, the next step consists of the segmentation of the signal in large units, which are generally phonemic or phonetic units or homogeneous acoustic segments, such as diphones. Finally, the identification of these segments is done. The importance of a correct initial segmentation is great, otherwise the upper recognition level becomes too complex because it works with fuzzy constraints. One widely used approach consists of a recognition-based segmentation, in which the signal is assumed to be described by models or cues (phonetic, articulatory, etc.) which are extracted by FFT or LPC analysis in overlapping windows of constant length. The segmentation is then obtained from a coarse labeling or from a function of the fluctuations of the cues. It has been proven [André-Obrecht, 1988] that an alternative useful first processing step consists of a detection of nonstationarities - namely a segmentation - *without recognition*, which results in a sequence of consecutive segments having lengths adapted to the local properties of the speech signal, and not of fixed-length overlapping segments as before. The main desired properties of such a segmentation algorithm are few missed detections and false alarms, and also low detection delay, although the actual change times and thus the delays cannot be easily stated. One of the features of interest in this approach with respect to the more classical fixed size overlapping moving window approach is that at the further step of recognition, the number of states of the underlying hidden Markov model that is used is approximately three times lower.

The relevant segmentation problem for speech processing is the spectral segmentation problem, as opposed to the segmentation in terms of changes in the mean level. The *spectral segmentation* problem can be approached in the following manner. Taking an on-line point of view, we assume that only one change

has to be detected at a time, and, exactly as in the seismic onset detection problem, we consider a scalar zero-mean signal $(y_k)_k$ described by the AR model (11.1.15). The AR coefficients and the variance of the excitation may change at an unknown time instant, and the problem of interest is the on-line detection of such a change, and the estimation of the change time t_0, and possibly of the AR parameters before and after change. Detection tools for solving this problem were presented in section 8.3. We now demonstrate the relevance of the divergence algorithm, and compare it to a particular approximate implementation of the GLR algorithm.

11.1.3.2 The Usefulness of the Divergence Algorithm

The divergence algorithm has been recognized to be helpful in continuous speech processing for recognition purposes [André-Obrecht, 1988, André-Obrecht and Su, 1988, André-Obrecht, 1990] and for coding [Di Francesco, 1990]. Let us explain the main features of the behavior of this algorithm when applied to continuous speech signals, and the main experimental properties of this algorithm as they result from the processing of large data bases.

Typical behavior of the divergence decision function We discussed the implementation issues related to the divergence algorithm in section 8.6. As shown in figure 8.6, the implementation of this algorithm for processing real data requires the use of two identification methods inside two different data windows. The following choice has proven satisfactory. Inside the growing window, we use the approximated least-squares Burg identification algorithm in lattice form [Basseville, 1986] for estimating the model M_0. Inside the fixed-size sliding window, we use the so-called autocorrelation identification method [Markel and Gray, 1976] for the model M_1. Recall that such an implementation requires that the divergence algorithm is inhibited during a period of time at least equal to the length of the sliding window; this time interval is referred to as a dead zone.

Because the goal of this recognition-oriented segmentation is to obtain segments with length less than the average duration of a phoneme, the size of the sliding window is chosen to be equal to 20 ms. Typically, this gives 256 sample points for the signals sampled at the sampling frequency 12.8 KHz and 160 for the above mentioned noisy speech signals. The AR order selected here is equal to 16. We comment further on this choice next when discussing robustness issues.

Let us now discuss the choice of the two remaining tuning parameters, namely the minimum magnitude of spectral change ν in (8.3.13) and the threshold h in the corresponding cumulative sum decision function. The choice of ν has been dictated by the experimental result shown in figure 11.11. In this figure, we show that different spectral changes in the speech signals are reflected in different ways on the divergence cumulative sums : the slope of the cumulative sum is not the same inside the voiced and unvoiced segments, respectively, and the magnitude of the changes, as reflected in the deviation of the divergence cumulative sum with respect to its maximum, is not constant. What is remarkable, however, is that it turns out that only *two* different choices of the pair (ν, h) are necessary for processing continuous speech signals :

$$\begin{array}{rcll} (\nu, h) & = & (0.2, 40) & \text{in voiced zones} \\ (\nu, h) & = & (0.8, 80) & \text{in unvoiced zones} \end{array} \tag{11.1.16}$$

A very rough detector for choosing between the two pairs is simply activated during the dead zone of the divergence algorithm. When the signal is highly corrupted by noise (see the examples that follow), only the first pair of tuning parameters is used.

An example of segmentation obtained by this tuning of the parameters is shown in figure 11.12. In this figure we show that a phonetic event is not detected, and this fact is true whatever the choice of ν and h is. It turns out that, as we explained in chapter 8, some of these events that are not detected when the signal is

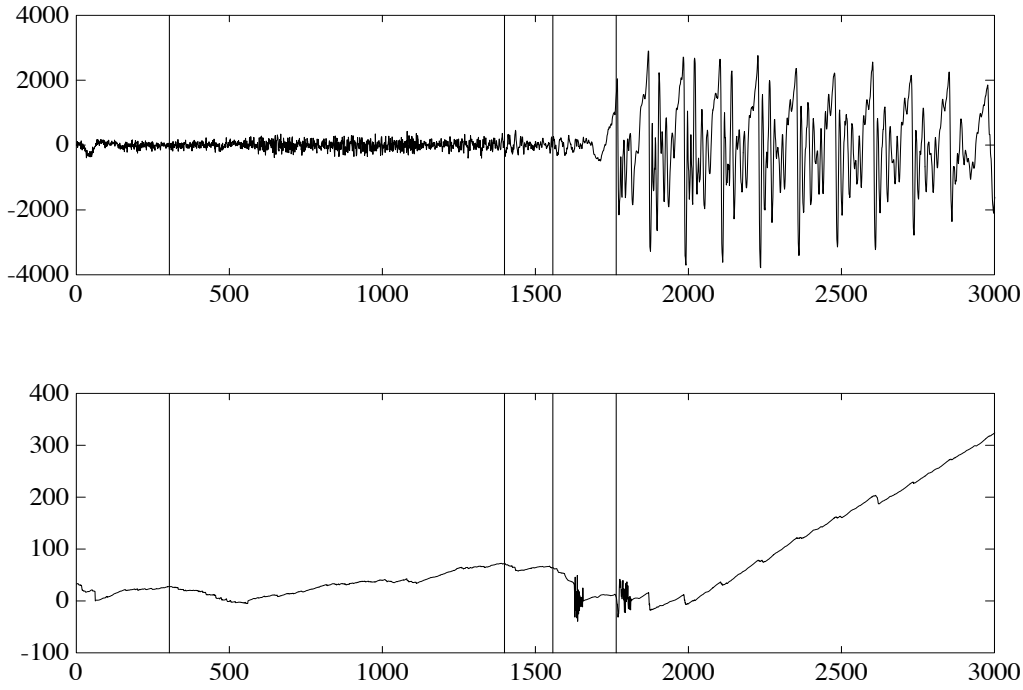

Figure 11.11 Behavior of the divergence cumulative sum on a speech signal with a constant value of ν : the slope of the cumulative sum is greater in the voiced segments than in the unvoiced segments.

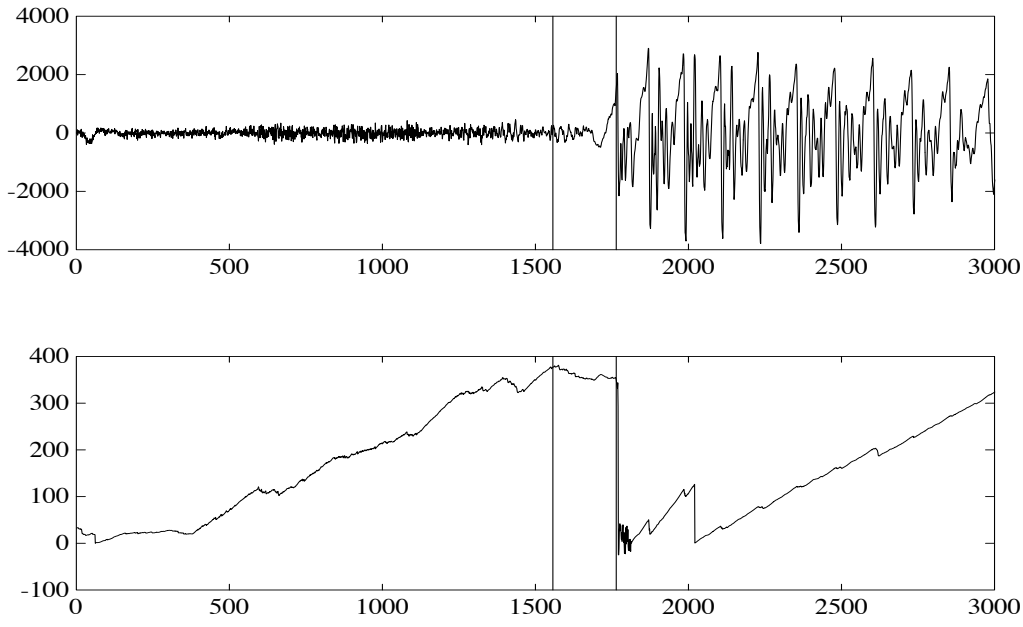

Figure 11.12 The divergence algorithm on the same speech signal with the pairs ν, h.

processed forward are detected when the signal is processed backward. This is the lack of symmetry of the detection algorithm. For this reason, and when the length of a segment is greater than a prespecified value - related to mean length of a phoneme - a backward processing is activated during a time interval with a length less than the length of this long segment. From now on, all the results of segmentation of continuous speech signals that we show are obtained with the aid of this forward-backward divergence algorithm. The interested reader is referred to [André-Obrecht, 1988] for further details.

Experimental properties of the divergence algorithm The following properties of the divergence algorithm when applied to continuous speech signals can be deduced from processing a large number of sentences pronounced by different speakers and under various noise levels :

- the tuning values (11.1.16) for ν and h do not depend upon the *speaker* (male or female);
- the tuning values for ν and h do not depend upon the *noise level*;
- the tuning values for ν and h do not depend upon the *sampling frequency*;
- the tuning values for ν do not depend upon the *quantization rate* of the signal; only the threshold h has to be decreased when this rate is, for example, equal to 8 bits instead of 16 as before;
- the tuning values for ν and h do not depend upon the *AR order* used in the algorithm; this is discussed with the robustness issues next.

An example of processing of a noisy speech signal is shown in figures 11.13 and 11.14. The upper row of these figures depicts a speech signal recorded inside a car and the result of the segmentation algorithm with the values of ν, h taken as in (11.1.16). The lower row of these figures depicts the same sentence, which has been low-pass filtered to remove the noise, and the result of the segmentation of this filtered signal with *again* the choice (11.1.16). It results from these two figures that, as far as the segmentation itself is concerned, the pre-filtering of the noisy speech signal is not necessary : the segmentation results that are obtained with or without this prefiltering operation - and with the aid of the *same* tuning values - are quite similar.

11.1.3.3 Comparison Between the Divergence and GLR Algorithms

We mentioned in chapter 8 that an approximate implementation of the GLR algorithm was proposed in [Appel and von Brandt, 1983, von Brandt, 1983], with the advantage of being far less time-consuming. The key idea is to decouple the detection and the estimation of the change time, which is done at a second step. We refer the reader to [Appel and von Brandt, 1983, von Brandt, 1983, André-Obrecht, 1988] for further details. Both the divergence algorithm and this implementation of the GLR algorithm have been run on the same speech sentences [André-Obrecht, 1988]. An example of such processing of a given sentence is shown in figures 11.15 and 11.16. In these figures, it is obvious that the behavior of the GLR decision function is far less smooth than the behavior of the divergence decision function, which leads us to suspect difficulties in tuning the value of the threshold h in the approximate GLR algorithm. Actually, this problem arises in these comparative experiments, which leads us to prefer the divergence algorithm for speech processing.

11.1.3.4 Discussion: Modeling Robustness Issues

Even though the usefulness of AR and ARMA models for spectral analysis has been demonstrated for many types of signals, it has to be kept in mind that in the present framework of recognition-oriented change detection, the AR or ARMA models to be used are nothing but a tool for the detection of such changes, and have not necessarily the same orders as the models used for characterizing the various segments. For

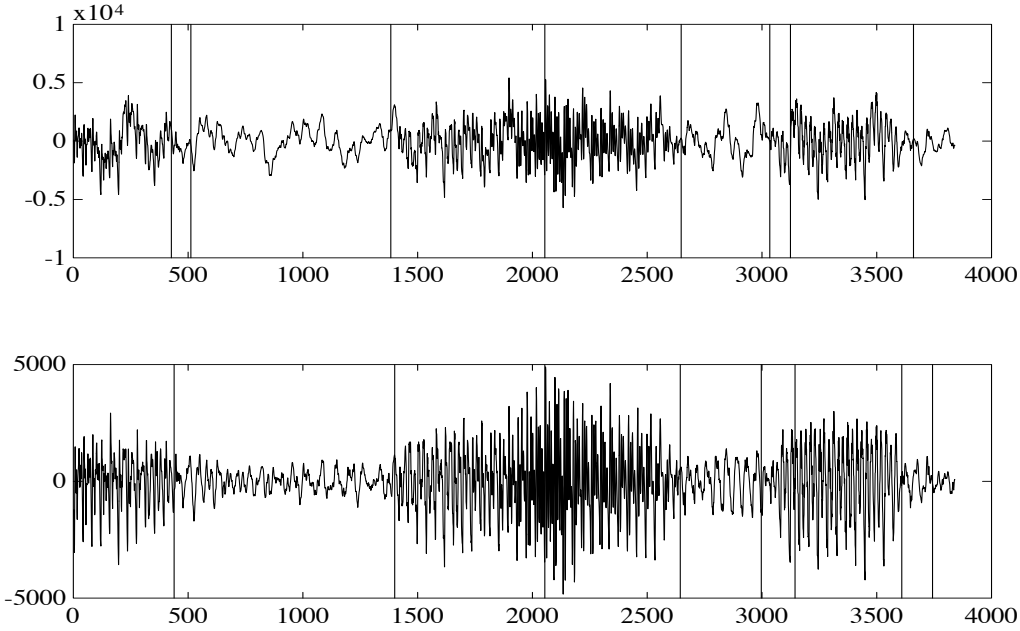

Figure 11.13 Segmentation of a noisy speech signal, without (upper row) or with (lower row) prefiltering. The vertical lines indicate the estimated change times.

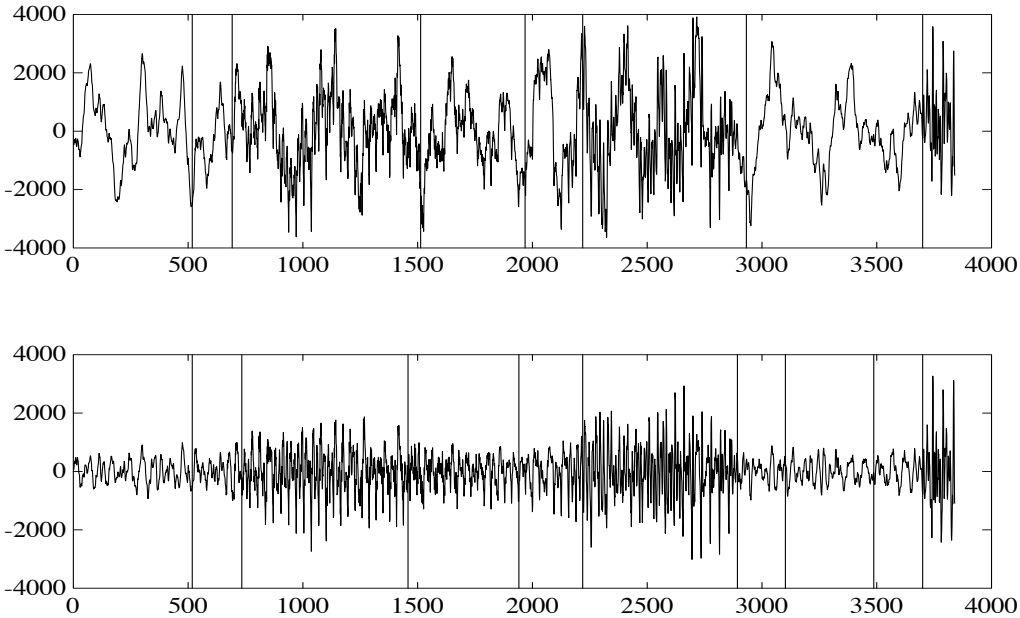

Figure 11.14 Segmentation of a noisy speech signal (contd.). The vertical lines indicate the estimated change times.

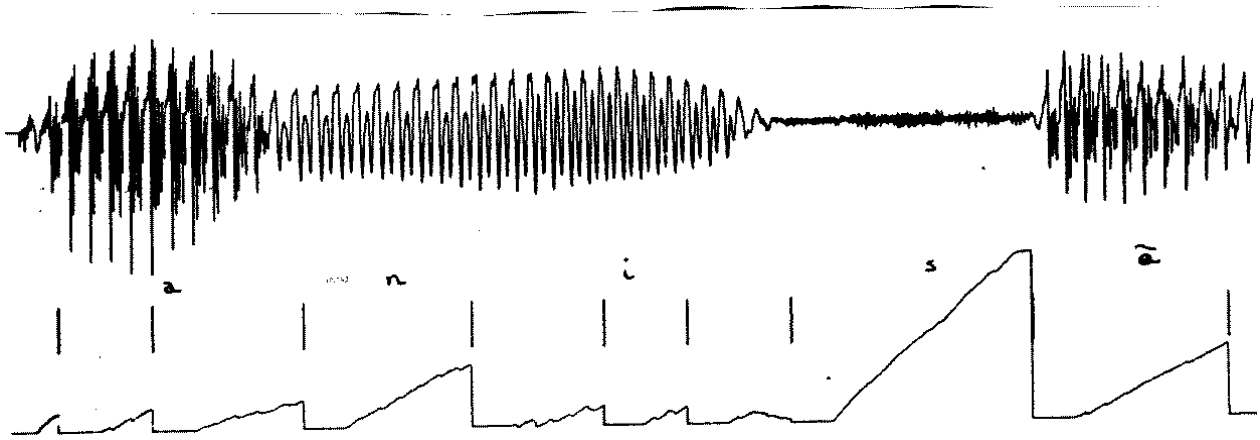

Figure 11.15 Divergence algorithm.

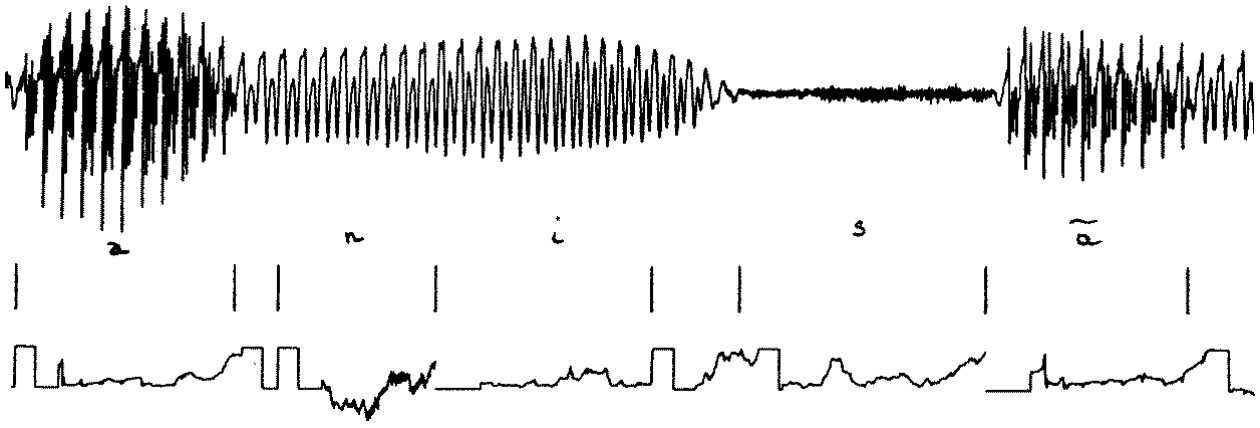

Figure 11.16 Approximate GLR algorithm.

example, robust segmentation results can be obtained with the aid of AR models of order 2 on continuous speech signals that are usually analyzed by linear filters of order 12 to 16 [André-Obrecht, 1988]. This can be seen in figures 11.17 and 11.18 where we compare, for the *filtered* version of the above-mentioned noisy speech signal, the segmentation obtained with an AR order equal to 16 to the segmentation obtained with the aid of the same tuning values but with an AR order equal to 2. Very few differences exist between the two segmentation results. The main comment that can be made is that, apart from alarms that are set when using the order 16 and not when using the order 2, the alarms are set at extremely close time instants with both orders. The most surprising fact is that this result equally holds when considering the *noisy* signal itself, as is obvious in figures 11.19 and 11.20.

Furthermore, such a segmentation algorithm allows us to detect several types of events [André-Obrecht, 1988], and not only abrupt changes between two stationary segments. Typically, it allows us to detect onsets of gradual spectral changes, such as a fluctuation of energy inside a frequency band, or a drift fluctuation of a formant, or a loss of formantic structure.

Finally, let us comment upon the *estimation* issue concerning the models before and after change, which are to be used for recognition purposes. In many real cases, when the characteristics of the signal before and after change are of interest - for example, for classification and recognition purposes - it is necessary to reidentify these characteristics inside the detected segments, and not simply to use the output of the filter(s) involved in the detection algorithm. The key reason for this is that, in practice, the ideal model (11.1.15) used in the algorithm is only an approximation, and, because a real signal can be seen as a sequence of slowly time-varying segments, a global reidentification of each entire segment is necessary.

11.1.4 Vibration Monitoring of Mechanical Systems

As we explained in chapter 1, the problem of vibration monitoring of mechanical structures and rotating machines under usual operating conditions is of key practical importance. The use of artificial excitations or stopping the machine, which is required by many monitoring and maintenance procedures, is often prohibitive in terms of costs and feasibility. The interest of a sensor-based monitoring procedure lies in its ability to extract detection and diagnosis information from the measurements that are taken under the usual operating conditions, namely without changing the rotation speed of the machine, and under usual surrounding excitations.

In this subsection, we first show that the vibration monitoring problem is nothing but the problem of detecting and diagnosing changes in the eigenstructure of a nonstationary multivariable system in state-space form, or equivalently in the AR part of a multivariable ARMA model with nonstationary MA part. Then we report numerical results obtained for a simulated system.

11.1.4.1 Vibration Monitoring and Changes in the Eigenstructure

We assume that a vibrating structure may be decomposed into finite elements and has a linear behavior. Under this assumption, the continuous time model of such a system is as follows :

$$\left\{ \begin{array}{rcl} M\ddot{Z}_t + C\dot{Z}_t + KZ_t & = & E_t \\ Y_t & = & LZ_t \end{array} \right. \tag{11.1.17}$$

where

the first equation is the well-known *forces = mass* $\times$ *acceleration* relation;

Z_t is the vector of the positions of the m discretized elements of the structure;

$-C\dot{Z}_t$ is the friction force and C the damping matrix;

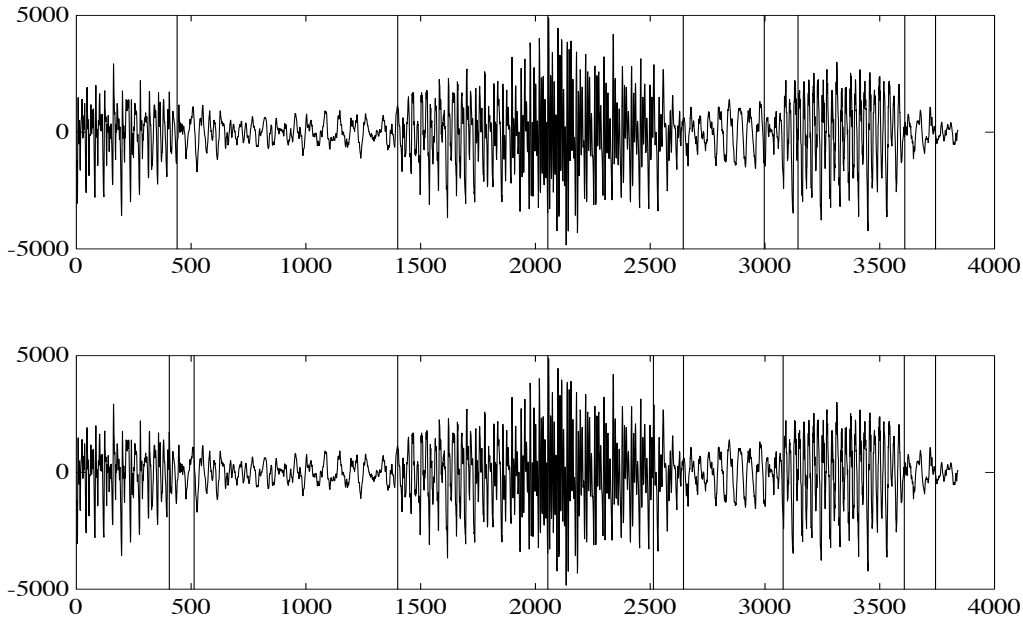

Figure 11.17 Segmentation of the filtered speech signal corresponding to the noisy signal of figure 11.19, with AR order 16 (upper row) and with AR order 2 (lower row). The vertical lines indicate the estimated change times.

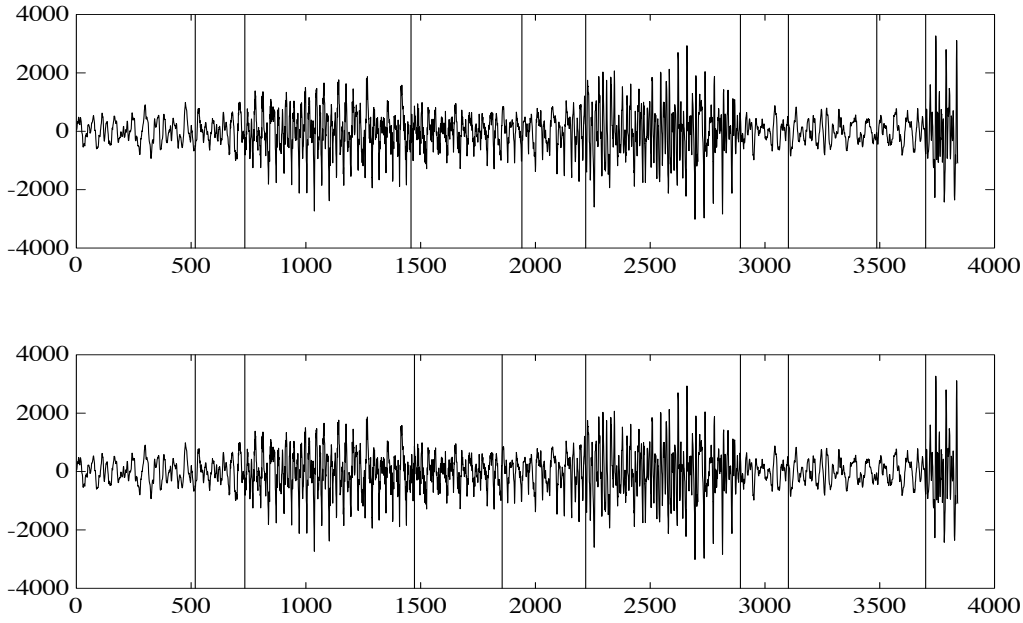

Figure 11.18 Segmentation with different orders (contd.) : the filtered speech signal corresponding to the noisy signal of figure 11.20 .

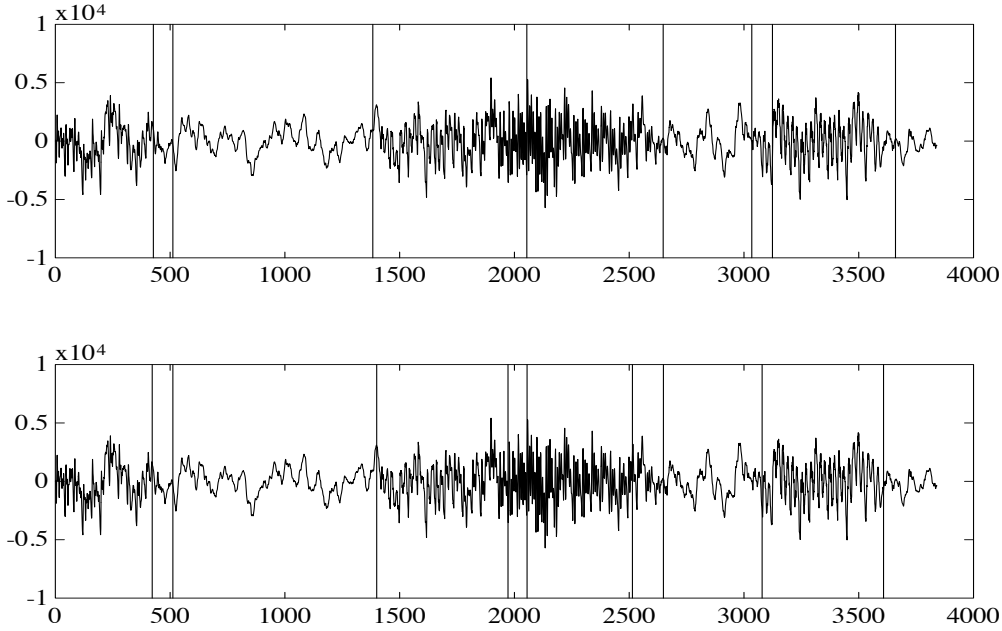

Figure 11.19 Segmentation of the noisy speech signal, with AR order 16 (upper row) and with AR order 2 (lower row). The vertical lines indicate the estimated change times.

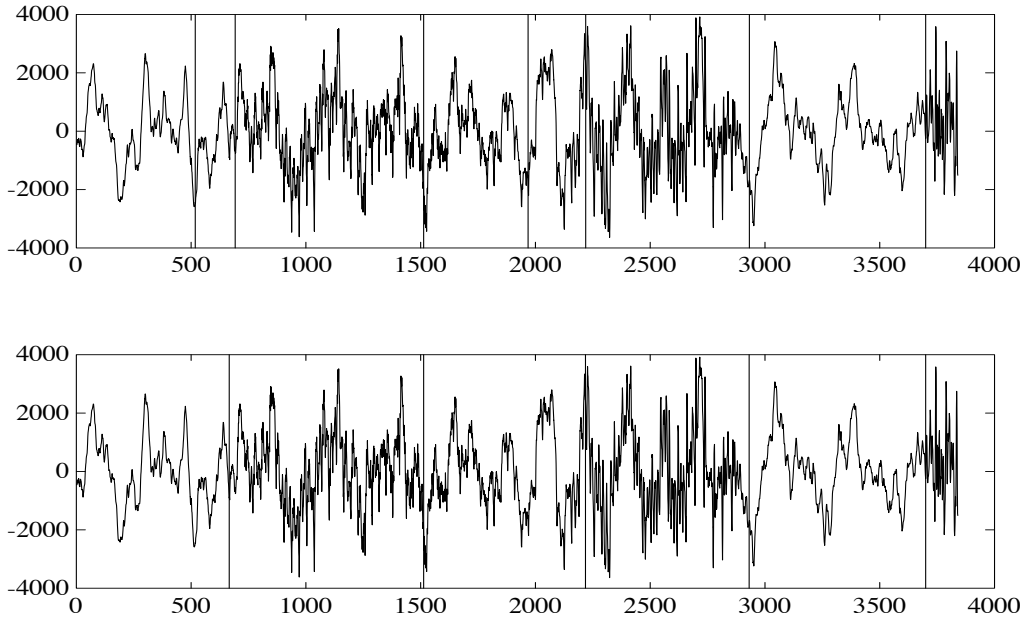

Figure 11.20 Segmentation with different orders (contd.) : the noisy speech signal .

$-KZ_t$ is the stiffness force and K the stiffness matrix;

E_t is the external (nonmeasured) force, which is simulated by a *nonstationary white noise,* with covariance matrix Σ_t;

Y_t is the vector of the measurements, of dimension smaller than the dimension of Z_t, and the matrix L specifies which node displacements of the structure are measured, namely where the sensors are.

In the appendix, numerical values for the matrices M, K, and C are given for a particular simulated system. After sampling with period δ and transformation of the second-order differential equation (11.1.17) into a first-order system, we obtain the following discrete time state-space model :

$$\left\{ \begin{array}{rcll} X_{k+1} & = & FX_k + W_k, & \dim X_k = n = 2m, \ \ \mathrm{cov}(W_k) = Q_k \\ Y_k & = & HX_k, & \dim Y_k = r \end{array} \right. \tag{11.1.18}$$

where

$$\begin{aligned} X &= \begin{pmatrix} Z \\ \dot{Z} \end{pmatrix} \\ F &= e^{\tilde{F}\delta} \\ \tilde{F} &= \begin{pmatrix} 0 & I \\ -M^{-1}K & -M^{-1}C \end{pmatrix} \\ H &= \begin{pmatrix} L & 0 \end{pmatrix} \\ W_k &= \int_k^{k+\delta} e^{\tilde{F}(k+\delta-t)} \begin{pmatrix} 0 \\ M^{-1}E_t \end{pmatrix} dt \\ Q_k &= \int_k^{k+\delta} e^{\tilde{F}t} \tilde{\Sigma}_t e^{\tilde{F}^T t}\, dt \\ \tilde{\Sigma}_t &= \begin{pmatrix} 0 & 0 \\ 0 & M^{-1}\Sigma_t M^{-1} \end{pmatrix} \end{aligned} \tag{11.1.19}$$

Let us emphasize that the generation of the equivalent white noise sequence W_k, given the covariance matrix Σ of the white noise on the 18 masses system, has to be done carefully; in other words, the integration in the formula giving Q_k as a function of Σ_k has to be done in a fine way.

By definition, the *vibrating characteristics* (μ, ψ_μ) of a vibrating system are given by

$$\begin{array}{rcl} \det(M\mu^2 + C\mu + K) & = & 0 \\ (M\mu^2 + C\mu + K)\psi_\mu & = & 0 \end{array} \tag{11.1.20}$$

The eigenvalues λ and eigenvectors ϕ_λ of the state transition matrix F are related to the vibrating characteristics (11.1.20) through

$$\begin{array}{rcl} \lambda & = & e^{\delta\mu} \\ H\phi_\lambda & = & L\psi_\mu \end{array} \tag{11.1.21}$$

In other words, the eigenstructure of the state transition matrix F contains all the vibrating characteristics of the mechanical structure. The state transition matrix F can be factorized as

$$F = \Phi e^D \Phi^{-1} \tag{11.1.22}$$

where

$$D = \begin{pmatrix} \Delta & 0 \\ 0 & \bar{\Delta} \end{pmatrix} \tag{11.1.23}$$

and

$$\Phi = \begin{pmatrix} \Psi & \bar{\Psi} \\ \Psi\Delta & \bar{\Psi}\bar{\Delta} \end{pmatrix} \tag{11.1.24}$$

and where $\Delta = \mathrm{diag}(\mu)$ and Ψ contains the ψ in its columns.

Now, as we explained in subsection 3.2.4, the state-space model (11.1.18) is equivalent to a multivariable ARMA model :

$$Y_k = \sum_{i=1}^{p} A_i Y_{k-i} + \sum_{j=0}^{q} B_j(k) V_{k-j} \tag{11.1.25}$$

where the AR part can be obtained by solving the linear system of equations :

$$HF^p = \sum_{i=1}^{p} A_i H F^{p-i} \tag{11.1.26}$$

and where the nonstationary state noise W_k in (11.1.18) is reflected only in the MA matrix coefficients $B_j(k)$ for which we note the time dependence explicitly. Therefore, the eigenvalues λ and eigenvectors ϕ_λ of the state transition matrix F are also solutions of

$$\left(\lambda^p I_r - \sum_{i=1}^{p} \lambda^{p-i} A_i \right) H\phi_\lambda = 0 \tag{11.1.27}$$

These pairs $(\lambda, H\phi_\lambda)$ are called *modal signature*.

Consequently, identifying and monitoring the set of pairs $(\lambda, H\phi_\lambda)$ given by (11.1.27) is equivalent to the same tasks for the set of $(\mu, L\psi_\mu)$, which are the *observed* part of the vibrating characteristics given in (11.1.20). We solved this problem in chapter 9.

11.1.4.2 Experimental Results

We now show what kind of numerical results can be obtained with the aid of the instrumental statistic when applied to the simulated system made of 18 masses described in the appendix, and depicted in figure 11.21. The contribution of Marc Prevosto from the French Research Institute for the Sea (IFREMER), who provided us with the corresponding simulated data, is gratefully acknowledged.

Detection Recall that the instrumental test consists of computing the following statistic :

$$\bar{\mathbf{Y}}_{\mathbf{N}} = \sqrt{N} \left(\mathcal{H}^T_{p+1,p} \otimes I_r \right) \begin{pmatrix} -\check{\underline{\theta}} \\ I_r \end{pmatrix} \tag{11.1.28}$$

$$\check{\underline{\theta}}^T = \begin{pmatrix} A_p & \dots & A_1 \end{pmatrix} \tag{11.1.29}$$

The resulting χ^2 test

$$(\chi^k_j)^2 = \bar{\mathbf{Y}}^T_N \, \hat{\Sigma}^{-1}_N \, (\mathcal{H}^T_{p,p} \otimes I_r) \, \left[(\mathcal{H}^T_{p,p} \otimes I_r)^T \hat{\Sigma}^{-1}_N (\mathcal{H}^T_{p,p} \otimes I_r) \right]^{-1} \, (\mathcal{H}^T_{p,p} \otimes I_r)^T \, \hat{\Sigma}^{-1}_N \, \bar{\mathbf{Y}}_N \tag{11.1.30}$$

where

$$\hat{\Sigma}_N(\underline{\theta}^*) = \frac{1}{N} \sum_{k=1}^{N} \sum_{i=-p+1}^{p-1} \check{\mathcal{Y}}^{k-p}_{k-2p+1} (\check{\mathcal{Y}}^{k-i-p}_{k-i-2p+1})^T \otimes e_k e^T_{k-i} \tag{11.1.31}$$

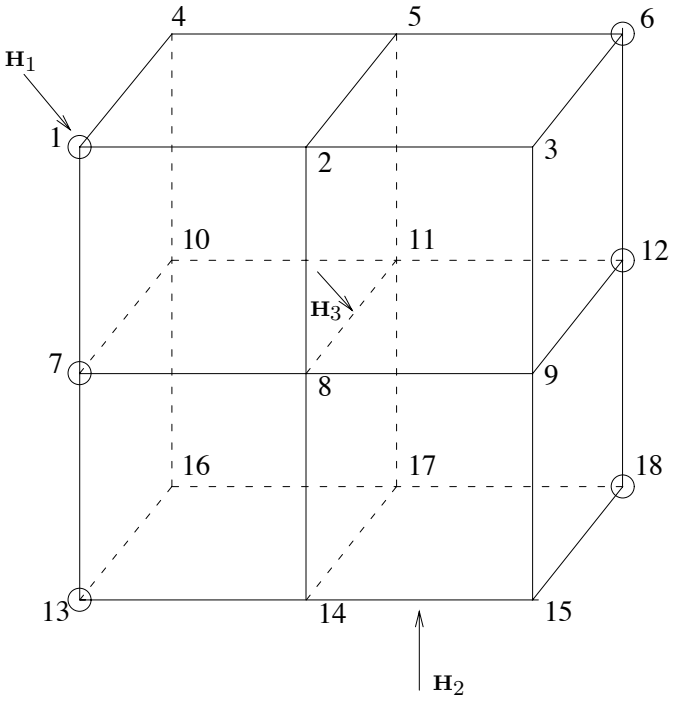

Figure 11.21 The 18 mass and spring system. Fault $\mathbf{H}_1$: change in the mass 1; Fault $\mathbf{H}_2$: change in the stiffness of the connection to the ground; Fault $\mathbf{H}_3$: cutoff in the connection between the masses 8 and 11.

has $2mr = 36r$ degrees of freedom, where r is the number of sensors, and the following noncentrality parameter :

$$\lambda = \Upsilon^T \left(\mathcal{H}_{p,p}^T \otimes I_r\right)^T \hat{\Sigma}_N^{-1} \left(\mathcal{H}_{p,p}^T \otimes I_r\right) \Upsilon \tag{11.1.32}$$

This test has been computed for 43 possible sensor locations, representing all the subsets of sensors corresponding to $r = 2, 3, 4$, and 6 and to at least one sensor on each of the two opposite "legs" 1-7-13 and 6-12-18 of the structure. These sensor locations are given in table 11.4.

The test (11.1.30), computed under the four hypotheses, namely the no-change situation and each of the three faulty situations given in the appendix, is depicted in figure 11.22. Recall that the mean value of the instrumental test under the no-change situation is the number of degrees of freedom in (11.1.30), namely $36r$, where r is the number of sensors (ranging from 2 to 6).

These results show that some sensor locations lead to very poor detection, in the sense that no detection is possible (nearly the same value under the no-change and under the faulty situations). Since the poor locations depend upon the considered fault, it can be of practical interest to use moving sensors, along the leg of an offshore platform, for example. These figures also show that the third faulty situation, which was *a priori* thought to be nondetectable, is detected by the instrumental test in a small number of sensor locations, and with a lower threshold (lower ratio between nonfaulty and faulty situations).

Frequency diagnosis The sensitivity technique described in section 9.3 for the diagnosis of changes in terms of the eigenfrequencies and eigenvectors, was first tested in [Basseville *et al*., 1986] in the case of *scalar* signals, using the AR(4) and ARMA(4, 3) models, the latter one with a nonstationary MA part. The poles are pairwise conjugate, and changes in only one of them are simulated. The result is that the detection and diagnosis of *small* changes - of order of magnitude of 1% - in the eigenfrequencies are possible, provided

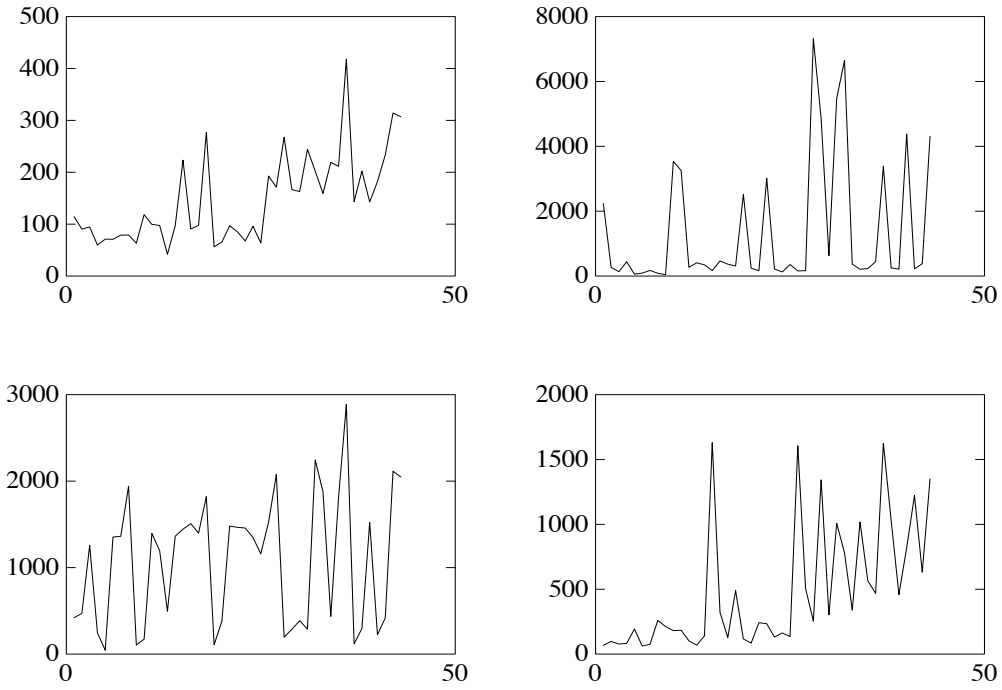

Figure 11.22 The instrumental test for the 18 mass system. Upper left : no change; upper right : $\mathbf{H}_1$, change in mass 1; lower left : $\mathbf{H}_2$, change in the stiffness of the connection to the ground; lower right : $\mathbf{H}_3$, change in the stiffness of the connection 8-11.

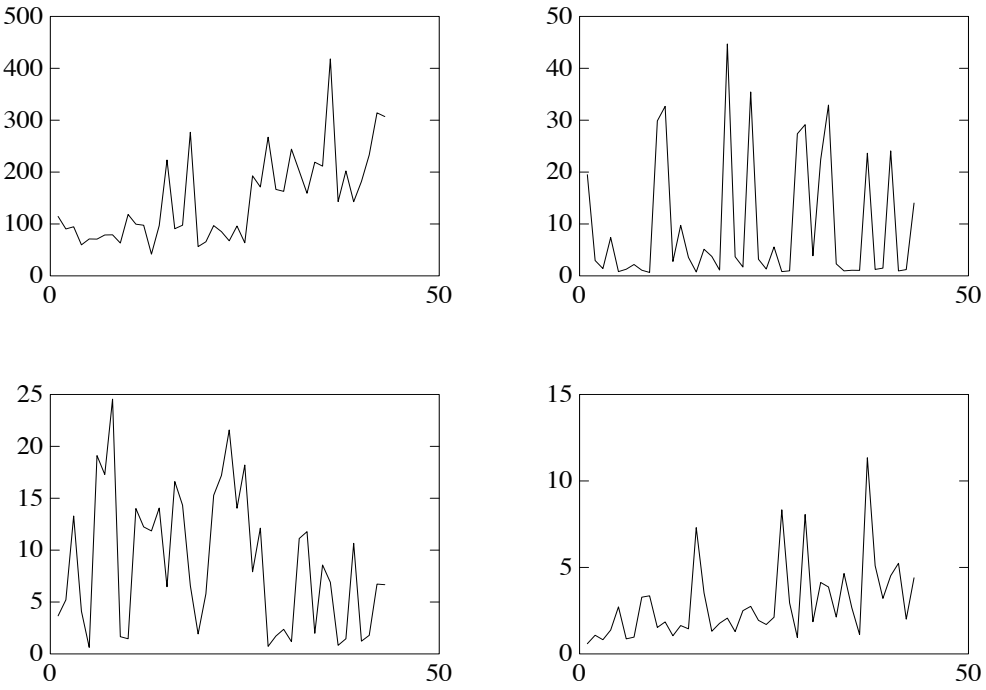

Figure 11.23 Values of the test under the no-change situation, and ratios between the change and no-change situations.

that these poles are slightly damped. This may be explained by the property of the Fisher information matrix in this case, which goes to a diagonal matrix when the poles all go toward the unit circle. It also appears that a change in a damped frequency can be masked, and thus not even detected, by the presence of a slightly damped frequency. We discussed this detectability issue in section 8.5 with the aid of the behavior of the Kullback information when the poles go toward or move off the unit circle again.

In the case of *multidimensional* signals, simulations have shown that as soon as there exist less sensors than degrees of freedom in the masses and springs system - which is most of the time in practice - there exists a coupling between the various eigen components of the system. The interest of a diagnosis directly in terms of the mechanical characteristics and using an aggregation for clustering changes that are not distinguishable with the aid of the instrumental test is thus obvious.

Mechanical diagnosis As explained in section 9.3, the basic idea of the diagnosis in terms of the mechanical parameters M, K consists of using a sensitivity technique coupled to the instrumental test again, but first also using a clustering procedure, based upon the *same* metric as that of the instrumental χ^2-test (11.1.30). A given resulting cluster contains elementary changes in M, K that cannot be discriminated by the sensitivity test. Recall that we do not diagnose changes in terms of the damping coefficients, mainly because these coefficients are less well identified than the others. But our technique could be used for these damping coefficients as well.

For the above-mentioned 18 mass system, the results obtained for the mechanical diagnosis are the following. Let us consider a sensor location under which both the last two faulty situations - change in the stiffness of the connection to the ground and cutoff of the connection 8-11 - are detected, selection $7, 13, 12$, for example. The clustering process results in a set of 14 classes containing the 92 possible elementary

Table 11.3 The global and sensitivity tests for vibration monitoring.

change type	$\mathbf{H}_2$	$\mathbf{H}_3$	
global test	1514.53	1623.37	
sensitiv.1	3.79	157.34	
sensitiv.2	193.42	18.35	
sensitiv.3	254.66	3.10	
sensitiv.4	58.13	195.52	
sensitiv.5	2.44	2.16	
sensitiv.6	0.14	1122.13	
sensitiv.7	0.11	1510.76	$\leftarrow \mathbf{H}_3$
sensitiv.8	410.60	0.06	
sensitiv.9	104.69	232.43	
sensitiv.10	0.01	120.69	
sensitiv.11	1414.67	2.28	$\leftarrow \mathbf{H}_2$
sensitiv.12	1.87	1362.60	
sensitiv.13	0.86	738.66	
sensitiv.14	1.01	748.42	

changes in M and K. The numerical values of the global instrumental test and of the sensitivity tests under both hypotheses are given in table 11.3. The decision strategy consists of selecting the sensitivity test with maximal value, and diagnosing the change in terms of the mechanical elements that form the class underlying the corresponding sensitivity test. These classes are depicted with the aid of arrows for stiffness coefficients and bullets for masses in figures 11.24 and 11.25 for hypotheses $\mathbf{H}_2$ and $\mathbf{H}_3$, respectively. These figures show a quite satisfactory physical coherence, leading to a correct diagnosis. Note that for hypothesis $\mathbf{H}_2$, it is physically difficult to discriminate between masses and stiffness coefficients at the same nodes, and remember that hypothesis $\mathbf{H}_3$ was *a priori* thought to be nondetectable.

In [Devauchelle-Gach, 1991, Basseville *et al.*, 1993] are reported further extensive simulations concerning a vertical steel clamped-free beam which is fixed at the bottom and free at the top end, and excited by a vibrator producing a white noise-like excitation.

11.2 Examples of Potential Areas of Application

In this section, we discuss several potential areas of application of change detection algorithms. In some cases, especially in quality control and biomedical signal processing, change detection algorithms have already been used, but we do not show results of processing real data; instead we refer to the relevant literature.

11.2.1 Statistical Quality Control

First we recall that, from a historical point of view, the first change detection problems arose for on-line statistical quality control. Quality control is concerned with many application areas and plays an important role in modern industries. Let us describe the key issues in this topic [Himmelblau, 1978] :

- Considering the various factors that influence the production in many types of factories, the measured values (concentrations of chemical components, temperatures, pressures, geometrical shapes, etc.)

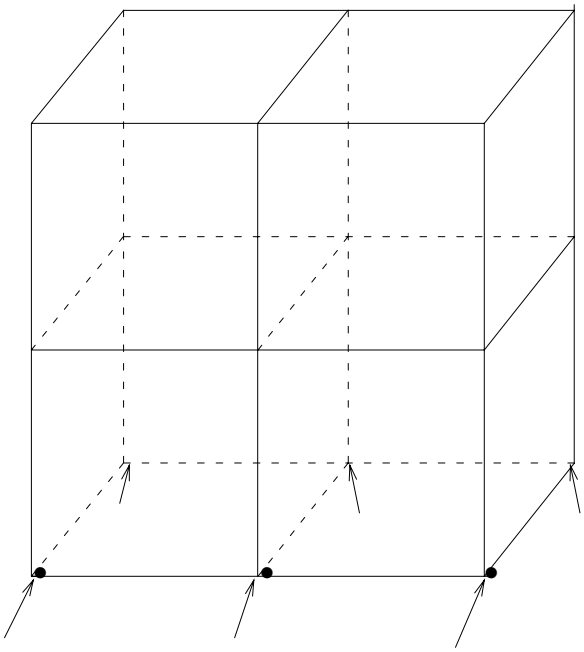

Figure 11.24 The content of the class 11 diagnosing $\mathbf{H}_2$; bullets indicate masses and arrows indicate stiffness coefficients.

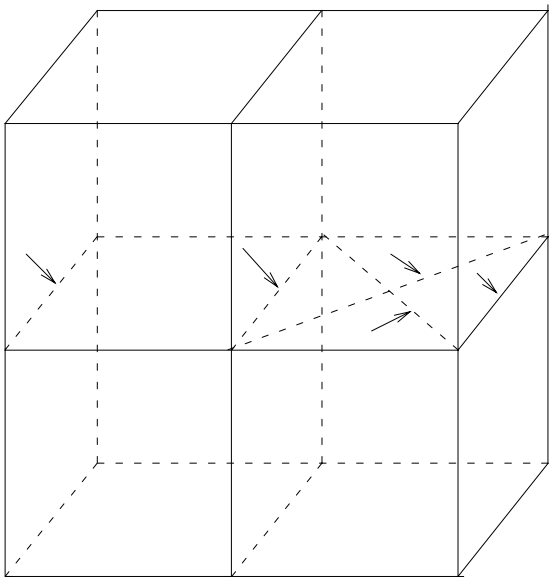

Figure 11.25 The content of the class 7 diagnosing $\mathbf{H}_3$; arrows indicate stiffness coefficients.

should be considered *random*, and thus *statistical* change detection algorithms are relevant for the purpose of quality control.

- Different faults in technological processes lead to nonrandom and deterministic changes in model parameters.
- The probability distributions of the measured time series depend upon the quality of the raw input material and of the production. Therefore, the on-line detection of changes in these distributions can help in the early detection of a decrease in the quality indexes of the output products.

Let us now describe what can be obtained in quality control when using change detection algorithms, and how to tune these algorithms.

11.2.1.1 Change Detection for Quality Control

Change detection algorithms can be used in quality control for several purposes :

- generation of alarms after which the technological process has to be stopped, checked, and repaired if necessary;
- generation of alarms for the attention of the operators;
- classification of the output products according to different quality levels, as results from the use of a two-sided detection algorithm, for example;
- investigation of various types of faults in a technological process, in view of its modernization.

The criteria of such a type of processing are as follows. First, it should be clear that, most of the time a short detection delay is highly desirable to prevent the process from getting into an out-of-control state. Second, the occurrence of false alarms results in additional costs because unnecessary checking and repairing actions are taken. More relevant criteria should include these costs for a joint optimization of the technological process itself and of the monitoring procedure. But, the design of optimum decision rules corresponding to such types of economical criteria is a difficult problem for which closed form solutions seldom exist. For this reason, a possible solution to statistical quality control consists of using statistical change detection algorithms tuned with the criteria of small false alarms rate and delay for detection. Such tuning helps in reducing the overall costs. Moreover, additional costs saving can be achieved by using more sophisticated change detection algorithms than the Shewhart's, GMA, FMA, or CUSUM charts traditionally used in this area. The detection of a change in the mean can be more efficiently achieved if the data correlations are taken into account, as we did, for example, in chapter 7. Moreover, algorithms for detecting changes in spectral properties, such as those discussed in chapters 8 and 9, should be helpful for many measurements.

11.2.1.2 Economic Design of Change Detection Algorithms

We now discuss a particular statistical model and change detection problem for which we describe the main ideas of optimal design with respect to joint costs of the process and the monitoring itself [Taylor, 1968, Chiu, 1974]. We assume that the technological process can be in two states : "in control" or 0 and "out of control" or 1. Moreover, we assume that the state 1 is absorbing, and that there exists a known matrix of transition probabilities between these two states. From the state 0 until state 1 is first reached, the duration t of one life cycle of the joint technological and monitoring process satisfies the following relation :

$$\mathbf{E}(t) = \bar{t}_r + [\mathbf{E}(N_f) + 1]\, \bar{t}_c + \mathbf{E}(N_m)\ \delta + \bar{\tau}^* \delta \tag{11.2.1}$$

where $\bar{t}_r$ and $\bar{t}_c$ are the mean durations of the repairing and checking actions, respectively, N_f is the number of false alarms, N_m is the number of sample points before the change time t_0, and δ is the sampling period. The mean joint cost of such a life cycle is

$$\mathbf{E}(\kappa) = \kappa_r + [\mathbf{E}(N_f) + 1]\,\kappa_c - p\,\mathbf{E}(t_0) + \{[\mathbf{E}(N_m) + \bar{\tau}^*]\,\delta - \mathbf{E}(t_0)\}\,c \tag{11.2.2}$$

where κ_r and κ_c are the costs of the repairing and checking actions, respectively, p is the profit rate of the technological process, and c is the cost rate resulting from the out of control operation. The average cost is estimated as

$$\frac{\mathbf{E}(\kappa)}{\mathbf{E}(t)} \tag{11.2.3}$$

Note here that, as is obvious from the previous equations, this ratio is a function of the mean delay $\bar{\tau}^*$ and of the mean time between false alarms $\bar{T}$ (through N_f). This average cost can be used as an economical criterion for tuning change detection algorithms in order to have a minimal average cost of the joint process and monitoring system.

11.2.2 Biomedical Signal Processing

Like most real signals, biomedical signals exhibit various types of nonstationarities, and the interest in automatic segmentation and detection procedures in this field has been recognized for a long time [Mathieu, 1976, Bodenstein and Praetorius, 1977, Bohlin, 1977, Cohen, 1987]. From a historical point of view, it is well known that several investigations for designing change detection algorithms have been motivated by the automatic processing of biomedical signals, for example, the electroencephalogram (EEG). The shifted log-likelihood decision function (8.2.30)-(8.3.19) was proposed independently in [R.Jones *et al.*, 1970, Borodkin and Mottl', 1976, Segen and Sanderson, 1980] for detecting spikes and segmenting EEG signals. More recently, an approximate implementation of the GLR algorithm was designed in [Appel and von Brandt, 1983, von Brandt, 1983, von Brandt, 1984] for the segmentation of EEG signals again.

The usefulness of change detection algorithms for processing electrocardiograms (ECG) was demonstrated in [Corge and Puech, 1986]. Let us describe here some of the main features of this investigation. The values of the fetal cardiac rhythm signal are the lengths of the time intervals between two consecutive cardiac pulsations. This signal is known to contain the following information :

- the basic cardiac rhythm, which is the low-frequency component of the signal;
- typical peaks exhibiting the accelerations and decelerations in the beating of the heart;
- time-varying spectral characteristics in the high-frequency components.

In [Corge and Puech, 1986], the following automatic processing of this signal is performed :

- The measured signal is first low-pass filtered. The basic rhythm and the peaks are then detected and/or estimated with the aid of a sliding window empirical estimate of the mean value coupled with the CUSUM algorithm for detecting a change in the mean. Some heuristics are used for the validation of the most significant peaks and the estimation of the basic rhythm.
- The high-frequency component of the signal is then extracted by a simple difference between the initial signal and the basic rhythm estimated in the previous step. Changes in the spectral characteristics of the initial signal are detected with the aid of the divergence algorithm applied to this high-pass filtered version.
- The segments obtained in such a way are then characterized by estimated parameter values and classified. The resulting information concerns the alternance between waking and sleeping stages, and is characteristic of the state of the nervous system of the fetus.

11.2.3 Fault Detection in Chemical Processes

It results from the increasing complexity of continuous chemical processes that any break in the normal operation of a process implies high ecological and economical losses. Moreover, several huge catastrophes occurred in such complex processes, which motivated further investigations of new mathematical tools for the early detection of small faults, which can be sources of subsequent catastrophic situations [Himmelblau, 1970, Himmelblau, 1978, Patton *et al*., 1989]. Recent developments in the theory of change detection should prove useful for fault detection in the equipment and instrumentation of chemical processes. Actually, the traditional distinction [Himmelblau, 1978] between the use of model-free statistical control charts on one hand and parametric models on the other hand should no longer exist, because of the available theory and algorithms that we describe in this book.

Let us describe some possible examples of application of typical change detection algorithms to this type of processes :

- Detection of a change in the mean : Typical measurements in chemical processes contain correlations, and the traditional assumption of independence of the data when using control charts is no longer valid in this case. It is thus of interest to use the algorithms described in chapter 7 for this purpose.
- Detection of an increase in the variance : A typical strategy for reducing the production costs consists of using the linear control theory to minimize the variance of some specific quantities (concentrations, masses, etc.). This allows the producer to reach actual mean values of these quantities close to a given target value. But it should be clear that the detection of any increase in variance is of importance to avoid decreases in production quality. The algorithms described in chapter 8 should be useful in achieving this goal.
- Detection of changes in serial correlations : This problem is again related to the use of optimal control theory. An optimal controller is usually designed to ensure a specific profile of the correlation function of the output signals [Aström and Wittenmark, 1984]. Moreover, both the prediction error and the output signal in this case behave as moving average processes. This MA model can thus be used as a reference model for a change detection algorithm in order to detect changes in the profile of the correlation function. These changes indicate a loss of optimality of the controller.
- Detection of changes in regression models : In continuous-type processes, lots of equipment can be modeled as $Y_k = HX_k + V_k$, where X is the input and Y is the output. This model arises typically from the use of balance equations. On the other hand, several different types of measurement systems in such processes can be described in the same manner, exactly as we discussed for the inertial navigation system. Moreover, as we explained in chapter 7, many additive change detection problems in more complex models can be reduced to additive change detection problems in such regression models.
- Detection of additive changes in state-space models : An alternative model for continuous-type processes is the state-space model, derived, for example, from *dynamic* balance equations. This type of model can be used to detect faults in sensors and actuators, and the algorithms for detecting additive changes in state-space models described in subsection 7.2.4 are useful for this purpose.

Appendix : Models for Vibration Monitoring

In this appendix, we describe two models for simulated mechanical systems.

In [Kumamaru *et al*., 1989], a sampled damp oscillator is used, modeled by the following discrete time transfer function :

$$G(z) = \frac{b_1^0 z + b_2^0}{z^2 + a_1^0 z + a_2^0}$$

corresponding to the continuous time transfer function :

$$G(s) = \frac{\omega^2}{s^2 + 2\zeta\omega s + \omega^2}$$

where $\omega = 1$, the sampling interval is $1/2$, and the damping coefficient ζ varies between 0.2 and 0.5. A fault is here a change in this damping coefficient.

In [Basseville *et al*., 1987a], a more complex simulated system is used, which has been proven useful for testing change detection and diagnosis algorithms for large mechanical vibrating structures, such as offshore platforms, because the state-space model of this simulated system has dimension 36 and thus is large enough for testing the algorithms in a nontrivial situation. This system is a tied down system of 18 masses of one degree of freedom - six in each of three horizontal planes - connected by springs, as shown in figure 11.21, with known weights, stiffness, and damping coefficients.

The matrices M, K, and C are as follows :

$$M = \mathrm{diag}(128, 64, 64, 64, 64, 64, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32)$$

K is given by

$$\begin{pmatrix}
134 & -80 & 0 & -26 & -20 & 0 & -4 & -2 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-80 & 236 & -80 & -20 & -26 & -20 & -2 & -4 & -2 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -80 & 134 & 0 & -20 & -26 & 0 & -2 & -4 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\
-26 & -20 & 0 & 134 & -80 & 0 & -2 & 0 & 0 & -4 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-20 & -26 & -20 & -80 & 236 & -80 & 0 & -2 & 0 & -4 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -20 & -26 & 0 & -80 & 134 & 0 & 0 & -2 & 0 & -2 & -4 & 0 & 0 & 0 & 0 & 0 & 0 \\
-4 & -2 & 0 & -2 & 0 & 0 & 104 & -80 & 0 & -8 & 0 & 0 & -4 & -2 & 0 & -2 & 0 & 0 \\
-2 & -4 & -2 & 0 & -2 & 0 & -80 & 192 & -80 & 0 & -8 & -4 & -2 & -4 & -2 & 0 & -2 & 0 \\
0 & -2 & -4 & 0 & 0 & -2 & 0 & -80 & 108 & 0 & -4 & -8 & 0 & -2 & -4 & 0 & 0 & -2 \\
-2 & 0 & 0 & -4 & -2 & 0 & -8 & 0 & 0 & 104 & -80 & 0 & -2 & 0 & 0 & -4 & -2 & 0 \\
0 & -2 & 0 & -2 & -4 & -2 & 0 & -8 & -4 & -80 & 192 & -80 & 0 & -2 & 0 & -2 & -4 & -2 \\
0 & 0 & -2 & 0 & -2 & -4 & 0 & -4 & -8 & 0 & -80 & 108 & 0 & 0 & -2 & 0 & -2 & -4 \\
0 & 0 & 0 & 0 & 0 & 0 & -4 & -2 & 0 & -2 & 0 & 0 & 104 & -80 & 0 & -8 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -2 & -4 & -2 & 0 & -2 & 0 & -80 & 190 & -80 & 0 & -8 & -4 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & -4 & 0 & 0 & -2 & 0 & -80 & 108 & 0 & -4 & -8 \\
0 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & -4 & -2 & 0 & -8 & 0 & 0 & 104 & -80 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & -2 & -4 & -2 & 0 & -8 & -4 & -80 & 190 & -80 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & -2 & -4 & 0 & -4 & -8 & 0 & -80 & 108
\end{pmatrix}$$

$$C = \alpha M + \beta K$$

where $\alpha = 0.01$ and $\beta = 0.001$. The covariance matrix of the excitation is

$$Q = \mathrm{diag}(0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 3., 3., 3., 3., 3., 3., 1., 1., 1., 1., 1., 1.)$$

Finally,

$$L = (1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1)$$

In other words, six-dimensional signals corresponding to the displacements of the masses $1, 7, 13, 6, 12$, and 18 are generated.

The experiments concern 43 possible sensor locations, representing all the subsets of sensors corresponding to $r = 2, 3, 4$, and 6 and to at least one sensor on each of the two opposite "legs" 1-7-13 and 6-12-18 of the structure. These sensor locations are given in table 11.4. The **faults** that are simulated are indicated by the arrows in figure 11.21 and are as follows :

Table 11.4 The tested sensor locations for vibration monitoring.

1-1	1	6		2-2	1	7	6	12		
	1	12			1	7	6	18		
	1	18			1	7	12	18		
	7	6			1	13	6	12		
	7	12			1	13	6	18		
	7	18			1	13	12	18		
	13	6			7	13	6	12		
	13	12			7	13	6	18		
	13	18			7	13	12	18		
1-2	1	6	12	1-3	1	7	13	6		
	1	6	18		1	7	13	12		
	1	12	18		1	7	13	18		
	7	6	12		1	6	12	18		
	7	6	18		7	6	12	18		
	7	12	18		13	6	12	18		
	13	6	12							
	13	6	18							
	13	12	18							
2-1	1	7	6	3-3	1	7	13	6	12	18
	1	7	12							
	1	7	18							
	1	13	6							
	1	13	12							
	1	13	18							
	7	13	6							
	7	13	12							
	7	13	18							

- $\mathbf{H}_1$: *change of 15% in mass 1*, namely M_1 decreases from 128 to 110;
- $\mathbf{H}_2$: *change of 12% in the stiffness of the connection to the ground*, namely the stiffness coefficients of the connections of the floor masses to the ground decrease from 8 to 7;
- $\mathbf{H}_3$: *cutoff of the connection between masses 8 and 11*, namely the stiffness coefficient of this connection is set to 0.

The problem to be solved is the detection and the diagnosis of these faults.

11.3 Notes and References

Section 11.1

Navigation Systems The literature concerning model-based fault detection in inertial navigation systems is quite wide (e.g., [Newbold and Ho, 1968, Clark *et al*., 1975, Willsky *et al*., 1975, Satin and Gates, 1978, Kerr, 1980, Kerr, 1982, Huddle, 1983, Kerr, 1987, Sturza, 1988, Jeerage, 1990]). Recent investigation of the usefulness of change detection algorithms in this area can be found in [Varavva *et al*., 1988, Kireichikov *et al*., 1990, Nikiforov *et al*., 1991, Nikiforov *et al*., 1993].

Seismic Signals The use of autoregressive models for processing seismic signals in seismology was introduced in [Tjostheim, 1975]. Change detection algorithms were introduced in this area in [Kushnir *et al*., 1983, Morita and Hamaguchi, 1984, Nikiforov and Tikhonov, 1986, Pisarenko *et al*., 1987, Nikiforov *et al*., 1989, Mikhailova *et al*., 1990, Tikhonov *et al*., 1990].

Speech Signals The use of the divergence algorithm for speech signal recognition was proposed in [André-Obrecht, 1988]. In [Di Francesco, 1990], the divergence decision function is also used for processing continuous speech signals in view of coding and transmission. The idea is that, using a segmentation algorithm as a first processing step allows the design of a coder with time-varying rate, especially in the voiced segments. The main application underlying this investigation is the storage of vocal messages.

Vibration Monitoring We refer the reader to the Notes and References of chapter 9 for comments about the use of the noncentrality parameter of the instrumental test for deriving a quantitative criterion for the optimal sensor location problem. Examples of segmentation of signals in vibration mechanics can be found in [Gersch, 1986].

Section 11.2

Quality Control On-line quality control has been the historical source of change detection theory, and therefore the literature concerning the use of change detection algorithms in this area is quite wide. The main references are [Aroian and Levene, 1950, Girshick and Rubin, 1952, Barnard, 1959, Goldsmith and Whitfield, 1961, N.Johnson and Leone, 1962, Woodward and Goldsmith, 1964, Taylor, 1968, Van Dobben De Bruyn, 1968, Bissell, 1969, Phillips, 1969, Gibra, 1975, Montgomery, 1980, Vance, 1983, Montgomery, 1985, Duncan, 1986, Wetherill and Brown, 1991]

Biomedical Signals Change detection algorithms were used for processing biomedical signals in [R.Jones *et al*., 1970, Mathieu, 1976, Borodkin and Mottl', 1976, Bodenstein and Praetorius, 1977, Gustafson *et al*., 1978, Ishii *et al*., 1979, Segen and Sanderson, 1980, Appel and von Brandt, 1983, von

Brandt, 1983, Cohen, 1987]. Recognition-oriented biomedical signal processing is also treated in [Sanderson and Segen, 1980, Mottl' *et al.*, 1983].

Chemical Processes For obvious safety reasons, there were early investigations about fault detection in chemical processes, as can be seen from the book [Himmelblau, 1970]. More recent references are [Himmelblau, 1978, Watanabe and Himmelblau, 1982, Isermann, 1984].

Other Applications Change detection methods have been used in many other application domains, such as the following :

- incident detection on freeways [Willsky *et al.*, 1980];
- edge detection in image processing [Basseville *et al.*, 1981];
- geophysical signal processing [Basseville and Benveniste, 1983a];
- tracking maneuvering targets [Korn *et al.*, 1982, Favier and Smolders, 1984];
- design of reconfigurable flight control systems [Caglayan *et al.*, 1988];
- leak detection in pipelines [Isermann, 1984];
- control of air conditioning systems [Usoro, 1985];
- structural changes in econometry [Poirier, 1976, Shaban, 1980, Broemeling, 1982, Broemeling and Tsurumi, 1987, Krishnaiah and Miao, 1988];
- prediction of water municipal demand [Sastri, 1987];
- improvement of the tracking ability of adaptive algorithms [Perriot-Mathonna, 1984, Hägglund, 1983, Chen and Norton, 1987, Mariton *et al.*, 1988, Benveniste *et al.*, 1990, Wahnon and Berman, 1990].

Bibliography

H. AKAIKE (1974). Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes. *Annals Institute Statistical Mathematics*, vol.26, pp.363-387.

B.D.O. ANDERSON and J.B. MOORE (1979). *Optimal Filtering*. Information and System Sciences Series, Prentice Hall, Englewood Cliffs, NJ.

T.W. ANDERSON (1971). *The Statistical Analysis of Time Series*. Series in Probability and Mathematical Statistics, Wiley, New York.

R. ANDRE-OBRECHT (1988). A new statistical approach for the automatic segmentation of continuous speech signals. *IEEE Trans. Acoustics, Speech, Signal Processing*, vol.ASSP-36, no 1, pp.29-40.

R. ANDRE-OBRECHT (1990). Reconnaissance automatique de parole à partir de segments acoustiques et de modèles de Markov cachés. *Proc. Journées Etude de la Parole*, Montréal, May 1990 (in French).

R. ANDRE-OBRECHT and H.Y. SU (1988). Three acoustic labellings for phoneme based continuous speech recognition. *Proc. Speech'88*, Edinburgh, UK, pp.943-950.

U. APPEL and A. VON BRANDT (1983). Adaptive sequential segmentation of piecewise stationary time series. *Information Sciences*, vol.29, no 1, pp.27-56.

L.A. AROIAN and H. LEVENE (1950). The effectiveness of quality control procedures. *Jal American Statistical Association*, vol.45, pp.520-529.

K.J. ASTRÖM and B. WITTENMARK (1984). *Computer Controlled Systems : Theory and Design*. Information and System Sciences Series, Prentice Hall, Englewood Cliffs, NJ.

M. BAGSHAW and R.A. JOHNSON (1975a). The effect of serial correlation on the performance of CUSUM tests - Part II. *Technometrics*, vol.17, no 1, pp.73-80.

M. BAGSHAW and R.A. JOHNSON (1975b). The influence of reference values and estimated variance on the ARL of CUSUM tests. *Jal Royal Statistical Society*, vol.37(B), no 3, pp.413-420.

M. BAGSHAW and R.A. JOHNSON (1977). Sequential procedures for detecting parameter changes in a time-series model. *Jal American Statistical Association*, vol.72, no 359, pp.593-597.

R.K. BANSAL and P. PAPANTONI-KAZAKOS (1986). An algorithm for detecting a change in a stochastic process. *IEEE Trans. Information Theory*, vol.IT-32, no 2, pp.227-235.

G.A. BARNARD (1959). Control charts and stochastic processes. *Jal Royal Statistical Society*, vol.B.21, pp.239-271.

A.E. BASHARINOV and B.S. FLEISHMAN (1962). *Methods of the statistical sequential analysis and their radiotechnical applications*. Sovetskoe Radio, Moscow (in Russian).

M. BASSEVILLE (1978). Déviations par rapport au maximum: formules d'arrêt et martingales associées. *Compte-rendus du Séminaire de Probabilités*, Université de Rennes I.

M. BASSEVILLE (1981). Edge detection using sequential methods for change in level - Part II : Sequential detection of change in mean. *IEEE Trans. Acoustics, Speech, Signal Processing*, vol.ASSP-29, no 1, pp.32-50.

M. BASSEVILLE (1982). A survey of statistical failure detection techniques. In *Contribution à la Détection Séquentielle de Ruptures de Modèles Statistiques*, Thèse d'Etat, Université de Rennes I, France (in English).

M. BASSEVILLE (1986). The two-models approach for the on-line detection of changes in AR processes. In *Detection of Abrupt Changes in Signals and Dynamical Systems* (M. Basseville, A. Benveniste, eds.). Lecture Notes in Control and Information Sciences, LNCIS 77, Springer, New York, pp.169-215.

M. BASSEVILLE (1988). Detecting changes in signals and systems - A survey. *Automatica*, vol.24, pp.309-326.

M. BASSEVILLE (1989). Distance measures for signal processing and pattern recognition. *Signal Processing*, vol.18, pp.349-369.

M. BASSEVILLE and A. BENVENISTE (1983a). Design and comparative study of some sequential jump detection algorithms for digital signals. *IEEE Trans. Acoustics, Speech, Signal Processing*, vol.ASSP-31, no 3, pp.521-535.

M. BASSEVILLE and A. BENVENISTE (1983b). Sequential detection of abrupt changes in spectral characteristics of digital signals. *IEEE Trans. Information Theory*, vol.IT-29, no 5, pp.709-724.

M. BASSEVILLE and A. BENVENISTE, eds. (1986). *Detection of Abrupt Changes in Signals and Dynamical Systems*. Lecture Notes in Control and Information Sciences, LNCIS 77, Springer, New York.

M. BASSEVILLE and I. NIKIFOROV (1991). A unified framework for statistical change detection. *Proc. 30th IEEE Conference on Decision and Control*, Brighton, UK.

M. BASSEVILLE, B. ESPIAU and J. GASNIER (1981). Edge detection using sequential methods for change in level - Part I : A sequential edge detection algorithm. *IEEE Trans. Acoustics, Speech, Signal Processing*, vol.ASSP-29, no 1, pp.24-31.

M. BASSEVILLE, A. BENVENISTE and G. MOUSTAKIDES (1986). Detection and diagnosis of abrupt changes in modal characteristics of nonstationary digital signals. *IEEE Trans. Information Theory*, vol.IT-32, no 3, pp.412-417.

M. BASSEVILLE, A. BENVENISTE, G. MOUSTAKIDES and A. ROUGÉE (1987a). Detection and diagnosis of changes in the eigenstructure of nonstationary multivariable systems. *Automatica*, vol.23, no 3, pp.479-489.

M. BASSEVILLE, A. BENVENISTE, G. MOUSTAKIDES and A. ROUGÉE (1987b). Optimal sensor location for detecting changes in dynamical behavior. *IEEE Trans. Automatic Control*, vol.AC-32, no 12, pp.1067-1075.

M. BASSEVILLE, A. BENVENISTE, B. GACH-DEVAUCHELLE, M. GOURSAT, D. BONNECASE, P. DOREY, M. PREVOSTO and M. OLAGNON (1993). Damage monitoring in vibration mechanics : issues in diagnostics and predictive maintenance. *Mechanical Systems and Signal Processing*, vol.7, no 5, pp.401-423.

R.V. BEARD (1971). *Failure Accommodation in Linear Systems through Self-reorganization*. Ph.D.Thesis, Dept. Aeronautics and Astronautics, MIT, Cambridge, MA.

A. BENVENISTE and J.J. FUCHS (1985). Single sample modal identification of a nonstationary stochastic process. *IEEE Trans. Automatic Control*, vol.AC-30, no 1, pp.66-74.

A. BENVENISTE, M. BASSEVILLE and G. MOUSTAKIDES (1987). The asymptotic local approach to change detection and model validation. *IEEE Trans. Automatic Control*, vol.AC-32, no 7, pp.583-592.

A. BENVENISTE, M. METIVIER and P. PRIOURET (1990). *Adaptive Algorithms and Stochastic Approximations*. Series on Applications of Mathematics, (A.V. Balakrishnan, I. Karatzas, M. Yor, eds.). Springer, New York.

A. BENVENISTE, M. BASSEVILLE, L. EL GHAOUI, R. NIKOUKHAH and A.S. WILLSKY (1992). An optimum robust approach to statistical failure detection and identification. IFAC World Conference, Sydney, July 1993.

R.H. BERK (1973). Some asymptotic aspects of sequential analysis. *Annals Statistics*, vol.1, no 6, pp.1126-1138.

R.H. BERK (1975). Locally most powerful sequential test. *Annals Statistics*, vol.3, no 2, pp.373-381.

P. BILLINGSLEY (1968). *Convergence of Probability Measures*. Wiley, New York.

A.F. BISSELL (1969). Cusum techniques for quality control. *Applied Statistics*, vol.18, pp.1-30.

M.E. BIVAIKOV (1991). Control of the sample size for recursive estimation of parameters subject to abrupt changes. *Automation and Remote Control*, no 9, pp.96-103.

R.E. BLAHUT (1987). *Principles and Practice of Information Theory*. Addison-Wesley, Reading, MA.

I.F. BLAKE and W.C. LINDSEY (1973). Level-crossing problems for random processes. *IEEE Trans. Information Theory*, vol.IT-19, no 3, pp.295-315.

G. BODENSTEIN and H.M. PRAETORIUS (1977). Feature extraction from the encephalogram by adaptive segmentation. *Proc. IEEE*, vol.65, pp.642-652.

T. BOHLIN (1977). Analysis of EEG signals with changing spectra using a short word Kalman estimator. *Mathematical Biosciences*, vol.35, pp.221-259.

W. BÖHM and P. HACKL (1990). Improved bounds for the average run length of control charts based on finite weighted sums. *Annals Statistics*, vol.18, no 4, pp.1895-1899.

T. BOJDECKI and J. HOSZA (1984). On a generalized disorder problem. *Stochastic Processes and their Applications*, vol.18, pp.349-359.

L.I. BORODKIN and V.V. MOTTL' (1976). Algorithm for finding the jump times of random process equation parameters. *Automation and Remote Control*, vol.37, no 6, Part 1, pp.23-32.

A.A. BOROVKOV (1984). *Theory of Mathematical Statistics - Estimation and Hypotheses Testing*, Naouka, Moscow (in Russian). Translated in French under the title *Statistique Mathématique - Estimation et Tests d'Hypothèses*, Mir, Paris, 1987.

G.E.P. BOX and G.M. JENKINS (1970). *Time Series Analysis, Forecasting and Control*. Series in Time Series Analysis, Holden-Day, San Francisco.

A. VON BRANDT (1983). Detecting and estimating parameters jumps using ladder algorithms and likelihood ratio test. *Proc.ICASSP*, Boston, MA, pp.1017-1020.

A. VON BRANDT (1984). *Modellierung von Signalen mit Sprunghaft Veränderlichem Leistungsspektrum durch Adaptive Segmentierung*. Doctor-Engineer Dissertation, München, RFA (in German).

S. BRAUN, ed. (1986). *Mechanical Signature Analysis - Theory and Applications*. Academic Press, London.

L. BREIMAN (1968). *Probability*. Series in Statistics, Addison-Wesley, Reading, MA.

G.S. BRITOV and L.A. MIRONOVSKI (1972). Diagnostics of linear systems of automatic regulation. *Tekh. Kibernetics*, vol.1, pp.76-83.

B.E. BRODSKIY and B.S. DARKHOVSKIY (1992). *Nonparametric Methods in Change-point Problems*. Kluwer Academic, Boston.

L.D. BROEMELING (1982). *Jal Econometrics*, vol.19, Special issue on structural change in Econometrics.

L.D. BROEMELING and H. TSURUMI (1987). *Econometrics and Structural Change*. Dekker, New York.

D. BROOK and D.A. EVANS (1972). An approach to the probability distribution of Cusum run length. *Biometrika*, vol.59, pp.539-550.

J. BRUNET, D. JAUME, M. LABARRÈRE, A. RAULT and M. VERGÉ (1990). *Détection et Diagnostic de Pannes*. Traité des Nouvelles Technologies, Série Diagnostic et Maintenance, Hermès, Paris (in French).

S.P. BRUZZONE and M. KAVEH (1984). Information tradeoffs in using the sample autocorrelation function in ARMA parameter estimation. *IEEE Trans. Acoustics, Speech, Signal Processing*, vol.ASSP-32, no 4, pp.701-715.

A.K. CAGLAYAN (1980). Necessary and sufficient conditions for detectability of jumps in linear systems. *IEEE Trans. Automatic Control*, vol.AC-25, no 4, pp.833-834.

A.K. CAGLAYAN and R.E. LANCRAFT (1983). Reinitialization issues in fault tolerant systems. *Proc. American Control Conf.*, pp.952-955.

A.K. CAGLAYAN, S.M. ALLEN and K. WEHMULLER (1988). Evaluation of a second generation reconfiguration strategy for aircraft flight control systems subjected to actuator failure/surface damage. *Proc. National Aerospace and Electronic Conference*, Dayton, OH.

P.E. CAINES (1988). *Linear Stochastic Systems*. Series in Probability and Mathematical Statistics, Wiley, New York.

M.J. CHEN and J.P. NORTON (1987). Estimation techniques for tracking rapid parameter changes. *Intern. Jal Control*, vol.45, no 4, pp.1387-1398.

W.K. CHIU (1974). The economic design of cusum charts for controlling normal mean. *Applied Statistics*, vol.23, no 3, pp.420-433.

E.Y. CHOW (1980). *A Failure Detection System Design Methodology*. Ph.D.Thesis, M.I.T., L.I.D.S., Cambridge, MA.

E.Y. CHOW and A.S. WILLSKY (1984). Analytical redundancy and the design of robust failure detection systems. *IEEE Trans. Automatic Control*, vol.AC-29, no 3, pp.689-691.

Y.S. CHOW, H. ROBBINS and D. SIEGMUND (1971). *Great Expectations : The Theory of Optimal Stopping*. Houghton-Mifflin, Boston.

R.N. CLARK, D.C. FOSTH and V.M. WALTON (1975). Detection of instrument malfunctions in control systems. *IEEE Trans. Aerospace Electronic Systems*, vol.AES-11, pp.465-473.

A. COHEN (1987). *Biomedical Signal Processing - vol.1: Time and Frequency Domain Analysis; vol.2: Compression and Automatic Recognition*. CRC Press, Boca Raton, FL.

J. CORGE and F. PUECH (1986). Analyse du rythme cardiaque foetal par des méthodes de détection de ruptures. *Proc. 7th INRIA Int. Conf. Analysis and optimization of Systems*. Antibes, FR (in French).

D.R. COX and D.V. HINKLEY (1986). *Theoretical Statistics*. Chapman and Hall, New York.

D.R. COX and H.D. MILLER (1965). *The Theory of Stochastic Processes*. Wiley, New York.

S.V. CROWDER (1987). A simple method for studying run-length distributions of exponentially weighted moving average charts. *Technometrics*, vol.29, no 4, pp.401-407.

H. CSÖRGÖ and L. HORVÁTH (1988). Nonparametric methods for change point problems. In *Handbook of Statistics* (P.R. Krishnaiah, C.R. Rao, eds.), vol.7, Elsevier, New York, pp.403-425.

R.B. DAVIES (1973). Asymptotic inference in stationary gaussian time series. *Advances Applied Probability*, vol.5, no 3, pp.469-497.

J.C. DECKERT, M.N. DESAI, J.J. DEYST and A.S. WILLSKY (1977). F-8 DFBW sensor failure identification using analytical redundancy. *IEEE Trans. Automatic Control*, vol.AC-22, no 5, pp.795-803.

M.H. DE GROOT (1970). *Optimal Statistical Decisions*. Series in Probability and Statistics, McGraw-Hill, New York.

J. DESHAYES and D. PICARD (1979). Tests de ruptures dans un modèle. *Compte-Rendus de l'Académie des Sciences*, vol.288, Ser.A, pp.563-566 (in French).

J. DESHAYES and D. PICARD (1983). *Ruptures de Modèles en Statistique*. Thèses d'Etat, Université de Paris-Sud, Orsay, France (in French).

J. DESHAYES and D. PICARD (1986). Off-line statistical analysis of change-point models using non parametric and likelihood methods. In *Detection of Abrupt Changes in Signals and Dynamical Systems* (M. Basseville, A. Benveniste, eds.). Lecture Notes in Control and Information Sciences, LNCIS 77, Springer, New York, pp.103-168.

B. DEVAUCHELLE-GACH (1991). *Diagnostic Mécanique des Fatigues sur les Structures Soumises à des Vibrations en Ambiance de Travail*. Thèse de l'Université Paris IX Dauphine (in French).

B. DEVAUCHELLE-GACH, M. BASSEVILLE and A. BENVENISTE (1991). Diagnosing mechanical changes in vibrating systems. *Proc. SAFEPROCESS'91*, Baden-Baden, FRG, pp.85-89.

R. DI FRANCESCO (1990). Real-time speech segmentation using pitch and convexity jump models: application to variable rate speech coding. *IEEE Trans. Acoustics, Speech, Signal Processing*, vol.ASSP-38, no 5, pp.741-748.

X. DING and P.M. FRANK (1990). Fault detection via factorization approach. *Systems and Control Letters*, vol.14, pp.431-436.

J.L. DOOB (1953). *Stochastic Processes*. Wiley, New York.

V. DRAGALIN (1988). Asymptotic solutions in detecting a change in distribution under an unknown parameter. *Statistical Problems of Control*, Issue 83, Vilnius, pp.45-52.

B. DUBUISSON (1990). *Diagnostic et Reconnaissance des Formes*. Traité des Nouvelles Technologies, Série Diagnostic et Maintenance, Hermès, Paris (in French).

A.J. DUNCAN (1986). *Quality Control and Industrial Statistics*, 5th edition. Richard D.Irwin, Inc., Homewood, IL.

J. DURBIN (1971). Boundary-crossing probabilities for the Brownian motion and Poisson processes and techniques for computing the power of the Kolmogorov-Smirnov test. *Jal Applied Probability*, vol.8, pp.431-453.

J. DURBIN (1985). The first passage density of the crossing of a continuous Gaussian process to a general boundary. *Jal Applied Probability*, vol.22, no 1, pp.99-122.

A. EMAMI-NAEINI, M.M. AKHTER and S.M. ROCK (1988). Effect of model uncertainty on failure detection : the threshold selector. *IEEE Trans. Automatic Control*, vol.AC-33, no 12, pp.1106-1115.

J.D. ESARY, F. PROSCHAN and D.W. WALKUP (1967). Association of random variables with applications. *Annals Mathematical Statistics*, vol.38, pp.1466-1474.

W.D. EWAN and K.W. KEMP (1960). Sampling inspection of continuous processes with no autocorrelation between successive results. *Biometrika*, vol.47, pp.263-280.

G. FAVIER and A. SMOLDERS (1984). Adaptive smoother-predictors for tracking maneuvering targets. *Proc. 23rd Conf. Decision and Control*, Las Vegas, NV, pp.831-836.

W. FELLER (1966). *An Introduction to Probability Theory and Its Applications*, vol.2. Series in Probability and Mathematical Statistics, Wiley, New York.

R.A. FISHER (1925). Theory of statistical estimation. *Proc. Cambridge Philosophical Society*, vol.22, pp.700-725.

M. FISHMAN (1988). Optimization of the algorithm for the detection of a disorder, based on the statistic of exponential smoothing. In *Statistical Problems of Control*, Issue 83, Vilnius, pp.146-151.

R. FLETCHER (1980). *Practical Methods of Optimization*, 2 volumes. Wiley, New York.

P.M. FRANK (1990). Fault diagnosis in dynamic systems using analytical and knowledge based redundancy - A survey and new results. *Automatica*, vol.26, pp.459-474.

P.M. FRANK (1991). Enhancement of robustness in observer-based fault detection. *Proc. SAFEPROCESS'91*, Baden-Baden, FRG, pp.275-287.

P.M. FRANK and J. WÜNNENBERG (1989). Robust fault diagnosis using unknown input observer schemes. In *Fault Diagnosis in Dynamic Systems - Theory and Application* (R. Patton, P. Frank, R. Clark, eds.). International Series in Systems and Control Engineering, Prentice Hall International, London, UK, pp.47-98.

K. FUKUNAGA (1990). *Introduction to Statistical Pattern Recognition*, 2d ed. Academic Press, New York.

S.I. GASS (1958). *Linear Programming : Methods and Applications*. McGraw Hill, New York.

W. GE and C.Z. FANG (1989). Extended robust observation approach for failure isolation. *Int. Jal Control*, vol.49, no 5, pp.1537-1553.

W. GERSCH (1986). Two applications of parametric time series modeling methods. In *Mechanical Signature Analysis - Theory and Applications* (S. Braun, ed.), chap.10. Academic Press, London.

J.J. GERTLER (1988). Survey of model-based failure detection and isolation in complex plants. *IEEE Control Systems Magazine*, vol.8, no 6, pp.3-11.

J.J. GERTLER (1991). Analytical redundancy methods in fault detection and isolation. *Proc. SAFEPROCESS'91*, Baden-Baden, FRG, pp.9-22.

B.K. GHOSH (1970). *Sequential Tests of Statistical Hypotheses*. Addison-Wesley, Cambridge, MA.

I.N. GIBRA (1975). Recent developments in control charts techniques. *Jal Quality Technology*, vol.7, pp.183-192.

J.P. GILMORE and R.A. MCKERN (1972). A redundant strapdown inertial reference unit (SIRU). *Jal Spacecraft*, vol.9, pp.39-47.

M.A. GIRSHICK and H. RUBIN (1952). A Bayes approach to a quality control model. *Annals Mathematical Statistics*, vol.23, pp.114-125.

A.L. GOEL and S.M. WU (1971). Determination of the ARL and a contour nomogram for CUSUM charts to control normal mean. *Technometrics*, vol.13, no 2, pp.221-230.

P.L. GOLDSMITH and H. WHITFIELD (1961). Average run lengths in cumulative chart quality control schemes. *Technometrics*, vol.3, pp.11-20.

G.C. GOODWIN and K.S. SIN (1984). *Adaptive Filtering, Prediction and Control*. Information and System Sciences Series, Prentice Hall, Englewood Cliffs, NJ.

R.M. GRAY and L.D. DAVISSON (1986). *Random Processes : a Mathematical Approach for Engineers*. Information and System Sciences Series, Prentice Hall, Englewood Cliffs, NJ.

C. GUEGUEN and L.L. SCHARF (1980). Exact maximum likelihood identification for ARMA models: a signal processing perspective. *Proc. 1st EUSIPCO*, Lausanne.

D.E. GUSTAFSON, A.S. WILLSKY, J.Y. WANG, M.C. LANCASTER and J.H. TRIEBWASSER (1978). ECG/VCG rhythm diagnosis using statistical signal analysis. Part I: Identification of persistent rhythms. Part II: Identification of transient rhythms. *IEEE Trans. Biomedical Engineering*, vol.BME-25, pp.344-353 and 353-361.

F. GUSTAFSSON (1991). Optimal segmentation of linear regression parameters. *Proc. IFAC/IFORS Symp. Identification and System Parameter Estimation*, Budapest, pp.225-229.

T. HÄGGLUND (1983). *New Estimation Techniques for Adaptive Control*. Ph.D.Thesis, Lund Institute of Technology, Lund, Sweden.

T. HÄGGLUND (1984). Adaptive control of systems subject to large parameter changes. *Proc. IFAC 9th World Congress*, Budapest.

P. HALL and C.C. HEYDE (1980). *Martingale Limit Theory and its Application*. Probability and Mathematical Statistics, a Series of Monographs and Textbooks, Academic Press, New York.

W.J. HALL, R.A. WIJSMAN and J.K. GHOSH (1965). The relationship between sufficiency and invariance with applications in sequential analysis. *Ann. Math. Statist.*, vol.36, pp.576-614.

E.J. HANNAN and M. DEISTLER (1988). *The Statistical Theory of Linear Systems*. Series in Probability and Mathematical Statistics, Wiley, New York.

J.D. HEALY (1987). A note on multivariate CuSum procedures. *Technometrics*, vol.29, pp.402-412.

D.M. HIMMELBLAU (1970). *Process Analysis by Statistical Methods*. Wiley, New York.

D.M. HIMMELBLAU (1978). *Fault Detection and Diagnosis in Chemical and Petrochemical Processes*. Chemical Engineering Monographs, vol.8, Elsevier, Amsterdam.

W.G.S. HINES (1976a). A simple monitor of a system with sudden parameter changes. *IEEE Trans. Information Theory*, vol.IT-22, no 2, pp.210-216.

W.G.S. HINES (1976b). Improving a simple monitor of a system with sudden parameter changes. *IEEE Trans. Information Theory*, vol.IT-22, no 4, pp.496-499.

D.V. HINKLEY (1969). Inference about the intersection in two-phase regression. *Biometrika*, vol.56, no 3, pp.495-504.

D.V. HINKLEY (1970). Inference about the change point in a sequence of random variables. *Biometrika*, vol.57, no 1, pp.1-17.

D.V. HINKLEY (1971). Inference about the change point from cumulative sum-tests. *Biometrika*, vol.58, no 3, pp.509-523.

D.V. HINKLEY (1971). Inference in two-phase regression. *Jal American Statistical Association*, vol.66, no 336, pp.736-743.

J.R. HUDDLE (1983). Inertial navigation system error-model considerations in Kalman filtering applications. In *Control and Dynamic Systems* (C.T. Leondes, ed.), Academic Press, New York, pp.293-339.

J.S. HUNTER (1986). The exponentially weighted moving average. *Jal Quality Technology*, vol.18, pp.203-210.

I.A. IBRAGIMOV and R.Z. KHASMINSKII (1981). *Statistical Estimation - Asymptotic Theory*. Applications of Mathematics Series, vol.16. Springer, New York.

R. ISERMANN (1984). Process fault detection based on modeling and estimation methods - A survey. *Automatica*, vol.20, pp.387-404.

N. ISHII, A. IWATA and N. SUZUMURA (1979). Segmentation of nonstationary time series. *Int. Jal Systems Sciences*, vol.10, pp.883-894.

J.E. JACKSON and R.A. BRADLEY (1961). Sequential χ^2 and T^2 tests. *Annals Mathematical Statistics*, vol.32, pp.1063-1077.

B. JAMES, K.L. JAMES and D. SIEGMUND (1988). Conditional boundary crossing probabilities with applications to change-point problems. *Annals Probability*, vol.16, pp.825-839.

M.K. JEERAGE (1990). Reliability analysis of fault-tolerant IMU architectures with redundant inertial sensors. *IEEE Trans. Aerospace and Electronic Systems*, vol.AES-5, no.7, pp.23-27.

N.L. JOHNSON (1961). A simple theoretical approach to cumulative sum control charts. *Jal American Statistical Association*, vol.56, pp.835-840.

N.L. JOHNSON and F.C. LEONE (1962). Cumulative sum control charts: mathematical principles applied to their construction and use. Parts I,II,III. *Industrial Quality Control*, vol.18, pp.15-21; vol.19, pp.29-36; vol.20, pp.22-28.

R.A. JOHNSON and M. BAGSHAW (1974). The effect of serial correlation on the performance of CUSUM tests - Part I. *Technometrics*, vol.16, no.1, pp.103-112.

H.L. JONES (1973). *Failure Detection in Linear Systems*. Ph.D.Thesis, Dept. Aeronautics and Astronautics, MIT, Cambridge, MA.

R.H. JONES, D.H. CROWELL and L.E. KAPUNIAI (1970). Change detection model for serially correlated multivariate data. *Biometrics*, vol.26, no 2, pp.269-280.

M. JURGUTIS (1984). Comparison of the statistical properties of the estimates of the change times in an autoregressive process. In *Statistical Problems of Control*, Issue 65, Vilnius, pp.234-243 (in Russian).

T. KAILATH (1980). *Linear Systems*. Information and System Sciences Series, Prentice Hall, Englewood Cliffs, NJ.

L. V.KANTOROVICH and V.I. KRILOV (1958). *Approximate Methods of Higher Analysis*. Interscience, New York.

S. KARLIN and H.M. TAYLOR (1975). *A First Course in Stochastic Processes*, 2d ed. Academic Press, New York.

S. KARLIN and H.M. TAYLOR (1981). *A Second Course in Stochastic Processes*. Academic Press, New York.

D. KAZAKOS and P. PAPANTONI-KAZAKOS (1980). Spectral distance measures between gaussian processes. *IEEE Trans. Automatic Control*, vol.AC-25, no 5, pp.950-959.

K.W. KEMP (1958). Formula for calculating the operating characteristic and average sample number of some sequential tests. *Jal Royal Statistical Society*, vol.B-20, no 2, pp.379-386.

K.W. KEMP (1961). The average run length of the cumulative sum chart when a V-mask is used. *Jal Royal Statistical Society*, vol.B-23, pp.149-153.

K.W. KEMP (1967a). Formal expressions which can be used for the determination of operating characteristics and average sample number of a simple sequential test. *Jal Royal Statistical Society*, vol.B-29, no 2, pp.248-262.

K.W. KEMP (1967b). A simple procedure for determining upper and lower limits for the average sample run length of a cumulative sum scheme. *Jal Royal Statistical Society*, vol.B-29, no 2, pp.263-265.

D.P. KENNEDY (1976). Some martingales related to cumulative sum tests and single server queues. *Stochastic Processes and Appl.*, vol.4, pp.261-269.

T.H. KERR (1980). Statistical analysis of two-ellipsoid overlap test for real time failure detection. *IEEE Trans. Automatic Control*, vol.AC-25, no 4, pp.762-772.

T.H. KERR (1982). False alarm and correct detection probabilities over a time interval for restricted classes of failure detection algorithms. *IEEE Trans. Information Theory*, vol.IT- 24, pp.619-631.

T.H. KERR (1987). Decentralized filtering and redundancy management for multisensor navigation. *IEEE Trans. Aerospace and Electronic systems*, vol.AES-23, pp.83-119. Minor corrections on p.412 and p.599 (May and July issues, respectively).

R.A. KHAN (1978). Wald's approximations to the average run length in cusum procedures. *Jal Statistical Planning and Inference*, vol.2, no 1, pp.63-77.

R.A. KHAN (1979). Some first passage problems related to cusum procedures. *Stochastic Processes and Applications*, vol.9, no 2, pp.207-215.

R.A. KHAN (1981). A note on Page's two-sided cumulative sum procedures. *Biometrika*, vol.68, no 3, pp.717-719.

V. KIREICHIKOV, V. MANGUSHEV and I. NIKIFOROV (1990). Investigation and application of CUSUM algorithms to monitoring of sensors. In *Statistical Problems of Control*, Issue 89, Vilnius, pp.124-130 (in Russian).

G. KITAGAWA and W. GERSCH (1985). A smoothness prior time-varying AR coefficient modeling of non-stationary covariance time series. *IEEE Trans. Automatic Control*, vol.AC-30, no 1, pp.48-56.

N. KLIGIENE (1980). Probabilities of deviations of the change point estimate in statistical models. In *Statistical Problems of Control*, Issue 83, Vilnius, pp.80-86 (in Russian).

N. KLIGIENE and L. TELKSNYS (1983). Methods of detecting instants of change of random process properties. *Automation and Remote Control*, vol.44, no 10, Part II, pp.1241-1283.

J. KORN, S.W. GULLY and A.S. WILLSKY (1982). Application of the generalized likelihood ratio algorithm to maneuver detection and estimation. *Proc. American Control Conf.*, Arlington, VA, pp.792-798.

P.R. KRISHNAIAH and B.Q. MIAO (1988). Review about estimation of change points. In *Handbook of Statistics* (P.R. Krishnaiah, C.R. Rao, eds.), vol.7, Elsevier, New York, pp.375-402.

P. KUDVA, N. VISWANADHAM and A. RAMAKRISHNAN (1980). Observers for linear systems with unknown inputs. *IEEE Trans. Automatic Control*, vol.AC-25, no 1, pp.113-115.

S. KULLBACK (1959). *Information Theory and Statistics*. Wiley, New York (also Dover, New York, 1968).

K. KUMAMARU, S. SAGARA and T. SÖDERSTRÖM (1989). Some statistical methods for fault diagnosis for dynamical systems. In *Fault Diagnosis in Dynamic Systems - Theory and Application* (R. Patton, P. Frank, R. Clark, eds.). International Series in Systems and Control Engineering, Prentice Hall International, London, UK, pp.439-476.

A. KUSHNIR, I. NIKIFOROV and I. SAVIN (1983). Statistical adaptive algorithms for automatic detection of seismic signals - Part I : One-dimensional case. In *Earthquake Prediction and the Study of the Earth Structure*, Naouka, Moscow (*Computational Seismology*, vol.15), pp.154-159 (in Russian).

L. LADELLI (1990). Diffusion approximation for a pseudo-likelihood test process with application to detection of change in stochastic system. *Stochastics and Stochastics Reports*, vol.32, pp.1-25.

T.L. LAÏ (1974). Control charts based on weighted sums. *Annals Statistics*, vol.2, no 1, pp.134-147.

T.L. LAÏ (1981). Asymptotic optimality of invariant sequential probability ratio tests. *Annals Statistics*, vol.9, no 2, pp.318-333.

D.G. LAINIOTIS (1971). Joint detection, estimation, and system identification. *Information and Control*, vol.19, pp.75-92.

M.R. LEADBETTER, G. LINDGREN and H. ROOTZEN (1983). *Extremes and Related Properties of Random Sequences and Processes*. Series in Statistics, Springer, New York.

L. LE CAM (1960). Locally asymptotically normal families of distributions. *Univ. California Publications in Statistics*, vol.3, pp.37-98.

L. LE CAM (1986). *Asymptotic Methods in Statistical Decision Theory*. Series in Statistics, Springer, New York.

E.L. LEHMANN (1986). *Testing Statistical Hypotheses*, 2d ed. Wiley, New York.

J.P. LEHOCZKY (1977). Formulas for stopped diffusion processes with stopping times based on the maximum. *Annals Probability*, vol.5, no 4, pp.601-607.

H.R. LERCHE (1980). *Boundary Crossing of Brownian Motion*. Lecture Notes in Statistics, vol.40, Springer, New York.

L. LJUNG (1987). *System Identification - Theory for the User*. Information and System Sciences Series, Prentice Hall, Englewood Cliffs, NJ.

L. LJUNG and S. GUNNARSSON (1990). Adaptation and tracking in system identification - A survey. *Automatica*, vol.26, no 1, pp.7-22.

M. LOEVE (1964). *Probability Theory*, 3d ed. Van Nostrand, Princeton, NJ. Also 4th edition in two volumes in the Graduate Texts in Mathematics Series, Springer, New York.

G. LORDEN (1970). On excess over the boundary. *Annals Mathematical Statistics*, vol.41, no 2, pp.520-527.

G. LORDEN (1971). Procedures for reacting to a change in distribution. *Annals Mathematical Statistics*, vol.42, pp.1897-1908.

G. LORDEN (1973). Open-ended tests for Koopman-Darmois families. *Annals Statistics*, vol.1, no 4, pp.633-643.

G. LORDEN and I. EISENBERGER (1973). Detection of failure rate increases. *Technometrics*, vol.15, no 1, pp.167-175.

X.-C. LOU, A.S. WILLSKY and G.C. VERGHEESE (1986). Optimally robust redundancy relations for failure detection in uncertain systems. *Automatica*, vol.22, no 3, pp.333-344.

J.M. LUCAS and R.B. CROSIER (1982). Fast initial response for CUSUM quality control schemes : give your CUSUM a head start. *Technometrics*, vol.24, no 3, pp.199-205.

D.G. LUENBERGER (1984). *Linear and Nonlinear Programming*, 2d ed. Addison Wesley, Reading, MA.

V.YA. LUMEL'SKY (1972). Algorithm for detecting the time of change in properties of a random process. *Automation and Remote Control*, vol.33, no 11, Part I, pp.1620-1625.

R.J. MACAULAY and E. DENLINGER (1973). A decision-directed adaptive tracker. *IEEE Trans. Aerospace and Electronic Systems*, vol.AES-9, pp.229-236.

G.P. MACCORMICK (1983). *Nonlinear Programming: Theory, Algorithms and Applications*. Wiley, New York.

C.A. MACKEN and H.M. TAYLOR (1977). On deviations from the maximum in a stochastic process. *SIAM Jal Applied Mathematics*, vol.32, no 1, pp.96-104.

I.B. MACNEILL (1974). Tests for change of parameters at unknown times and distributions of some related functionals on Brownian motions. *Annals Statistics*, vol.2, no 5, pp.950-962.

M. MARITON (1990). *Jump Linear Systems in Automatic Control*. Dekker, New York.

M. MARITON, P. BERTRAND and C. YANG (1988). Systems subject to Markovian parameter changes: detection and adaptive estimation. *Proc. 12th IMACS World Congress Scientific Computation*, Paris, July 1988, vol.1.

J.D. MARKEL and A.H. GRAY (1976). *Linear Prediction of Speech*. Springer, New York.

M.A. MASSOUMNIA (1986). A geometric approach to the synthesis of failure detection filters. *IEEE Trans. Automatic Control*, vol.AC-31, no 5, pp.839-846.

M.A. MASSOUMNIA and W.E. VANDER VELDE (1988). Generating parity relations for detecting and identifying control systems component failures. *Jal Guidance, Control and Dynamics*, vol.11, no 1, pp.60-65.

M.A. MASSOUMNIA, G.C. VERGHEESE and A.S. WILLSKY (1989). Failure detection and identification. *IEEE Trans. Automatic Control*, vol.AC-34, no 3, pp.316-321.

M. MATHIEU (1976). *Analyse de l'Electroencéphalogramme par Prédiction Linéaire*. Thèse Docteur Ingénieur, ENST, Paris, France (in French).

R.K. MEHRA (1970). On the identification of variances and adaptive Kalman filtering. *IEEE Trans. Automatic Control*, vol.AC-15, no 2, pp.175-184.

R.K. MEHRA and J. PESCHON (1971). An innovations approach to fault detection and diagnosis in dynamic systems. *Automatica*, vol.7, pp.637-640.

W.C. MERRILL (1985). Sensor failure detection for jet engines using analytical redundancy. *Jal Guidance, Control and Dynamics*, vol.8, pp.673-682.

T.G. MIKHAILOVA, I.N. TIKHONOV, V.A. MANGUSHEV and I.V. NIKIFOROV (1990). Automatic identification of overlapping of signals produced by two subsequent shocks on a seismogram. *Vulcanology and Seismology*, no 5, pp.94-102.

L.A. MIRONOVSKI (1979). Functional diagnosis of linear dynamic systems. *Automation and Remote Control*, vol.40, pp.120-128.

L.A. MIRONOVSKI (1980). Functional diagnosis of dynamic systems - A survey. *Automation and Remote Control*, vol.41, pp.1122-1143.

D.C. MONTGOMERY (1980). The economic design of control charts: a review and literature survey. *Jal Quality Technology*, vol.12, pp.75-87.

D.C. MONTGOMERY (1985). *Introduction to Statistical Quality Control*. Wiley, New York.

Y. MORITA and H. HAMAGUCHI (1984). Automatic detection of onset time of seismic waves and its confidence interval using the autoregressive model fitting. *ZISIN*, vol.37, no 2, pp.281-293.

V.V. MOTTL', I.B. MUCHNIK and V.G. YAKOVLEV (1983). Optimal segmentation of experimental curves. *Automation and Remote Control*, vol.44, pp.1035-1044.

G. MOUSTAKIDES (1986). Optimal procedures for detecting changes in distributions. *Annals Statistics*, vol.14, pp.1379-1387.

G. MOUSTAKIDES and A. BENVENISTE (1986). Detecting changes in the AR parameters of a nonstationary ARMA process. *Stochastics*, vol.16, pp.137-155.

G. MOUSTAKIDES, M. BASSEVILLE, A. BENVENISTE and G. LE VEY (1988). Diagnosing mechanical changes in vibrating systems. Research Report IRISA no 436/INRIA no 942.

J. NADLER and N.B. ROBBINS (1971). Some characteristics of Page's two-sided procedure for detecting a change in location parameter. *Annals Mathematical Statistics*, vol.42, no 2, pp.538-551.

P.M. NEWBOLD and Y.C. HO (1968). Detection of changes in the characteristics of a Gauss-Markov process. *IEEE Trans. Aerospace and Electronic Systems*, vol.AES-4, no 5, pp.707-718.

I.V. NIKIFOROV (1975). Sequential analysis applied to autoregression processes. *Automation and Remote Control*, vol.36, no 8, pp.174-178.

I.V. NIKIFOROV (1978). A statistical method for detecting the time at which the sensor properties change. *Proc. IMEKO Symp. Applications of Statistical Methods in Measurement*, Leningrad, pp.1-7.

I.V. NIKIFOROV (1979). Cumulative sums for detection of changes in random process characteristics. *Automation and Remote Control*, vol.40, no 2, Part 1, pp.48-58.

I.V. NIKIFOROV (1980). Modification and analysis of the cumulative sum procedure. *Automation and Remote Control*, vol.41, no 9, pp.74-80.

I.V. NIKIFOROV (1983). *Sequential Detection of Abrupt Changes in Time Series Properties*. Naouka, Moscow (in Russian).

I.V. NIKIFOROV (1986). Sequential detection of changes in stochastic systems. In *Detection of Abrupt Changes in Signals and Dynamical Systems* (M. Basseville, A. Benveniste, eds.). Lecture Notes in Control and Information Sciences, LNCIS 77, Springer, New York, pp.216-258.

I.V. NIKIFOROV (1991). Sequential detection of changes in stochastic processes. *Proc. IFAC/IFORS Symp. Identification and System Parameter Estimation*, Budapest, H., pp.11-19.

I.V. NIKIFOROV and I.N. TIKHONOV (1986). Application of change detection theory to seismic signal processing. In *Detection of Abrupt Changes in Signals and Dynamical Systems* (M. Basseville, A. Benveniste, eds.). Lecture Notes in Control and Information Sciences, LNCIS 77, Springer, New York, pp.355-373.

I.V. NIKIFOROV, I.N. TIKHONOV and T.G. MIKHAILOVA (1989). *Automatic On-line Processing of Seismic Station Data - Theory and Applications*. Far Eastern Dept of USSR Academy of Sciences, Vladivostok, USSR (in Russian).

I.V. NIKIFOROV, V. VARAVVA and V. KIREICHIKOV (1991). Application of statistical fault detection algorithms for navigation systems monitoring. *Proc. SAFEPROCESS'91*, Baden-Baden, FRG, pp.351-356.

I.V. NIKIFOROV, V. VARAVVA and V. KIREICHIKOV (1993). Application of statistical fault detection algorithms to navigation systems monitoring. To appear in *Automatica*.

A. NOVIKOV and B. ERGASHEV (1988). Analytical approach to the calculation of moving average characteristics. In *Statistical Problems of Control*, Issue 83, Vilnius, pp.110-114 (in Russian).

E.S. PAGE (1954a). Continuous inspection schemes. *Biometrika*, vol.41, pp.100-115.

E.S. PAGE (1954b). An improvement to Wald's approximation for some properties of sequential tests. *Jal Royal Statistical Society*, vol.B-16, no 1, pp.136-139.

E.S. PAGE (1954c). Control charts for the mean of a normal population. *Jal Royal Statistical Society*, vol.B-16, no 1, pp.131-135.

E.S. PAGE (1955). Control charts with warning lines. *Biometrika*, vol.42, no 2, pp.241-257.

E.S. PAGE (1957). Estimating the point of change in a continuous process. *Biometrika*, vol.44, pp.248-252.

E.S. PAGE (1962). A modified control chart with warning limits. *Biometrika*, vol.49, pp.171-176.

P.M. PARDALOS and J.B. ROSEN (1987). *Constrained Global Optimization: Algorithms and Applications*. Lecture Notes in Computer Sciences, vol.268. Springer, New York.

R.J. PATTON and J. CHEN (1991). A review of parity space approaches to fault diagnosis. *Proc. SAFEPROCESS'91*, Baden-Baden, FRG, pp.239-256.

R.J. PATTON and S.W. WILLCOX (1987). A robust method for fault diagnosis using parity space eigenstructure assignment. In *Fault Detection and Reliability - Knowledge-Based and Other Approaches* (M.G. Singh, K.S. Hindi, G. Schmidt, S. Tzafestas, eds.). International Series on Systems and Control, vol.9, Pergamon, Oxford, UK, pp.155-164.

R.J. PATTON, P. FRANK and R. CLARK, eds. (1989). *Fault Diagnosis in Dynamic Systems - Theory and Application*, International Series in Systems and Control Engineering, Prentice Hall International, London, UK.

L.F. PAU (1981). *Failure Diagnosis and Performance Monitoring*. Control and Systems Theory Series of Monographs and Textbooks, Dekker, New York.

L. PELKOWITZ (1987). The general discrete time disorder problem. *Stochastics*, vol.20, pp.89-110.

D. PERRIOT-MATHONNA (1984). Recursive stochastic estimation of parameters subject to random jumps. *IEEE Trans. Automatic Control*, vol.AC-29, pp.962-969.

M.J. PHILLIPS (1969). A survey of sampling procedures for continuous production. *Jal Royal Statistical Society*, vol.A-132, no 2, pp.205-228.

D. PICARD (1985). Testing and estimating change-points in time series. *Advances in Applied Probability*, vol.17, pp.841-867.

J.A. PIERCE (1989). Omega. *IEEE Trans. Aerospace and Electronic Systems*, vol.AES-4, no.7, pp.4-13.

M.S. PINSKER (1964). *Information and Information Stability of Random Variables and Processes*. Holden Day, San Francisco.

V.F. PISARENKO, A.F. KUSHNIR and I.V. SAVIN (1987). Statistical adaptive algorithms for estimation of onset moments of seismic phases. *Physics of the Earth and Planetary Interiors*, vol.47, pp.4-10.

D.J. POIRIER (1976). *The Econometrics of Structural Change*. North-Holland, Amsterdam.

M. POLLAK (1985). Optimal detection of a change in distribution. *Annals Statistics*, vol.13, pp.206-227.

M. POLLAK (1987). Average run lengths of an optimal method of detecting a change in distribution. *Annals Statistics*, vol.15, pp.749-779.

M. POLLAK and D. SIEGMUND (1975). Approximations to the expected sample size of certain sequential tests. *Annals Statistics*, vol.3, no 6, pp.1267-1282.

M. POLLAK and D. SIEGMUND (1985). A diffusion process and its application to detecting a change in the drift of a Brownian motion. *Biometrika*, vol.72, pp.267-280.

M. POLLAK and D. SIEGMUND (1991). Sequential decision of a change in a normal mean when the initial value is unknown. *Annals Statistics*, vol.19, no 1, pp.394-416.

B.T. POLYAK (1987). *Introduction to Optimization*. Optimization Software, Inc., New York.

H.V. POOR (1988). *An Introduction to Signal Detection and Estimation*. Springer Texts in Electrical Engineering, Springer, New York.

J.E. POTTER and M.C. SUMAN (1977). Thresholdless redundancy management with arrays of skewed instruments. *Integrity in Electronic Flight Control Systems*, vol.AGARDOGRAPH-224, pp.15.11-15.25.

R.R. PRAIRIE and W.J. ZIMMER (1970). Continuous sampling plans based on cumulative sums. *Applied Statistics*, vol.19, pp.222-230.

R.E. QUANDT (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *Jal American Statistical Association*, vol.53, pp.873-880.

R.E. QUANDT (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *Jal American Statistical Association*, vol.55, pp.324-330.

A. RAY (1989). Sequential testing for fault detection in multiply redundant measurements systems. *ASME Jal Dynamic Systems, Measurement and Control*, vol.111, pp.329-332.

A. RAY and M. DESAI (1984). A calibration and estimation filter for multiply redundant measurements systems. *ASME Jal Dynamic Systems, Measurement and Control*, vol.106, pp.149-156.

A. RAY and M. DESAI (1986). A redundancy management procedure for fault detection and isolation. *ASME Jal Dynamic Systems, Measurement, and Control*, vol.108, pp.248-254.

A. RAY and R. LUCK (1991). An introduction to sensor signal validation in redundant measurement systems. *IEEE Control Magazine*, vol.11, no 2, pp.44-49.

A. RAY, M. DESAI and J. DEYST (1983). Fault detection and isolation in a nuclear reactor. *Jal Energy*, vol.7, no 1, pp.79-85.

M.R. REYNOLDS (1975). Approximations to the average run length in cumulative sum control charts. *Technometrics*, vol.17, no 1, pp.65-71.

Y. RITOV (1990). Decision theoretic optimality of the cusum procedure. *Annals Statistics*. vol.18, no 3, pp.1464-1469.

H. ROBBINS (1970). Statistical methods related to the law of the iterated logarithm. *Annals Mathematical Statistics*, vol.41, no 5, pp.1397-1409.

H. ROBBINS and D. SIEGMUND (1970). Boundary crossing probabilities for the Wiener process and sample sums. *Annals Mathematical Statistics*, vol.41, no 5, pp.1410-1429.

L.G. ROBERTS (1965). Machine perception of 3-dimensional solids. In *Optical and Electro-optical Information Processing* (J.T. Tipett *et al*., eds.). MIT Press, Cambridge, MA, pp.159-197.

S.W. ROBERTS (1958). Properties of control chart zone tests. *B.S.T.J*., vol.37, pp.83-114.

S.W. ROBERTS (1959). Control charts based on geometric moving averages. *Technometrics*, vol.1, pp.239-250.

S.W. ROBERTS (1966). A comparison of some control chart procedures. *Technometrics*, vol.8, pp.411-430.

P.B. ROBINSON and T.Y. HO (1978). Average run lengths of geometric moving average charts by numerical methods. *Technometrics*, vol.20, pp.85-93.

A. ROUGÉE, M. BASSEVILLE, A. BENVENISTE and G. MOUSTAKIDES (1987). Optimum robust detection of changes in the AR part of a multivariable ARMA process. *IEEE Trans. Automatic Control*, vol.AC-32, no 12, pp.1116-1120.

G.G. ROUSSAS (1972). *Contiguity of Probability Measures, Some Applications in Statistics*. Cambridge University Press, New York.

A.C. SANDERSON and J. SEGEN (1980). Hierarchical modeling of EEG signals. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.PAMI-2, pp.405-414.

T. SASTRI (1987). Sequential method of change detection and adaptive prediction of municipal water demand. *Int. Jal System Sciences*, vol.18, pp.1029-1049.

A.L. SATIN and R.L. GATES (1978). Evaluation of parity equations for gyro failure detection and isolation. *Jal Guidance and Control*, vol.1, pp.14-20.

G.A.F. SEBER (1977). *Linear regression analysis*. Series in probability and mathematical statistics, Wiley, New York.

J. SEGEN and A.C. SANDERSON (1980). Detecting changes in a time series. *IEEE Trans. Information Theory*, vol.IT-26, no 2, pp.249-255.

S.A. SHABAN (1980). Change point problem and two-phase regression: an annotated bibliography. *Intern. Statistical Review*, vol.48, no 1, pp.83-93.

J.E. SHAPIRO (1979). *Mathematical Programming Structures and Algorithms*. Wiley, New York.

W.A. SHEWHART (1931). *Economic Control of Quality Manufactured Product*. D.Van Nostrand Reinhold, Princeton, NJ.

A.N. SHIRYAEV (1961). The problem of the most rapid detection of a disturbance in a stationary process. *Soviet Math. Dokl.*, no 2, pp.795-799.

A.N. SHIRYAEV (1963). On optimum methods in quickest detection problems. *Theory Probability and Applications*, vol.8, no 1, pp.22-46.

A.N. SHIRYAEV (1965). Some exact formulas in a disorder process. *Theory Probability and Applications*, vol.10, no 3, pp.348-354.

A.N. SHIRYAEV (1978). *Optimal Stopping Rules*. Springer, New York.

A.N. SHIRYAEV (1984). *Probability*. Springer, Graduate Texts in Mathematics, New York.

D. SIEGMUND (1975). Error probabilities and average sample number of the sequential probability ratio test. *Jal Royal Statistical Society*, vol.B-37, pp.394-401.

D. SIEGMUND (1979). Corrected diffusion approximations in certain random walk. *Advanced Applied Probability*, vol.11, pp.701-709.

D. SIEGMUND (1985a). Corrected diffusion approximations and their applications. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. Le Cam and R.A. Olshen, eds.), vol.II, pp.599-617.

D. SIEGMUND (1985b). *Sequential Analysis - Tests and Confidence Intervals*. Series in Statistics, Springer, New York.

M.G. SINGH, K.S. HINDI, G. SCHMIDT and S.G. TZAFESTAS, eds. (1987). *Fault Detection and Reliability - Knowledge-Based and Other Approaches*. International Series on Systems and Control, vol.9, Pergamon Press, Oxford, UK.

T. SÖDERSTRÖM and P. STOÏCA (1989). *System Identification*. International Series in Systems and Control Engineering, Prentice Hall International, London, UK.

M.A. STURZA (1988). Navigation system integrity monitoring using redundant measurements. *Navigation (Jal Institute of Navigation)*, vol.35, no 4, pp.483-501.

A.L. SWEET and J.C. HARDIN (1970). Solutions for some diffusion processes with two barriers. *Jal Applied Probability*, vol.7, pp.423-431.

S. TANAKA (1989). Diagnosability of systems and optimal sensor location. In *Fault diagnosis in Dynamic Systems - Theory and Application* (R. Patton, P. Frank, R. Clark, eds.). International Series in Systems and Control Engineering, Prentice Hall International, London, UK, pp.155-188.

S. TANAKA and P.C. MÜLLER (1990). Fault detection in linear discrete dynamic systems by a pattern recognition of a generalized likelihood ratio. *Jal Dynamic Systems, Measurement, and Control*, vol.112, pp.276-282.

S. TANAKA and P.C. MÜLLER (1992). Fault detection in linear discrete dynamic systems by a reduced order generalized likelihood ratio method. To appear in *Intern. Jal Systems Science*.

H.M. TAYLOR (1967). Statistical control of a gaussian process. *Technometrics*, vol.9, no 1, pp.29-41.

H.M. TAYLOR (1968). The economic design of cumulative sum control charts. *Technometrics*, vol.10, no 3, pp.479-488.

H.M. TAYLOR (1975). A stopped Brownian motion formula. *Annals Probability*, vol.3, no 2, 234-246.

L. TELKSNYS, ed. (1987). *Detection of Changes in Random Processes*. Optimization Software, Inc., Publications Division, New York.

I.N. TIKHONOV, T.G. MIKHAILOVA and I.V. NIKIFOROV (1990). Results of the experimental exploitation of an automatic seismic station. *News of the Far Eastern Dept of USSR Academy of Sciences*, no.6, pp.70-76 (in Russian).

D. TJOSTHEIM (1975). Autoregressive representation of seismic P-wave signals with an application to the problem of short period discriminant. *Geophysical Jal Royal Astronomical Society*, vol.43, no 2, pp.269-291.

J.K. TUGNAIT (1982). Adaptive estimation and identification for discrete systems with Markov jump parameters. *IEEE Trans. Automatic Control*, vol.AC-27, no 5, pp.1054-1065.

S.G. TZAFESTAS, M.G. SINGH and G. SCHMIDT, eds. (1987). *System Fault Diagnosis, Reliability and Related Knowledge-based Approaches* - vol.1. *Fault Diagnostics and Reliability* - vol.2. *Knowledge-based and Fault-tolerant Techniques*. Reidel, Dordrecht, Holland.

P.B. USORO, I.C. SCHICK and S. NEGAHDARIPOUR (1985). An innovation-based methodology for HVAC system fault detection. *Trans. ASME*, vol.107, pp.284-289.

L.C. VANCE (1983). A bibliography of statistical quality control charts techniques, 1970-1980. *Jal Quality Technology*, vol.15, pp.59-62.

D.S. VAN DOBBEN DE BRUYN (1968). *Cumulative Sum Tests: Theory and Practice*. Hafner, New York.

V. VARAVVA, V. KIREICHIKOV and I. NIKIFOROV (1988). Application of change detection algorithms to monitoring of measurement systems of moving object. In *Statistical Problems of Control*, Vilnius, Issue 83, pp.169-174 (in Russian).

A.V. VASILOPOULOS and A.P. STAMBOULIS (1978). Modification of control charts limits in the presence of data correlation. *Jal Quality Technology*, vol.10, no 1, pp.20-29.

M. VIDYASAGAR (1985). *Control System Synthesis: a Factorization Approach*. MIT Press, Cambridge, MA.

N. VISWANADHAM and K.D. MINTO (1988). Robust observer design with application to fault detection. *Proc. American Control Conf.*, Atlanta, Georgia, pp.1393-1399.

N. VISWANADHAM and R. SRICHANDER (1987). Fault detection using unknown-input observers. *Control-Theory and Advanced Technology*, vol.3, no 2, pp.91-101.

N. VISWANADHAM, J.H. TAYLOR and E.C. LUCE (1987a). A frequency domain approach to failure detection and isolation with application to GE-21 turbine engine control systems. *Control-Theory and Advanced Technology*, vol.3, no 1, pp.45-72.

N. VISWANADHAM, V.V.S. SARMA and M.G. SINGH (1987b). *Reliability of Computer and Control Systems*. Systems and Control Series, vol.8, North-Holland, Amsterdam.

S.E. VOROBEICHIKOV and V.V. KONEV (1984). Sequential method of detecting change in random processes of recurrence type. *Automation and Remote Control*, vol.45, no 5, Part I, pp.568-577.

S.E. VOROBEICHIKOV and V.V. KONEV (1988). Change detection of parameters of autoregressive processes with unknown noise distribution. In *Statistical Problems of Control*, Vilnius, Issue 83, pp.175-180 (in Russian).

E. WAHNON and N. BERMAN (1990). Tracking algorithm designed by the asymptotic local approach. *IEEE Trans. Automatic Control*, vol.AC-35 , no 4, pp.440-443.

E. WAHNON, M. BASSEVILLE and A. BENVENISTE (1991a). On failure detection and identification : an optimum robust min-max approach. *Proc. SAFEPROCESS'91*, Baden-Baden, FRG, pp.319-324.

E. WAHNON, A. BENVENISTE, L. EL GHAOUI and R. NIKOUKHAH (1991b). An optimum robust approach to statistical failure detection and identification. *Proc. 30th Conf. Decision and Control*, Brighton, UK, pp.650-655.

A. WALD (1947). *Sequential Analysis*. Wiley, New York.

K. WATANABE (1990). A multiple model adaptive filtering approach to fault diagnosis in stochastic systems. In *Fault Diagnosis in Dynamic Systems - Theory and Application* (R. Patton, P. Frank, R. Clark, eds.). International Series in Systems and Control Engineering, Prentice Hall International, London, UK, pp.411-438.

K. WATANABE and D.M. HIMMELBLAU (1982). Instrument fault detection in systems with uncertainties. *Int. Jal Systems Sciences*, vol.13, no 2, pp.137-158.

G.B. WETHERILL and D.W. BROWN (1991). *Statistical Process Control*. Chapman and Hall, London.

J.E. WHITE and J.L. SPEYER (1987). Detection filter design : spectral theory and algorithms. *IEEE Trans. Automatic Control*, vol.AC-32, no 7, pp.593-603.

A.S. WILLSKY (1976). A survey of design methods for failure detection in dynamic systems. *Automatica*, vol.12, pp.601-611.

A.S. WILLSKY (1986). Detection of abrupt changes in dynamic systems. In *Detection of Abrupt Changes in Signals and Dynamical Systems* (M. Basseville, A. Benveniste, eds.). Lecture Notes in Control and Information Sciences, LNCIS 77, Springer, New York, pp.27-49.

A.S. WILLSKY and H.L. JONES (1976). A generalized likelihood ratio approach to detection and estimation of jumps in linear systems. *IEEE Trans. Automatic Control*, vol.AC-21, no 1, pp.108-112.

A.S. WILLSKY, J.J. DEYST and B.S. CRAWFORD (1975). Two self-test methods applied to an inertial system problem. *Jal Spacecrafts and Rockets*, vol.12, pp.434-437.

A.S. WILLSKY, E.Y. CHOW, S.B. GERSHWIN, C.S. GREENE, P.K. HOUPT and A.L. KURKJIAN (1980). Dynamic model-based techniques for the detection of incidents on freeways. *IEEE Trans. Automatic Control*, vol.AC-25, no 3, pp.347-360.

W.H. WOODALL (1984). On the Markov chain approach to the two-sided cusum procedure. *Technometrics*, vol.25, pp.295-300.

W.H. WOODALL and M.M. NCUBE (1986). Multivariate cusum quality control procedures. *Technometrics*, vol.27, pp.285-292.

M. WOODROOFE (1982). *Nonlinear Renewal Theory in Sequential Analysis*. CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia.

R.H. WOODWARD and P.L. GOLDSMITH (1964). *Cumulative Sum Techniques*. Oliver and Boyd, Edinburgh, UK.

J. WÜNNENBERG (1990). *Observer-based Fault Detection in Dynamic Systems*. Thesis, Universitat-Gesamthochschule-Duisburg, FRG.

E. YASHCHIN (1985a). On a unified approach to the analysis of two-sided cumulative sum control schemes with headstarts. *Advanced Applied Probability*, vol.17, pp.562-593.

E. YASHCHIN (1985b). On the analysis and design of cusum-Shewhart control schemes. *IBM Jal Research and Development*, vol.29, no 4, pp.377-391.

S. ZACKS (1983). Survey of classical and Bayesian approaches to the change-point problem: fixed sample and sequential procedures of testing and estimation. In *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on his Sixtieth Birthday* (M.H. Rizvi, J. Rustagi, D. Siegmund, eds.). Academic Press, New York, pp.245-269.

Q. ZHANG (1991). *Contribution à la Surveillance de Procédés Industriels*. Thèse de l'Université de Rennes I (in French).

Q. ZHANG, M. BASSEVILLE and A. BENVENISTE (1994). Early warning of slight changes in systems and plants with application to condition based maintenance. *Automatica*, Special Issue on Statistical Methods in Signal Processing and Control, vol.30, no 1, pp.95-114.

A.A. ZHIGLJAVSKY and A.E. KRASKOVSKY (1988). *Detection of Changes in Stochastic Processes in Radiotechnical Problems*, Leningrad State University Publications, Leningrad (in Russian).

Index