Jean-Philippe Vert  
Jean-Philippe.Vert@mines.org  
Mines ParisTech / Curie Institute / Inserm  
1. Introduction

2. Cancer prognosis from DNA copy number variations

3. Diagnosis and prognosis from gene expression data

4. Conclusion
1. Introduction

2. Cancer prognosis from DNA copy number variations

3. Diagnosis and prognosis from gene expression data

4. Conclusion
Comparative Genomic Hybridization (CGH)

Motivation

- Comparative genomic hybridization (CGH) data measure the DNA copy number along the genome.
- Very useful, in particular in cancer research to observe systematically variants in DNA content.
Cancer prognosis: can we predict the future evolution?

Aggressive (left) vs non-aggressive (right) melanoma
CGH shows the (static) DNA

Cancer cells have also abnormal (dynamic) gene expression (= transcription)
Tissue profiling with DNA chips

Data

- Gene expression measures for more than 10k genes
- Measured typically on less than 100 samples of two (or more) different classes (e.g., different tumors)
Can we identify the cancer subtype? (diagnosis)
Can we predict the future evolution? (prognosis)
Given a **training set** of labeled data with...

learn a discrimination rule...

... in order to **predict** the label of new data
Machine learning (pattern recognition / supervised classification)

1. Given a **training set** of labeled data with...
2. **learn** a discrimination rule...
3. ... **in order to predict** the label of new data
1. Given a training set of labeled data with...
2. learn a discrimination rule...
3. ... in order to predict the label of new data
Machine learning (pattern recognition / supervised classification)

Given a **training set** of labeled data with...

learn a discrimination rule...

... in order to **predict** the label of new data
Machine learning in bioinformatics

Genome annotation, systems biology, personalized medicine...

Challenges
- Few samples
- High dimension
- Structured data
- Heterogeneous data
- Prior knowledge
- Fast and scalable implementations
- Interpretable models
Many applications
Multimedia, image, video, speech recognition, web, social network, online advertising, finance, biology, chemistry

Many tools
Linear discriminant analysis, logistic regression, decision trees, neural networks, support vector machines...
1. Define a large family of "candidate classifiers", e.g., linear predictors:

\[ f_\beta(x) = \beta^T x \quad \text{for } x \in \mathbb{R}^p \]

2. For any candidate classifier \( f_\beta \), quantify how "good" it is on the training set with some empirical risk, e.g.:

\[ R(\beta) = \frac{1}{n} \sum_{i=1}^{n} l(f_\beta(x_i), y_i) . \]

3. Choose \( \beta \) that achieves the minimum empirical risk, subject to some constraint:

\[ \min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C . \]
Define a large family of "candidate classifiers", e.g., linear predictors:

\[ f_\beta(x) = \beta^T x \quad \text{for} \ x \in \mathbb{R}^p \]

For any candidate classifier \( f_\beta \), quantify how "good" it is on the training set with some empirical risk, e.g.:

\[
R(\beta) = \frac{1}{n} \sum_{i=1}^{n} l(f_\beta(x_i), y_i).
\]

Choose \( \beta \) that achieves the minimum empirical risk, subject to some constraint:

\[
\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C.
\]
Define a large family of "candidate classifiers", e.g., linear predictors:

\[ f_\beta(x) = \beta^T x \quad \text{for } x \in \mathbb{R}^p \]

For any candidate classifier \( f_\beta \), quantify how "good" it is on the training set with some empirical risk, e.g.:

\[ R(\beta) = \frac{1}{n} \sum_{i=1}^{n} l(f_\beta(x_i), y_i) . \]

Choose \( \beta \) that achieves the minimum empirical risk, subject to some constraint:

\[ \min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C . \]
Why shrinkage classifiers?

$$\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C.$$
Why shrinkage classifiers?

$$\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C.$$
Why shrinkage classifiers?

$$\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C.$$
Why skrinkage classifiers?

\[
\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C.
\]
Why shrinkage classifiers?

\[
\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C.
\]
Why shrinkage classifiers?

$$\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C.$$
"Increases bias and decreases variance"

Common choices are

- $\Omega(\beta) = \sum_{i=1}^{p} \beta_i^2$ (ridge regression, SVM, ...)
- $\Omega(\beta) = \sum_{i=1}^{p} |\beta_i|$ (lasso, boosting, ...)
Including prior knowledge in the penalty?

\[
\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C.
\]
Including prior knowledge in the penalty?

$$\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C.$$
Including prior knowledge in the penalty?

\[
\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C.
\]
Including prior knowledge in the penalty?

$$\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C .$$

Diagram with points $b^{\text{est}}$ and $b^{*}$.
Including prior knowledge in the penalty?

$$\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C.$$
Including prior knowledge in the penalty?

\[
\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C.
\]
Further benefit: sparsity-inducing penalties

(Lasso)

$$\min_{\beta} R(\beta) + \lambda \sum_{i=1}^{p} |\beta_i|$$

Geometric interpretation with $p = 2$
1. Introduction

2. Cancer prognosis from DNA copy number variations

3. Diagnosis and prognosis from gene expression data

4. Conclusion
Chromosomal aberrations in cancer
Cancer prognosis: can we predict the future evolution?

Aggressive (left) vs non-aggressive (right) melanoma
For a CGH profile \( x \in \mathbb{R}^p \), we focus on linear classifiers, i.e., the sign of:

\[
f_\beta(x) = \beta^\top x.
\]

We expect \( \beta \) to be

- **sparse**: not all positions should be discriminative
- **piecewise constant**: within a selected region, all probes should contribute equally
The $\ell_1$ penalty (Tibshirani, 1996; Chen et al., 1998)

The solution of

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda \sum_{i=1}^{p} |\beta_i|$$

is usually sparse.

Geometric interpretation with $p = 2$
Promoting piecewise constant profiles penalty

The variable fusion penalty (Land and Friedman, 1996)

The solution of

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i|$$

is usually piecewise constant.

Geometric interpretation with $p = 2$
Fused Lasso signal approximator (Tibshirani et al., 2005)

\[
\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{p} (y_i - \beta_i)^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i|.
\]

- First term leads to sparse solutions
- Second term leads to piecewise constant solutions
Fused lasso for supervised classification (Rapaport et al., 2008)

\[
\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \ell \left( y_i, \beta^T x_i \right) + \lambda_1 \sum_{i=1}^p |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i| .
\]

where \( \ell \) is, e.g., the hinge loss \( \ell(y, t) = \max(1 - yt, 0) \).

**Implementation**

- When \( \ell \) is the hinge loss (fused SVM), this is a linear program -> up to \( p = 10^3 \sim 10^4 \)
- When \( \ell \) is convex and smooth (logistic, quadratic), efficient implementation with proximal methods -> up to \( p = 10^8 \sim 10^9 \)
Fused lasso for supervised classification (Rapaport et al., 2008)

\[
\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \ell \left( y_i, \beta^\top x_i \right) + \lambda_1 \sum_{i=1}^p |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i|.
\]

where \( \ell \) is, e.g., the hinge loss \( \ell(y, t) = \max(1 - yt, 0) \).

**Implementation**

- When \( \ell \) is the hinge loss (fused SVM), this is a linear program -> up to \( p = 10^3 \sim 10^4 \)
- When \( \ell \) is convex and smooth (logistic, quadratic), efficient implementation with proximal methods -> up to \( p = 10^8 \sim 10^9 \)
Example: predicting metastasis in melanoma
Diagnosis

Jean-Philippe Vert (ParisTech)

Machine learning in genomics

Mines M&S seminar

C-myb (U22376)
Proteasome iota (X59417)
MB-1 (U56259)
Cyclin D3 (M99287)
Myosin light chain (M31211)
RhoGAP48 (X74262)
SNF2 (D26156)
HIF-1 (S90223)
E2A (M31523)
Inducible protein (L47738)
Dynamin light chain (U32944)
Topoisomerase II B (Z15115)
IRF2 (X15949)
TFIIIE (X63469)
Acyl-Coenzyme A dehydrogenase (M91432)
SNF2 (U29175)
Ca2+-ATPase (Z68781)
SRP5 (U20990)
MCM3 (D38073)
Deoxyhypusine synthase (U26266)
Op 15 (M31503)
Rabaptin-5 (Y08612)
Heterochromatin protein p25 (U35851)
IL-7 receptor (M59606)
Adenosine deaminase (M13792)

Fumarylacetoacetate (M55150)
Zyxin (X99735)
LTG synthase (U20136)
LYN (M16038)
HoxA9 (U382759)
CD33 (M42197)
Adipin (M584526)
Leptin receptor (U12670)
Cystatin C (M427801)
Protocadherin 1 (X17042)
IL-8 precursor (Y00787)
Azurocidin (M96326)
ps2 (U46751)
CyP3 (M405254)
MCL1 (L00266)
ATPase (M62762)
IL-8 (M28130)
Cathepsin D (M61389)
Leptin (M57710)
MAD-3 (M59045)
CD11c (M8695)
Ebp72 (X85116)
Lysozyme (M19045)
Propertin (M3652)
Catalase (X04085)
Prognosis

Gene-Expression Profiling

A

St. Gallen Criteria

B

NO AT RISK
Good signature 50 57 54 45 31 22 12
Poor signature 91 72 55 41 26 17 9

NO AT RISK
Low risk 22 22 21 17 9 5 2
High risk 120 107 86 60 48 34 10

P=0.05

Probability of Remaining Metastasis-free

P<0.001

Years

Years

0 2 4 6 8 10 12

0 2 4 6 8 10 12
Gene selection, molecular signature

The idea

- We look for a **limited set** of genes that are sufficient for prediction.
- Selected genes should inform us about the underlying biology.
But... unstability of selected features

- Wang dataset: \( n = 286, \ p = 8141 \)
- Pearson correlation with the output on 2 random subsamples of 143 samples:
Comparison of feature selection methods...

Haury et al. (2011)
Gene networks

- Glycolysis / Gluconeogenesis
- Porphyrin and chlorophyll metabolism
- Riboflavin metabolism
- Folate biosynthesis
- Biosynthesis of steroids, ergosterol metabolism
- Lysine biosynthesis
- Phenylalanine, tyrosine and tryptophan biosynthesis
- Purine metabolism
- Oxidative phosphorylation, TCA cycle
- DNA and RNA polymerase subunits
- Nitrogen, asparagine metabolism
- N-Glycan biosynthesis
- Protein kinases
Gene networks and expression data

Motivation

- Basic biological functions usually involve the coordinated action of several proteins:
  - Formation of protein complexes
  - Activation of metabolic, signalling or regulatory pathways
- We know these groups through functional groups and protein networks

Shrinkage estimators with prior knowledge

\[
\min_{\beta} R(\beta) + \lambda \Omega(\beta)
\]

How to design penalties \( \Omega(\beta) \) to encode the following hypotheses:

1. Connected genes on a network should have similar weights
2. Select few genes that are connected or belong to same predefined functional groups
Gene networks and expression data

Motivation

- Basic biological functions usually involve the coordinated action of several proteins:
  - Formation of protein complexes
  - Activation of metabolic, signalling or regulatory pathways
- We know these groups through functional groups and protein networks

Shrinkage estimators with prior knowledge

\[ \min_{\beta} R(\beta) + \lambda \Omega(\beta) \]

How to design penalties \( \Omega(\beta) \) to encode the following hypotheses:
1. Connected genes on a network should have similar weights
2. Select few genes that are connected or belong to same predefined functional groups
Hypothesis 1: connected genes on a network should have similar weights

- Smooth weights on the graph (or more generally graph kernels)

\[ \Omega(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 \]

- Gene selection + smooth on the graph

\[ \Omega(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 + \sum_{i=1}^{p} |\beta_i| \]

- Gene selection + Piecewise constant on the graph (total variation)

\[ \Omega(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_{i=1}^{p} |\beta_i| \]
Glycolysis / Gluconeogenesis

Porphyrin and chlorophyll metabolism

Riboflavin metabolism

Folate biosynthesis

Biosynthesis of steroids, ergosterol metabolism

Lysine biosynthesis

Phenylalanine, tyrosine and tryptophan biosynthesis

Purine metabolism

Protein kinases

Nitrogen, asparagine metabolism

DNA and RNA polymerase subunits

Oxidative phosphorylation, TCA cycle

N Glycan biosynthesis
We are happy to see pathways appear.

However, in some cases, connected genes should have "opposite" weights (inhibition, pathway branching, etc...)

How to capture pathways without constraints on the weight similarities?
Selecting pre-defined groups of variables

Group lasso (Yuan & Lin, 2006)

If groups of covariates are likely to be selected together, the $\ell_1/\ell_2$-norm induces sparse solutions at the group level:

$$\Omega_{\text{group}}(\beta) = \sum_g \| \beta_g \|_2$$

$$\Omega(\beta_1, \beta_2, \beta_3) = \| (\beta_1, \beta_2) \|_2 + \| \beta_3 \|_2$$

$$= \sqrt{\beta_1^2 + \beta_2^2 + | \beta_3 |}$$
When groups overlap, the group Lasso

$$\Omega_{\text{group}}(\beta) = \sum_{g} \| \beta_g \|$$

puts groups to 0 $\Rightarrow$ the support of the solution is the complement of a union of groups

Alternatively, the following latent group Lasso promotes instead solutions with supports as union of predefined overlapping groups (Jacob et al., 2009):

$$\Omega_{\text{latent}}(\beta) = \sup_{\alpha \in \mathbb{R}^p : \forall g, \| \alpha_g \| \leq 1} \alpha^T \beta$$
Group Lasso vs latent group Lasso

Balls for $\Omega^G_{\text{group}} (\cdot)$ (middle) and $\Omega^{\text{latent}}_\cdot$ (right) for the groups $\mathcal{G} = \{\{1, 2\}, \{2, 3\}\}$ where $w_2$ is represented as the vertical coordinate.
Graph lasso vs kernel on graph

- **Graph lasso:**
  \[
  \Omega_{\text{group}}(\beta) = \sum_{i \sim j} \sqrt{\beta_i^2 + \beta_j^2} \quad \text{or} \quad \Omega_{\text{latent}}(\beta) = \sup_{\alpha \in \mathbb{R}^p : \forall i \sim j} \alpha^\top \beta \quad \text{s.t.} \quad \sqrt{\alpha_i^2 + \alpha_j^2} \leq 1
  \]
  constrains the **sparsity**, not the values

- **Graph kernel**
  \[
  \Omega_{\text{graph kernel}}(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2.
  \]
  constrains the values (**smoothness**), not the sparsity
## Preliminary results

### Breast cancer data

- Gene expression data for 8,141 genes in 295 breast cancer tumors.
- Canonical pathways from MSigDB containing 639 groups of genes, 637 of which involve genes from our study.

<table>
<thead>
<tr>
<th>METHOD</th>
<th>$\ell_1$</th>
<th>$\Omega_{\text{OVERLAP}}(\cdot)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERROR</td>
<td>0.38 ± 0.04</td>
<td>0.36 ± 0.03</td>
</tr>
<tr>
<td>MEAN # PATH.</td>
<td>130</td>
<td>30</td>
</tr>
</tbody>
</table>

Graph on the genes.

<table>
<thead>
<tr>
<th>METHOD</th>
<th>$\ell_1$</th>
<th>$\Omega_{\text{graph}}(\cdot)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERROR</td>
<td>0.39 ± 0.04</td>
<td>0.36 ± 0.01</td>
</tr>
<tr>
<td>AV. SIZE C.C.</td>
<td>1.03</td>
<td>1.30</td>
</tr>
</tbody>
</table>
Classical lasso signature

Graph showing gene interactions:
- EIF4G1
- AREG
- MMP9
- MMP7
- UBE2A
- RNF40
- POLD1
- POLD4
- RPS13
- RPS25
- RPS15
- RPL6
- EEF1A1
- PCSK6
- BTG2
- YWHAZ
- ADRA2B
- ADRB1
- NEDD9
- C2orf11
- TAT
- PDE6B
- TGFB2
- MYCBP
- GRP
- DLEU2
- ALDH3A2
- VEGFB
- PSMD7
- CXCL13
- FLT3
- PPAT
- ULK1
- SLC16A3
- AKR1C4
- BATF
- PLP2
- SYTL2
- CCNB2
- SLC39A7
- HYPK
- PDHB
- UBD
- FBXO2
- E2F1
- LRP5
- PIK3CG
- ZCCHC8
- NLRP2
- ANKZF1
- PRC1
- CTSL2
- TK1
- PTPN3
- CASC3
- IGFBP5
- RTN3
- DNAJB2
- CDH19
- GLRX2
Graph Lasso signature

Jean-Philippe Vert (ParisTech)

Machine learning in genomics

Mines M&S seminar 45 / 48
1. Introduction

2. Cancer prognosis from DNA copy number variations

3. Diagnosis and prognosis from gene expression data

4. Conclusion
Machine learning offers **many powerful tools** to learn predictive models from large sets of complex data.

**Specific developments** are required to solve complex problems that arise in bio-informatics.

Integration of prior knowledge in the penalization / regularization function is an efficient approach to fight the curse of dimension.

Requires **interdisciplinary collaborations** to incorporate expert knowledge at the heart of learning algorithms.

Many other applications not covered in this presentation!
Acknowledgements!

Franck Rapaport (MSKCC), Emmanuel Barillot, Andrei Zynoviev, Kevin Bleakley (INRIA), Fantine Mordelet (Duke), Anne-Claire Haury, Laurent Jacob (UC Berkeley) Guillaume Obozinski (INRIA)