Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2009)
71, *Part* 3, *pp.*

Covariance-regularized regression and classification for high dimensional problems

Daniela M. Witten and Robert Tibshirani

Stanford University, USA

[Received March 2008. Revised September 2008]

Summary. We propose covariance-regularized regression, a family of methods for prediction in high dimensional settings that uses a shrunken estimate of the inverse covariance matrix of the features to achieve superior prediction. An estimate of the inverse covariance matrix is obtained by maximizing the log-likelihood of the data, under a multivariate normal model, subject to a penalty; it is then used to estimate coefficients for the regression of the response onto the features. We show that ridge regression, the lasso and the elastic net are special cases of covariance-regularized regression, and we demonstrate that certain previously unexplored forms of covariance-regularized regression can outperform existing methods in a range of situations. The covariance-regularized regression framework is extended to generalized linear models and linear discriminant analysis, and is used to analyse gene expression data sets with multiple class and survival outcomes.

Keywords: Classification; Covariance regularization; $n \ll p$; Regression; Variable selection

1. Introduction

In high dimensional regression problems, where p, the number of features, is nearly as large as, or larger than, n, the number of observations, ordinary least squares regression does not provide a satisfactory solution. A remedy for the shortcomings of least squares is to modify the sum of squared errors criterion that is used to estimate the regression coefficients, by using penalties that are based on the magnitudes of the coefficients:

$$\hat{\beta} = \arg\min_{\beta}(\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_1\|\beta\|^{p_1} + \lambda_2\|\beta\|^{p_2}). \tag{1}$$

(Here, the notation $\|\beta\|^s$ is used to indicate $\Sigma_{i=1}^{p}|\beta_i|^s$.) Many popular regularization methods fall into this framework. For instance, when $\lambda_2 = 0$, $p_1 = 0$ gives best subset selection, $p_1 = 2$ gives ridge regression (Hoerl and Kennard, 1970) and $p_1 = 1$ gives the lasso (Tibshirani, 1996). More generally, for $\lambda_2 = 0$ and $p_1 \geqslant 0$, equation (1) defines the bridge estimators (Frank and Friedman, 1993). Equation (1) defines the naive elastic net in the case that $p_1 = 1$ and $p_2 = 2$ (Zou and Hastie, 2005). In this paper, we present a new approach to regularizing linear regression that involves applying a penalty, not to the sum of squared errors, but rather to the log-likelihood of the data under a multivariate normal model.

The least squares solution is $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. In multivariate normal theory, the entries of $(\mathbf{X}^T\mathbf{X})^{-1}$ that equal 0 correspond to pairs of variables that have no sample partial correlation; in other words, pairs of variables that are conditionally independent, given all of the other features in the data. Non-zero entries of $(\mathbf{X}^T\mathbf{X})^{-1}$ correspond to non-zero partial correlations. One way

Address for correspondence: Daniela M. Witten, Department of Statistics, Stanford University, 390 Serra Mall, Stanford, CA 94305, USA.
E-mail: dwitten@stanford.edu

 1369–7412/09/71000

to perform regularization of least squares regression is to shrink the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$; in fact, this is done by ridge regression, since the ridge solution can be written as $\hat{\beta}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$. Here, we propose a more general approach to shrinkage of the inverse covariance matrix. Our method involves estimating a regularized inverse covariance matrix by maximizing the log-likelihood of the data under a multivariate normal model, subject to a constraint on the elements of the inverse covariance matrix. In doing this, we attempt to distinguish between variables that truly are partially correlated with each other and variables that in fact have zero partial correlation. We then use this regularized inverse covariance matrix to obtain regularized regression coefficients. We call the class of regression methods that are defined by this procedure the *scout*.

In Section 2, we present the scout criteria and explain the method in greater detail. We also discuss connections between the scout and pre-existing regression methods. In particular, we show that ridge regression, the lasso and the elastic net are special cases of the scout. In addition, we present some specific members of the scout class that perform well relatively to pre-existing methods in a variety of situations. In Sections 3, 4 and 5, we demonstrate the use of these methods in regression, classification and generalized linear model settings on simulated data and on some gene expression data sets.

2. The scout method

2.1. The general scout family

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ denote an $n \times p$ matrix of data, where n is the number of observations and p the number of features. Let $\mathbf{y}$ denote a vector of length n, containing a response value for each observation. Assume that the columns of $\mathbf{X}$ are standardized, and that $\mathbf{y}$ is centred. We can create a matrix $\tilde{\mathbf{X}} = (\mathbf{X} \quad \mathbf{y})$, which has dimension $n \times (p+1)$. If we assume that $\tilde{\mathbf{X}} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma})$, then we can find the maximum likelihood estimator of the population inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ by maximizing

$$\log\{\det(\boldsymbol{\Sigma}^{-1})\} - \operatorname{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) \tag{2}$$

where

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{\mathbf{xx}} & \mathbf{S}_{\mathbf{xy}} \\ \mathbf{S}_{\mathbf{xy}}^T & S_{yy} \end{pmatrix}$$

is the empirical covariance matrix of $\tilde{\mathbf{X}}$. Assume for a moment that $\mathbf{S}$ is invertible. Then, the maximum likelihood estimator for $\boldsymbol{\Sigma}^{-1}$ is $\mathbf{S}^{-1}$ (we use the fact that $\mathrm{d}\log\{\det(\mathbf{W})\}/\mathrm{d}\mathbf{W} = \mathbf{W}^{-1}$ for a symmetric positive definite matrix $\mathbf{W}$). Let

$$\boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{\Theta}_{\mathbf{xx}} & \boldsymbol{\Theta}_{\mathbf{xy}} \\ \boldsymbol{\Theta}_{\mathbf{xy}}^T & \Theta_{yy} \end{pmatrix}$$

denote a symmetric estimate of $\boldsymbol{\Sigma}^{-1}$. The problem of regressing $\mathbf{y}$ onto $\mathbf{X}$ is closely related to the problem of estimating $\boldsymbol{\Sigma}^{-1}$, since the least squares coefficients for the regression equal $-\boldsymbol{\Theta}_{\mathbf{xy}}/\Theta_{yy}$ for $\boldsymbol{\Theta} = \mathbf{S}^{-1}$ (this follows from the partitioned inverse formula; see for example Mardia *et al.* (1979), page 459). If $p > n$, then some type of regularization is needed to estimate the regression coefficients, since $\mathbf{S}$ is not invertible. Even if $p < n$, we may want to shrink the least squares coefficients in some way to achieve superior prediction. The connection between estimation of $\boldsymbol{\Theta}$ and estimation of the least squares coefficients suggests the possibility that rather than shrinking the coefficients β by applying a penalty to the sum of squared errors for the regression of $\mathbf{y}$ onto $\mathbf{X}$, as is done in for example ridge regression or the lasso, we can obtain shrunken β-estimates through maximization of the penalized log-likelihood of the data.

To do this, we could estimate $\mathbf{\Sigma}^{-1}$ as $\mathbf{\Theta}$ that maximizes

$$\log\{\det(\mathbf{\Theta})\} - \mathrm{tr}(\mathbf{S\Theta}) - J(\mathbf{\Theta}) \tag{3}$$

where $J(\mathbf{\Theta})$ is a penalty function. For example, $J(\mathbf{\Theta}) = \|\mathbf{\Theta}\|^p$ denotes the sum of absolute values of the elements of $\mathbf{\Theta}$ if $p=1$, and it denotes the sum of squared elements of $\mathbf{\Theta}$ if $p=2$. Our regression coefficients would then be given by the formula $\beta = -\mathbf{\Theta_{xy}}/\Theta_{yy}$. However, recall that the ij-element of $\mathbf{\Theta}$ is 0 if and only if the partial correlation of $\tilde{\mathbf{x}}_i$ with $\tilde{\mathbf{x}}_j$ (conditional on all the other variables in $\tilde{\mathbf{X}}$) is 0. (This follows from the definition of the partial correlation, and again from the partitioned inverse formula.) Note that $\mathbf{y}$ is included in $\tilde{\mathbf{X}}$. So it does not make sense to regularize the elements of $\mathbf{\Theta}$ as presented above, because we really care about the partial correlations of pairs of variables given the other variables, as opposed to the partial correlations of pairs of variables given the other variables and the response.

For these reasons, rather than obtaining an estimate of $\mathbf{\Sigma}^{-1}$ by maximizing the penalized log-likelihood in equation (3), we estimate it via a two-stage maximization, given in the following algorithm for the *scout procedure for general penalty functions*.

Step 1: compute $\hat{\mathbf{\Theta}}_{\mathbf{xx}}$, which maximizes

$$\log\{\det(\mathbf{\Theta_{xx}})\} - \mathrm{tr}(\mathbf{S_{xx}\Theta_{xx}}) - J_1(\mathbf{\Theta_{xx}}). \tag{4}$$

Step 2: compute $\hat{\mathbf{\Theta}}$, which maximizes

$$\log\{\det(\mathbf{\Theta})\} - \mathrm{tr}(\mathbf{S\Theta}) - J_2(\mathbf{\Theta}), \tag{5}$$

where the top left $p \times p$ submatrix of $\hat{\mathbf{\Theta}}$ is constrained to equal $\hat{\mathbf{\Theta}}_{\mathbf{xx}}$, the solution to step 1.
Step 3: compute $\hat{\beta}$, defined by $\hat{\beta} = -\hat{\mathbf{\Theta}}_{\mathbf{xy}}/\hat{\Theta}_{yy}$.
Step 4: compute $\hat{\beta}^* = c\hat{\beta}$, where c is the coefficient for the regression of $\mathbf{y}$ onto $\mathbf{X}\hat{\beta}$.

$\hat{\beta}^*$ denotes the regularized coefficients that are obtained by using this new method. Step 1 of the scout procedure involves obtaining shrunken estimates of $(\mathbf{\Sigma_{xx}})^{-1}$ to smooth our estimates of which variables are conditionally independent. Step 2 involves obtaining shrunken estimates of $\mathbf{\Sigma}^{-1}$, conditional on $(\mathbf{\Sigma}^{-1})_{\mathbf{xx}} = \hat{\mathbf{\Theta}}_{\mathbf{xx}}$, the estimate that is obtained in step 1. Thus, we obtain regularized estimates of which predictors are dependent on $\mathbf{y}$, given all the other predictors. The scaling in the last step is performed because it has been found, empirically, to improve performance.

By penalizing the entries of the inverse covariance matrix of the predictors in step 1 of the scout procedure, we are attempting to distinguish between pairs of variables that truly are conditionally dependent, and pairs of variables that appear to be conditionally dependent only because of chance. We are searching, or *scouting*, for variables that truly are correlated with each other, conditional on all the other variables. Our hope is that sets of variables that truly are conditionally dependent will also be related to the response. In the context of a microarray experiment, where the variables are genes and the response is some clinical outcome, this assumption is reasonable: we seek genes that are part of a pathway related to the response. We expect that such genes will also be conditionally dependent. In step 2, we shrink our estimates of the partial correlation between each predictor and the response, given the shrunken partial correlations between the predictors that we estimated in step 1. In contrast with ordinary least squares regression, which uses the inverse of the empirical covariance matrix to compute regression coefficients, we jointly model the relationship that the p predictors have with each other and with the response to obtain shrunken regression coefficients.

We define the *scout family* of estimated coefficients for the regression of $\mathbf{y}$ onto $\mathbf{X}$ as the solutions $\hat{\beta}^*$ that are obtained in step 4 of the scout procedure. We refer to the penalized log-likelihoods in steps 1 and 2 of the scout procedure as the first and second *scout criteria*.

Table 1. Special cases of the scout

$J_1(\boldsymbol{\Theta}_{\mathbf{xx}})$	$J_2(\boldsymbol{\Theta})$	*Method*
0	0	Least squares
$\mathrm{tr}(\boldsymbol{\Theta}_{\mathbf{xx}})$	0	Ridge regression
$\mathrm{tr}(\boldsymbol{\Theta}_{\mathbf{xx}})$	$\|\boldsymbol{\Theta}\|^1$	Elastic net
0	$\|\boldsymbol{\Theta}\|^1$	Lasso
0	$\|\boldsymbol{\Theta}\|^2$	Ridge regression

In the rest of the paper, when we discuss properties of the scout, for ease of notation we shall ignore the scale factor in step 4 of the scout procedure. For instance, if we claim that two procedures yield the same regression coefficients, we more specifically mean that the regression coefficients are the same up to scaling by a constant factor.

Least squares, the elastic net, the lasso and ridge regression result from the scout procedure with appropriate choices of J_1 and J_2 (up to a scaling by a constant). Details are in Table 1. The first two results can be shown directly by differentiating the scout criteria, and the others follow from equation (11) in Section 2.4.

2.2. L_p-penalties

Throughout the remainder of this paper, with the exception of Section 3.2, we shall exclusively be interested in the case that $J_1(\boldsymbol{\Theta}) = \lambda_1 \|\boldsymbol{\Theta}\|^{p_1}$ and $J_2(\boldsymbol{\Theta}) = (\lambda_2/2)\|\boldsymbol{\Theta}\|^{p_2}$, where the norm is taken elementwise over the entries of $\boldsymbol{\Theta}$, and where $\lambda_1, \lambda_2 \geqslant 0$. For ease of notation, Scout(p_1, p_2) will refer to the solution to the scout criterion with J_1 and J_2 as just mentioned. If $\lambda_2 = 0$, then this will be indicated by Scout$(p_1, \cdot)$ and, if $\lambda_1 = 0$, then this will be indicated by Scout$(\cdot, p_2)$. Therefore, in the rest of this paper, the *scout procedure with L_p-penalties* will be as follows.

Step 1: compute $\hat{\boldsymbol{\Theta}}_{\mathbf{xx}}$, which maximizes

$$\log\{\det(\boldsymbol{\Theta}_{\mathbf{xx}})\} - \mathrm{tr}(\mathbf{S}_{\mathbf{xx}}\boldsymbol{\Theta}_{\mathbf{xx}}) - \lambda_1 \|\boldsymbol{\Theta}_{\mathbf{xx}}\|^{p_1}. \tag{6}$$

Step 2: compute $\hat{\boldsymbol{\Theta}}$, which maximizes

$$\log\{\det(\boldsymbol{\Theta})\} - \mathrm{tr}(\mathbf{S}\boldsymbol{\Theta}) - \frac{\lambda_2}{2}\|\boldsymbol{\Theta}\|^{p_2}, \tag{7}$$

where the top left $p \times p$ submatrix of $\hat{\boldsymbol{\Theta}}$ is constrained to equal $\hat{\boldsymbol{\Theta}}_{\mathbf{xx}}$, the solution to step 1. Because of this constraint, the penalty really is only being applied to the last row and column of $\hat{\boldsymbol{\Theta}}$.

Step 3: compute $\hat{\beta}$, defined by $\hat{\beta} = -\hat{\boldsymbol{\Theta}}_{\mathbf{xy}}/\hat{\Theta}_{yy}$.

Step 4: compute $\hat{\beta}^* = c\hat{\beta}$, where c is the coefficient for the regression of $\mathbf{y}$ onto $\mathbf{X}\hat{\beta}$.

2.3. Simple example

Here, we present a toy example in which $n = 20$ observations on $p = 19$ variables are generated under the model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, where $\beta_j = j$ for $j \leqslant 10$ and $\beta_j = 0$ for $j > 10$, and where $\varepsilon_i \sim N(0, 25)$ independent. In addition, the first 10 variables have correlation 0.5 with each other; the rest are uncorrelated.

Fig. 1 shows the average over 500 simulations of the least squares regression coefficients and the Scout$(1, \cdot)$ regression estimate. It is not surprising that the least squares method performs poorly in this situation, since n is barely larger than p. Scout$(1, \cdot)$ performs quite well; though

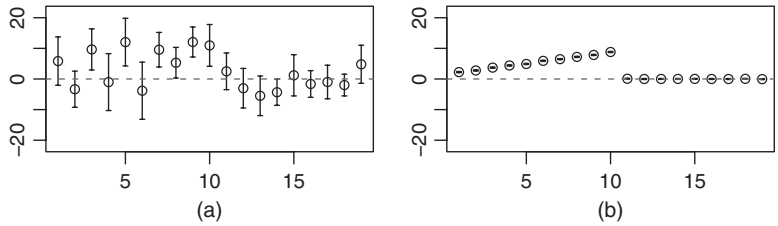

Fig. 1. Average coefficient estimates (over 500 repetitions) and 95% confidence intervals for data generated under a simple model (-------, $y = 0$): (a) $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$; (b) Scout(1, ·)

it results in coefficient estimates that are slightly biased, they have much lower variance. This simple example demonstrates that benefits can result from the use of a shrunken estimate of the inverse covariance matrix.

2.4. Maximization of the scout criteria with L_p-penalties

If $\lambda_1 = 0$, then the maximum of the first scout criterion is given by $(\mathbf{S_{xx}})^{-1}$ (if $\mathbf{S_{xx}}$ is invertible). In the case that $\lambda_1 > 0$ and $p_1 = 1$, maximization of the first Scout criterion has been studied extensively; see for example Meinshausen and Bühlmann (2006). The solution can be found via the 'graphical lasso', which is an efficient algorithm that was given by Banerjee *et al.* (2008) and Friedman *et al.* (2007) that involves iteratively regressing one row of the estimated covariance matrix onto the others, subject to an L_1-constraint, to update the estimate for that row.

If $\lambda_1 > 0$ and $p_1 = 2$, the solution to step 1 of the scout procedure is even easier. We want to find $\mathbf{\Theta_{xx}}$ that maximizes

$$\log\{\det(\mathbf{\Theta_{xx}})\} - \mathrm{tr}(\mathbf{S_{xx}}\mathbf{\Theta_{xx}}) - \lambda\|\mathbf{\Theta_{xx}}\|^2. \tag{8}$$

Differentiating with respect to $\mathbf{\Theta_{xx}}$, we see that the maximum solves

$$\mathbf{\Theta}_{\mathbf{xx}}^{-1} - 2\lambda\mathbf{\Theta_{xx}} = \mathbf{S_{xx}}. \tag{9}$$

This equation implies that $\mathbf{\Theta_{xx}}$ and $\mathbf{S_{xx}}$ share the same eigenvectors. Letting θ_i denote the ith eigenvalue of $\mathbf{\Theta_{xx}}$ and letting s_i denote the ith eigenvalue of $\mathbf{S_{xx}}$, it is clear that

$$\frac{1}{\theta_i} - 2\lambda\theta_i = s_i. \tag{10}$$

We can easily solve for θ_i and can therefore solve the first scout criterion exactly in the case $p_1 = 2$, in essentially just the computational cost of obtaining the eigenvalues of $\mathbf{S_{xx}}$.

It turns out that, if $p_2 = 1$ or $p_2 = 2$, then it is not necessary to maximize the second scout criterion directly, as there is an easier alternative, as follows.

Assumption 1. For $p_2 \in \{1, 2\}$, the solution to step 3 of the scout procedure is equal to the solution to the following equation, up to scaling by a constant:

$$\hat{\beta} = \arg\min_{\beta}(\beta^{\mathrm{T}}\hat{\Sigma}_{\mathbf{xx}}\beta - 2\mathbf{S}_{\mathbf{xy}}^{\mathrm{T}}\beta + \lambda_2\|\beta\|^{p_2}) \tag{11}$$

where $\hat{\mathbf{\Sigma}}_{\mathbf{xx}}$ is the inverse of the solution to step 1 of the scout procedure.

(The proof of assumption 1 is in Appendix A.1.1.) Therefore, we can replace steps 2 and 3 of the scout procedure with an L_{p_2}-regression. It is trivial to show that, if $\lambda_2 = 0$ in the scout procedure, then the scout solution is given by $\hat{\beta} = (\hat{\mathbf{\Sigma}}_{\mathbf{xx}})^{-1}\mathbf{S_{xy}}$. It also follows that, if $\lambda_1 = 0$, then the cases $\lambda_2 = 0$, $p_2 = 1$ and $p_2 = 2$ correspond to ordinary least squares regression (if the empirical covariance matrix is invertible), the lasso and ridge regression respectively.

Table 2. Maximization of the scout criteria: special cases

	$\lambda_2 = 0$	$p_2 = 1$	$p_2 = 2$
$\lambda_1 = 0$	Least squares	L_1-regression	L_2-regression
$p_1 = 1$	Graphical lasso	Graphical lasso + L_1-regression	Graphical lasso + L_2-regression
$p_1 = 2$	Eigenvalue problem	Elastic net	Eigenvalue problem + L_2-regression

Table 3. Timing comparisons for maximization of the scout criteria†

p	*Results (central processor unit s) for the following methods:*			
	Scout(1, ·)	*Scout(1, 1)*	*Scout(2, ·)*	*Scout(2, 1)*
500	1.685	1.700	0.034	0.072
1000	22.432	22.504	0.083	0.239
2000	241.289	241.483	0.260	0.466

†$\lambda_1 = \lambda_2 = 0.2$, $n = 100$, $\mathbf{X}$ dense and p the number of features.

In addition, we shall show in Section 2.5.1 that, if $p_1 = 2$ and $p_2 = 1$, then the scout can be rewritten as an elastic net problem with slightly different data; therefore, fast algorithms for solving the elastic net (Friedman *et al.*, 2008) can be used to solve Scout(2, 1). The methods for maximizing the scout criteria are summarized in Table 2.

We compared computation times for Scout$(2, \cdot)$, Scout$(1, \cdot)$, Scout(2, 1) and Scout(1, 1) on an example with $n = 100$, $\lambda_1 = \lambda_2 = 0.2$ and $\mathbf{X}$ dense. All timings were carried out on an Intel Xeon 2.80-GHz processor. Table 3 shows the number of central processor unit seconds required for each of these methods for a range of values of p (the number of features). For all methods, after the scout coefficients have been estimated for a given set of parameter values, estimation for different parameter values is faster because an approximate estimate of the inverse covariance matrix is available for use as an initial value (when $p_1 = 1$) or because the eigendecomposition has already been computed (when $p_1 = 2$).

Scout(p_1, p_2) involves the use of two tuning parameters, λ_1 and λ_2; in practice, these are chosen by cross-validating over a grid of (λ_1, λ_2) values. In Section 2.7, we present a Bayesian connection to the first scout criterion. The Joint Editor suggested that, as an alternative to cross-validating over λ_1, one could instead draw from the posterior distribution of $\boldsymbol{\Theta}_{\mathbf{xx}}$.

2.5. Properties of the scout

In this section, for ease of notation, we shall consider an equivalent form of the scout procedure that is obtained by replacing $\mathbf{S}_{\mathbf{xx}}$ with $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and $\mathbf{S}_{\mathbf{xy}}$ with $\mathbf{X}^{\mathrm{T}}\mathbf{y}$.

2.5.1. Similarities between scout, ridge regression and the elastic net

Let $\mathbf{U}_{n\times p}\mathbf{D}_{p\times p}\mathbf{V}^{\mathrm{T}}_{p\times p}$ denote the singular value decomposition of $\mathbf{X}$ with d_i the ith diagonal element of $\mathbf{D}$ and $d_1 \geqslant d_2 \geqslant \ldots \geqslant d_r > d_{r+1} = \ldots = d_p = 0$, where $r = \mathrm{rank}(\mathbf{X}) \leqslant \min(n, p)$.

Consider Scout$(2, p_2)$. As previously discussed, the first step in the scout procedure corresponds to finding $\boldsymbol{\Theta}$ that solves

$$\boldsymbol{\Theta}^{-1} - 2\lambda_1\boldsymbol{\Theta} = \mathbf{X}^{\mathrm{T}}\mathbf{X}. \tag{12}$$

Since $\boldsymbol{\Theta}$ and $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ therefore share the same eigenvectors, it follows that $\boldsymbol{\Theta}^{-1} = \mathbf{V}(\mathbf{D}^2 + \tilde{\mathbf{D}}^2)\mathbf{V}^{\mathrm{T}}$ where $\tilde{\mathbf{D}}^2$ is a $p \times p$ diagonal matrix with ith diagonal entry equal to $\frac{1}{2}\{-d_i{}^2 + \sqrt{(d_i^4 + 8\lambda_1)}\}$. It is not difficult to see that ridge regression, Scout$(2, \cdot)$ and Scout$(2, 2)$ result in similar regression coefficients:

$$\begin{aligned}
\hat{\beta}_{\mathrm{rr}} &= (\mathbf{V}(\mathbf{D}^2 + c\mathbf{I})\mathbf{V}^{\mathrm{T}})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y};\\
\hat{\beta}_{\mathrm{Scout}(2,\cdot)} &= (\mathbf{V}(\mathbf{D}^2 + \tilde{\mathbf{D}}^2)\mathbf{V}^{\mathrm{T}})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y};\\
\hat{\beta}_{\mathrm{Scout}(2,2)} &= (\mathbf{V}(\mathbf{D}^2 + \tilde{\mathbf{D}}^2 + \lambda_2\mathbf{I})\mathbf{V}^{\mathrm{T}})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.
\end{aligned} \tag{13}$$

Therefore, whereas ridge regression simply adds a constant to the diagonal elements of $\mathbf{D}$ in the least squares solution, Scout$(2, \cdot)$ instead adds a function that is monotone decreasing in the value of the diagonal element. (The consequences of this alternative shrinkage are explored under a latent variable model in Section 2.6.) Scout$(2, 2)$ is a compromise between Scout$(2, \cdot)$ and ridge regression.

In addition, we note that the solutions to the naive elastic net and Scout$(2, 1)$ are quite similar to each other:

$$\begin{aligned}
\hat{\beta}_{\mathrm{enet}} &= \arg\min_{\beta}\{\beta^{\mathrm{T}}\mathbf{V}(\mathbf{D}^2 + c\mathbf{I})\mathbf{V}^{\mathrm{T}}\beta - 2\beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \lambda_2\|\beta\|^1\}\\
&= \arg\min_{\beta}(\beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta - 2\beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \lambda_2\|\beta\|^1 + c\|\beta\|^2)\\
&= \arg\min_{\beta}(\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2\|\beta\|^1 + c\|\beta\|^2),\\
\hat{\beta}_{\mathrm{Scout}(2,1)} &= \arg\min_{\beta}\{\beta^{\mathrm{T}}\mathbf{V}(\mathbf{D}^2 + \tilde{\mathbf{D}}^2)\mathbf{V}^{\mathrm{T}}\beta - 2\beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \lambda_2\|\beta\|^1\}\\
&= \arg\min_{\beta}\{\beta^{\mathrm{T}}\mathbf{V}(\tfrac{1}{2}\mathbf{D}^2 + \tfrac{1}{2}\bar{\mathbf{D}}^2)\mathbf{V}^{\mathrm{T}}\beta - 2\beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \lambda_2\|\beta\|^1 + \sqrt{(2\lambda_1)}\|\beta\|^2\}\\
&= \arg\min_{\beta}\{\|\mathbf{y}^* - \mathbf{X}^*\beta\|^2 + \lambda_2\|\beta\|^1 + \sqrt{(2\lambda_1)}\|\beta\|^2\}
\end{aligned} \tag{14}$$

where $\bar{\mathbf{D}}^2$ is the diagonal matrix with elements $\sqrt{(d_i^4 + 8\lambda_1)} - \sqrt{(8\lambda_1)}$, and where

$$\mathbf{X}^* = \frac{1}{\sqrt{2}}\begin{pmatrix}\mathbf{X}\\ \bar{\mathbf{D}}\mathbf{V}^{\mathrm{T}}\end{pmatrix},$$
$$\mathbf{y}^* = \begin{pmatrix}\sqrt{2}\mathbf{y}\\ 0\end{pmatrix}.$$

From equation (14), it is clear that Scout(2, 1) solutions can be obtained by using software for the elastic net on data $\mathbf{X}^*$ and $\mathbf{y}^*$. In addition, given the similarity between the elastic net and Scout(2, 1) solutions, it is not surprising that Scout(2, 1) shares some of the elastic net's desirable properties, as is shown in Section 2.5.2.

2.5.2. Variable grouping effect

Zou and Hastie (2005) showed that, unlike the lasso, the elastic net and ridge regression have a variable grouping effect: correlated variables result in similar coefficients. The same is true of Scout(2, 1).

Assumption 2. Assume that the predictors are standardized and that $\mathbf{y}$ is centred. Let ρ denote the correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$, and let $\hat{\beta}$ denote the solution to Scout(2, 1). If $\hat{\beta}_i\hat{\beta}_j \neq 0$, then the following inequality holds:

$$|\hat{\beta}_i - \hat{\beta}_j| \leqslant \sqrt{\left\{\frac{2(1-\rho)}{\lambda_1}\right\}}\|\mathbf{y}\|. \tag{15}$$

The proof of assumption 2 is in Appendix A.1.2. Similar results hold for Scout$(2, \cdot)$ and Scout(2, 2), without the requirement that $\hat{\beta}_i\hat{\beta}_j \neq 0$.

2.5.3. *Connections to regression with orthogonal features*

Assume that the features are standardized, and consider the scout criterion with $p_1 = 1$. For λ_1 sufficiently large, the solution $\hat{\mathbf{\Theta}}_{\mathbf{xx}}$ to the first scout criterion (equation (6)) is a diagonal matrix with diagonal elements $1/(\lambda_1 + \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i)$. (More specifically, if $\lambda_1 \geqslant |\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j|$ for all $i \neq j$, then the scout criterion with $p_1 = 1$ results in a diagonal matrix; see Banerjee *et al.* (2008), theorem 4.) Thus, if $\hat{\beta}_i$ is the ith component of the Scout$(1, \cdot)$ solution, then $\hat{\beta}_i = \mathbf{x}_i^{\mathrm{T}}\mathbf{y}/(\lambda_1 + 1)$. If $\lambda_2 > 0$, then the resulting scout solutions with $p_2 = 1$ are given by a variation of the univariate soft thresholding formula for L_1-regression:

$$\hat{\beta}_i = \frac{1}{\lambda_1 + 1}\mathrm{sgn}(\mathbf{x}_i^{\mathrm{T}}\mathbf{y})\max\left(0, |\mathbf{x}_i^{\mathrm{T}}\mathbf{y}| - \frac{\lambda_2}{2}\right). \tag{16}$$

Similarly, if $p_2 = 2$, the resulting scout solutions are given by the formula

$$\hat{\beta} = (1 + \lambda_1 + \lambda_2)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{17}$$

Therefore, as the parameter λ_1 is increased, the solutions that are obtained range (up to a scaling) from the ordinary L_{p_2} multivariate regression solution to the regularized regression solution for orthonormal features.

2.6. *An underlying latent variable model*

Let $\mathbf{X}$ be an $n \times p$ matrix of n observations on p variables, and $\mathbf{y}$ an $n \times 1$ vector of response values. Suppose that $\mathbf{X}$ and $\mathbf{y}$ are generated under the following latent variable model:

$$\left.\begin{array}{c} \mathbf{X} = d_1\mathbf{u}_1\mathbf{v}_1^{\mathrm{T}} + d_2\mathbf{u}_2\mathbf{v}_2^{\mathrm{T}}, \\ d_1, d_2 > 0, \\ \mathbf{y} = \mathbf{u}_1 + \varepsilon, \\ \mathrm{var}(\varepsilon) = \sigma^2\mathbf{I}, \\ E(\varepsilon) = 0 \end{array}\right\} \tag{18}$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are the singular vectors of $\mathbf{X}$, and ε is an $n \times 1$ vector of noise.

Assumption 3. Under this model, if $d_1 > d_2$ and the tuning parameters for ridge regression and Scout$(2, \cdot)$ are chosen so that the resulting estimators have the same amount of bias, then the estimator that is given by Scout$(2, \cdot)$ will have lower variance.

The proof of assumption 3 is given in Appendix A.1.3. Note that, if $\mathbf{v}_1$ and $\mathbf{v}_2$ are sparse with non-overlapping regions of non-sparsity, then the model yields a block diagonal covariance matrix with two blocks, where one of the blocks of correlated features is associated with the outcome. In the case of gene expression data, these blocks could represent gene pathways, one of which is responsible for, and has expression that is correlated with, the outcome. Assumption

3 shows that, if the signal that is associated with the relevant gene pathway is sufficiently large, then Scout$(2,\cdot)$ will provide a benefit over ridge regression.

2.7. Bayesian connection to the first scout criterion

Consider the model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ where ε_i are independent Gaussian random variables. It is well known that ridge regression, the lasso and the elastic net can be viewed as the Bayes posterior modes under various priors, since they involve solving for β that minimizes a criterion of the form

$$(\mathbf{y} - \mathbf{X}\beta)^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\beta) + \lambda_1 \|\beta\|^{p_1} + \lambda_2 \|\beta\|^{p_2}. \tag{19}$$

Ridge regression corresponds to a normal prior on the elements of β, the lasso corresponds to a double-exponential prior and the elastic net corresponds to a prior that is a combination of the two.

Similarly, we can think of the solution to the first scout criterion as the Bayes mode of the posterior distribution given by $\mathbf{X} \sim N(0, \mathbf{\Sigma})$ and a prior on the elements of $\mathbf{\Sigma}^{-1}$, such that, for $i \leqslant j$, $(\mathbf{\Sigma}^{-1})_{ij}$ is independent and identically distributed with either a Gaussian distribution (if $p_1 = 2$) or a double-exponential distribution (if $p_1 = 1$). Formally, this would have the potential difficulty that draws from the prior distribution are not constrained to be positive semidefinite.

3. Numerical studies: regression via the scout

3.1. Simulated data

We compare the performance of ordinary least squares, the lasso, the elastic net, Scout(2,1) and Scout(1,1) on a suite of six simulated examples. The first four simulations are based on those used in Zou and Hastie (2005) and Tibshirani (1996). The fifth and sixth are of our own invention. All simulations are based on the model $\mathbf{y} = \mathbf{X}\beta + \sigma\varepsilon$ where $\varepsilon \sim N(0, \mathbf{I})$. For each simulation, each data set consists of a small training set, a small validation set (which is used to select the values of the various parameters) and a large test set. We indicate the size of the training, validation and test sets by using the notation $\cdot/\cdot/\cdot$. The simulations are as follows.

(a) Simulation 1: each data set consists of 20/20/200 observations, eight predictors with coefficients $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^{\mathrm{T}}$ and $\sigma = 3$. $\mathbf{X} \sim N(0, \mathbf{\Sigma})$, where $\Sigma_{ij} = 0.5^{|i-j|}$.
(b) Simulation 2: this simulation is as in simulation 1, except that $\beta_i = 0.85$ for all i.
(c) Simulation 3: each data set consists of 100/100/400 observations and 40 predictors. $\beta_i = 0$ for $i \in 1, \ldots, 10$ and for $i \in 21, \ldots, 30$; for all other i, $\beta_i = 2$. We also set $\sigma = 15$. $\mathbf{X} \sim N(0, \mathbf{\Sigma})$, where $\Sigma_{ij} = 0.5$ for $i \neq j$, and $\Sigma_{ii} = 1$.
(d) Simulation 4: each data set consists of 50/50/400 observations and 40 predictors. $\beta_i = 3$ for $i \in 1, \ldots, 15$ and $\beta_i = 0$ for $i \in 16, \ldots, 40$, and $\sigma = 15$. The predictors are generated as follows:

$$\mathbf{x}_i = \begin{cases} \mathbf{z}_1 + \varepsilon_i^x, & \mathbf{z}_1 \sim N(0, \mathbf{I}),\ i = 1, \ldots, 5; \\ \mathbf{z}_2 + \varepsilon_i^x, & \mathbf{z}_2 \sim N(0, \mathbf{I}),\ i = 6, \ldots, 10; \\ \mathbf{z}_3 + \varepsilon_i^x, & \mathbf{z}_3 \sim N(0, \mathbf{I}),\ i = 11, \ldots, 15. \end{cases} \tag{20}$$

Also, $\mathbf{x}_i \sim N(0, \mathbf{I})$ are independent and identically distributed for $i = 16, \ldots, 40$, and $\varepsilon_i^x \sim N(0, 0.01\mathbf{I})$ are independent and identically distributed for $i = 1, \ldots, 15$.
(e) Simulation 5: each data set consists of 50/50/400 observations and 50 predictors; $\beta_i = 2$ for $i < 9$ and $\beta_i = 0$ for $i \geqslant 9$. $\sigma = 6$ and $\mathbf{X} \sim N(0, \mathbf{\Sigma})$, where $\Sigma_{ij} = 0.5\mathbf{I}_{i,j \leqslant 9}$ for $i \neq j$, and $\Sigma_{ii} = 1$.
(f) Simulation 6: as in simulation 1, but $\beta = (3, 1.5, 0, 0, 0, 0, -1, -1)^{\mathrm{T}}$.

Table 4. Mean-squared error averaged over 200 simulated data sets for each simulation†

Simulation	*Results for the following methods:*				
	Least squares	*Lasso*	*Elastic net*	*Scout(1, 1)*	*Scout(2, 1)*
1	7.72 (0.46)	2.83 (0.16)	*2.28 (0.13)*	*2.22 (0.13)*	2.29 (0.13)
2	7.72 (0.46)	3.26 (0.13)	2.28 (0.11)	*1.31 (0.09)*	*1.54 (0.09)*
3	158.29 (3.66)	44.07 (0.80)	30.86 (0.47)	*20.44 (0.25)*	*18.94 (0.28)*
4	1094.84 (44.75)	54.79 (2.30)	*25.06 (1.62)*	30.21 (1.61)	*28.37 (1.52)*
5	—	10.91 (0.38)	2.46 (0.09)	*1.62 (0.09)*	*2.18 (0.11)*
6	7.72 (0.46)	2.95 (0.16)	2.34 (0.13)	*2.12 (0.11)*	*2.15 (0.11)*

†Standard errors are given in parentheses. For each simulation, the two methods with lowest average mean-squared errors are shown in italics. Least squares analysis was not performed for simulation 5, because $p = n$.

Table 5. Median L_2-distance over 200 simulated data sets for each simulation†

Simulation	*Results for the following methods:*				
	Least squares	*Lasso*	*Elastic net*	*Scout(1, 1)*	*Scout(2, 1)*
1	3.05 (0.10)	1.74 (0.05)	1.65 (0.08)	*1.58 (0.05)*	*1.62 (0.06)*
2	3.05 (0.10)	1.95 (0.02)	1.62 (0.03)	*0.90 (0.03)*	*1.04 (0.04)*
3	17.03 (0.22)	8.91 (0.09)	7.70 (0.06)	*6.15 (0.01)*	*5.83 (0.03)*
4	168.40 (5.13)	17.40 (0.16)	*3.85 (0.13)*	5.19 (2.3)	*3.80 (0.14)*
5	—	3.48 (0.06)	2.08 (0.06)	*1.15 (0.03)*	*1.55 (0.05)*
6	3.05 (0.10)	1.76 (0.06)	1.53 (0.05)	*1.48 (0.04)*	*1.50 (0.03)*

†The details are the same as in Table 4.

These simulations cover a variety of settings: simulations 1, 3, 4, 5 and 6 have sparse β, simulations 1, 2, 4, 5 and 6 have a sparse inverse covariance matrix and simulation 4 is characterized by groups of variables that contribute to the response. For each simulation, 200 data sets were generated, and the average mean-squared errors (with standard errors given in parentheses) are given in Table 4. The scout provides an improvement over the lasso in all simulations. Both scout methods result in lower mean-squared error than the elastic net in simulations 2, 3, 5 and 6; in simulations 1 and 4, the scout methods are quite competitive. Table 5 shows median L_2-distances between the true and estimated coefficients for each of the models.

Though Scout(2,1) and Scout(1,1) perform well relative to the elastic net and lasso in all six simulations, neither dominates the others in all cases. For a given application, we recommend selecting a regression method based on cross-validation error (with the *caveat* that Scout(1,1) is slow when the number of features is very large).

3.2. Scout using alternative covariance estimators

A referee asked whether a different estimator of the inverse covariance matrix of $\mathbf{X}$ could be used in step 2 of the scout procedure, rather than the solution to the first scout criterion. A large body of literature exists on estimation of covariance and inverse covariance matrices. Examples include James and Stein (1961), Haff (1979), Dey and Srinivasan (1985), Bickel and Levina (2008) and Rothman *et al.* (2008). Any covariance estimate can be plugged in for $\hat{\boldsymbol{\Sigma}}$ in the

Table 6. On simulation 2, comparison of two new estimators obtained by plugging in $\hat{\Sigma}^{JS}$ and $\hat{\Sigma}^{BL}$ to equation (21)†

Quantity	*Results for the following methods:*			
	Scout(JS, 1)	*Scout(BL, 1)*	*Scout(1, 1)*	*Scout(2, 1)*
Mean mean-squared error	3.79 (0.15)	2.42 (0.12)	1.31 (0.09)	1.54 (0.09)
Median L_2-distance	3.34 (0.14)	1.94 (0.10)	0.90 (0.03)	1.04 (0.04)

†Tuning parameter values were chosen by cross-validation, and standard errors are in parentheses.

equation from assumption 1:

$$\hat{\beta} = \arg\min_{\beta}(\beta^{T}\hat{\Sigma}\beta - 2\mathbf{S}_{\mathbf{xy}}^{T}\beta + \lambda_2\|\beta\|^{s}). \tag{21}$$

We explore that possibility here with two covariance estimates: the estimator of James and Stein (1961), and the hard thresholding estimator of Bickel and Levina (2008). The James–Stein estimator takes the form $\hat{\Sigma}^{JS} = \mathbf{TDT}^{T}$ where $\mathbf{T}$ is a lower triangular matrix with positive elements on the diagonal such that $\mathbf{TT}^{T} = \mathbf{X}^{T}\mathbf{X}$, and $\mathbf{D}$ is a diagonal matrix with diagonal elements $d_i = 1/(n + p + 1 - 2i)$. It is the constant risk minimax estimator under Stein's loss. $\hat{\Sigma}^{BL}$, the estimator of Bickel and Levina (2008), is obtained by hard thresholding each element of the empirical covariance matrix. With $s = 1$ in equation (22), the resulting methods (which we call Scout(JS, 1) and Scout(BL, 1)) are compared with Scout(2, 1) and Scout(1, 1) on simulation 2, described in Section 3.1. The results are shown in Table 6. In this example, Scout(JS, 1) and Scout(BL, 1) do not perform as well as Scout(1, 1) and Scout(2, 1).

3.3. Making use of observations without response values

In step 1 of the scout procedure, we estimate the inverse covariance matrix based on the training set $\mathbf{X}$-data and, in steps 2–4, we compute a penalized least squares solution based on that estimated inverse covariance matrix and $\widehat{\text{cov}}(\mathbf{X}, \mathbf{y})$. Step 1 of this procedure does not involve the response $\mathbf{y}$ at all.

Now, consider a situation in which we have access to a large amount of $\mathbf{X}$-data, but responses are known for only some of the observations. (For instance, this could be the case for a medical researcher who has clinical measurements on hundreds of cancer patients, but survival times for only dozens of patients.) More specifically, let $\mathbf{X}_1$ denote the observations for which there is an associated response $\mathbf{y}$, and let $\mathbf{X}_2$ denote the observations for which no response data are available. Then, we could estimate the inverse covariance matrix in step 1 of the scout procedure by using both $\mathbf{X}_1$ and $\mathbf{X}_2$, and perform step 2 by using $\widehat{\text{cov}}(\mathbf{X}_1, \mathbf{y})$. By also using $\mathbf{X}_2$ in step 1, we achieve a more accurate estimate of the inverse covariance matrix than would have been possible by using only $\mathbf{X}_1$.

Such an approach will not provide an improvement in all cases. For instance, consider the trivial case in which the response is a linear function of the predictors, $p < n$, and there is no noise: $\mathbf{y} = \mathbf{X}_1\beta$. Then, the least squares solution, using only $\mathbf{X}_1$ and not $\mathbf{X}_2$, is $\hat{\beta} = (\mathbf{X}_1^{T}\mathbf{X}_1)^{-1}\mathbf{X}_1^{T}\mathbf{y} = (\mathbf{X}_1^{T}\mathbf{X}_1)^{-1}\mathbf{X}_1^{T}\mathbf{X}_1\beta = \beta$. In this case, it clearly is best to use only $\mathbf{X}_1$ in estimating the inverse covariance matrix. However, one can imagine situations in which we can use $\mathbf{X}_2$ to obtain a more accurate estimate of the inverse covariance matrix.

Table 7. Making use of observations without response values: set-up

	Sample size	*Response description*
Training set	12	Available
Test set	200	Unavailable—must be predicted
Additional observations	36	Unavailable—not of interest

Consider a model in which a latent variable has generated some of the features, as well as the response. In particular, suppose that the data are generated as follows:

$$\left.\begin{aligned} x_{ij} &= 2u_i + \varepsilon_{ij}, && j = 1, \ldots, 5,\ i = 1, \ldots, n; \\ x_{ij} &= \varepsilon_{ij}, && j = 6, \ldots, 10,\ i = 1, \ldots, n; \\ y_i &= 8u_i + 4\varepsilon_i', && i = 1, \ldots, n. \end{aligned}\right\} \tag{22}$$

In addition, we let $\varepsilon_{ij}, \varepsilon_i', u_i \sim N(0,1)$ independent and identically distributed. The first five variables are 'signal' variables, and the rest are 'noise' variables. Suppose that we have three sets of observations: a training set of size $n = 12$, for which the $\mathbf{y}$-values are known, a test set of size $n = 200$, for which we wish to predict the $\mathbf{y}$-values, and an additional set of size $n = 36$ observations for which we do not know the $\mathbf{y}$-values and do not wish to predict them. This layout is shown in Table 7.

We compare the performances of the scout and other regression methods. The scout method is applied in two ways: using only the training set $\mathbf{X}$-values to estimate the inverse covariance matrix, and using also the observations without response values. All tuning parameter values are chosen by sixfold cross-validation. The results in Table 8 are the average mean-squared prediction errors obtained over 500 simulations. From Table 8, it is clear that both versions of the scout outperform all the other methods. In addition, using observations that do not have response values does result in a significant improvement.

Table 8. Results by making use of observations without response values†

Method	*Mean-squared prediction error*
Scout(1, ·) with additional observations	25.65 (0.38)
Scout(1, ·) without additional observations	29.92 (0.62)
Elastic net	32.38 (1.04)
Lasso	47.24 (3.58)
Least squares	1104.9 (428.84)
Null model	79.24 (0.3)

†Standard errors are shown in parentheses. The 'null model' predicts each test set outcome value by using the mean of the training set outcomes.

In this example, 12 labelled observations on 10 variables do not suffice to estimate the inverse covariance matrix reliably. The scout can make use of the observations that lack response values to improve the estimate of the inverse covariance matrix, thereby yielding superior predictions.

The use of unlabelled data for classification and regression is sometimes called *semisupervised learning*. The use of unlabelled observations for linear discriminant analysis dates back to O'Neill (1978); other classical work in this area can be found in McLachlan (1992). It is currently an active area of research in the statistical and machine learning communities. Liang *et al.* (2007) have presented an overview of the use of unlabelled data for prediction, as well as the underlying theory.

4. Classification via the scout

In classification problems, linear discriminant analysis can be used if $n > p$. However, when $p > n$, regularization of the within-class covariance matrix is necessary. Regularized linear discriminant analysis is discussed in Friedman (1989) and Guo *et al.* (2007). In Guo *et al.* (2007), the within-class covariance matrix is shrunken, as in ridge regression, by adding a multiple of the identity matrix to the empirical covariance matrix. Here, we instead estimate a shrunken within-class inverse covariance matrix by maximizing the log-likelihood of the data, under a multivariate normal model, subject to an L_p-penalty on its elements.

4.1. Details of extension of scout to classification

Consider a classification problem with K classes; each observation belongs to some class $k \in 1, \ldots, K$. Let $C(i)$ denote the class of training set observation i, which is denoted X_i. Our goal is to classify observations in an independent test set.

Let $\hat{\mu}_k$ denote the $p \times 1$ vector that contains the mean of observations in class k, and let

$$\mathbf{S}_{\mathrm{wc}} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:C(i)=k} (X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^{\mathrm{T}}$$

denote the estimated within-class covariance matrix (based on the training set) that is used for ordinary linear discriminant analysis. Then, the *scout procedure for classification* is as follows.

Step 1: compute the shrunken within-class inverse covariance matrix $\hat{\mathbf{\Sigma}}_{\mathrm{wc},\lambda}^{-1}$:

$$\hat{\mathbf{\Sigma}}_{\mathrm{wc},\lambda}^{-1} = \arg\max_{\mathbf{\Sigma}^{-1}} [\log\{\det(\mathbf{\Sigma}^{-1})\} - \mathrm{tr}(\mathbf{S}_{\mathrm{wc}}\mathbf{\Sigma}^{-1}) - \lambda \|\mathbf{\Sigma}^{-1}\|^s] \tag{23}$$

where λ is a shrinkage parameter.

Step 2: classify test set observation X to class k' if $k' = \arg\max_k \{\delta_k^{\lambda}(X)\}$, where

$$\delta_k^{\lambda}(X) = X^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{\mathrm{wc},\lambda}^{-1} \hat{\mu}_k - \tfrac{1}{2} \hat{\mu}_k^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{\mathrm{wc},\lambda}^{-1} \hat{\mu}_k + \log(\pi_k) \tag{24}$$

and π_k is the frequency of class k in the training set.

This procedure is analogous to LDA, but we have replaced $\mathbf{S}_{\mathrm{wc}}$ with a shrunken estimate.

This classification rule performs quite well on real microarray data (as is shown below) but has the drawback that it makes use of all the genes. We can remedy this in one of two ways. We can apply the method described above to only the genes with highest univariate rankings on the training data; this is done in the next section. Alternatively, we can apply an L_1-penalty in estimating the quantity $\hat{\mathbf{\Sigma}}_{\mathrm{wc},\lambda}^{-1} \hat{\mu}_k$; note (from equation (24)) that sparsity in this quantity will result in a classification rule that is sparse in the features. Details of this second method,

which is not implemented here, are given in Appendix A.2. We shall refer to the method that is detailed in equations (23) and (24) as Scout$(s, \cdot)$ because the penalized log-likelihood that is maximized in equation (23) is analogous to the first scout criterion in the regression case. The tuning parameter λ in equations (23) and (24) can be chosen via cross-validation.

4.2. Ramaswamy data

We assess the performance of this method on the Ramaswamy microarray data set, which was discussed in detail in Ramaswamy *et al.* (2001) and explored further in Zhu and Hastie (2004) and Guo *et al.* (2007). It consists of a training set of 144 samples and a test set of 54 samples, each of which contains measurements on 16063 genes. The samples are classified into 14 distinct cancer types. We compare the performance of Scout$(2, \cdot)$ with nearest shrunken centroids (Tibshirani *et al.*, 2002, 2003), L_2-penalized multiclass logistic regression (Zhu and Hastie, 2004), the support vector machine with one-*versus*-all classification (Ramaswamy *et al.*, 2001), regularized discriminant analysis (Guo *et al.*, 2007), random forests (Breiman, 2001) and k

Table 9. Comparison of methods on the Ramaswamy data†

Method	*Cross-validation error*	*Test error*	*Number of genes used*
Nearest shrunken centroids	35	17	5217
L_2-penalized multiclass logistic regression	29	15	16063
Support vector machine	26	14	16063
Regularized discriminant analysis	27	11	9141
k nearest neighbours	41	29	16063
Random forests	40	24	16063
Scout$(2, \cdot)$	22	11	16063

†All methods were performed on the cube-rooted data, after centring and scaling each patient.

Table 10. Comparison of methods on the Ramaswamy data†

Method	*Test error*	*Number of genes used*
Nearest shruken centroids	21	3999
L_2-penalized multiclass logistic regression	12	4000
Support vector machine	11	4000
Regularized discriminant analysis	10	3356
k nearest neighbours	17	4000
Random forests	17	4000
Scout$(2, \cdot)$	7	4000

†The methods were run on the cube-rooted data after centring and scaling the patients, using only the 4000 genes with highest training set F-statistics.

nearest neighbours. For each method, tuning parameter values were chosen by cross-validation. The results can be seen in Tables 9 and 10; Scout$(2, \cdot)$ performed quite well, especially when only the 4000 genes with highest training set F-statistics were used (Tusher *et al.*, 2001).

5. Extension to generalized linear models and the Cox model

We have discussed the application of the scout to classification and regression problems, and we have shown examples in which these methods perform well. In fact, the scout can also be used in fitting generalized linear models, by replacing the iteratively reweighted least squares step with a covariance-regularized regression. In particular, we discuss the use of the scout in the context of fitting a Cox proportional hazards model for survival data. We present an example involving four lymphoma microarray data sets in which the scout results in improved performance relative to other methods.

5.1. Details of extension of scout to the Cox model

Consider survival data of the form $(y_i, \mathbf{x}^i, \delta_i)$ for $i \in 1, \ldots, n$, where δ_i is an indicator variable that equals 1 if observation i is complete and 0 if censored, and $\mathbf{x}^i$ is a vector of predictors $(x_1^i, \ldots, x_p^i)$ for individual i. Failure times are $t_1 < t_2 < \ldots < t_k$; there are d_i failures at time t_i. We wish to estimate the parameter $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ in the proportional hazards model

$$\lambda(t|x) = \lambda_0(t) \exp\Big(\sum_j x_j \beta_j \Big).$$

We assume that censoring is non-informative. Letting $\eta = \mathbf{X}\beta$, D the set of indices of the failures, R_r the set of indices of the individuals at risk at time t_r and D_r the set of indices of the failures at t_r, the partial likelihood is given as follows (see for example Kalbfleisch and Prentice (1980)):

$$L(\boldsymbol{\beta}) = \prod_{r \in D} \frac{\exp\Big(\sum_{j \in D_r} \eta_j \Big)}{\Big\{ \sum_{j \in R_r} \exp(\eta_j) \Big\}^{d_r}}. \tag{25}$$

To fit the proportional hazards model, we must find the β that maximizes this partial likelihood. Let $l(\beta)$ denote the log-partial-likelihood, $\mathbf{u} = \partial l / \partial \eta$, and $\mathbf{A} = -\partial^2 l / \partial \eta \eta^{\mathrm{T}}$. The iteratively reweighted least squares algorithm that implements the Newton–Raphson method, for β_0 the value of β from the previous step, involves finding β that solves

$$\mathbf{X}^{\mathrm{T}} \mathbf{A} \mathbf{X} (\beta - \beta_0) = \mathbf{X}^{\mathrm{T}} \mathbf{u}. \tag{26}$$

This is equivalent to finding β that minimizes

$$\| \mathbf{y}^* - \mathbf{X}^* \beta^* \|^2 \tag{27}$$

where $\mathbf{X}^* = \mathbf{A}^{1/2} \mathbf{X}$, $\mathbf{y}^* = \mathbf{A}^{-1/2} \mathbf{u}$ and $\beta^* = \beta - \beta_0$ (Green, 1984).

The traditional iterative reweighted least squares algorithm involves solving the above least squares problem repeatedly, recomputing $\mathbf{y}^*$ and $\mathbf{X}^*$ at each step and setting β_0 equal to the solution β that is attained at the previous iteration. We propose to solve the above equation by using the scout, rather than by a simple linear regression. We have found empirically that good results are obtained if we initially set $\beta_0 = 0$, and then perform just one Newton–Raphson step (using the scout). This is convenient since, for data sets with many features,

solving a scout regression can be time consuming. Therefore, our implementation of the scout method for survival data involves simply performing one Newton–Raphson step, beginning with $\beta_0 = 0$.

Using the notation

$$\Theta = \begin{pmatrix} \mathbf{\Theta_{xx}} & \mathbf{\Theta_{xy}} \\ \mathbf{\Theta_{xy}^T} & \Theta_{yy} \end{pmatrix}$$

and

$$\mathbf{S} = \begin{pmatrix} \mathbf{X^T A X} & \mathbf{X^T u} \\ \mathbf{u^T X} & \mathbf{u^T A^{-1} u} \end{pmatrix},$$

the *scout procedure for the Cox model* for survival data is almost identical to the regression case, as follows.

Step 1: let $\hat{\mathbf{\Theta}}_{\mathbf{xx}}$ maximize

$$\log\{\det(\mathbf{\Theta_{xx}})\} - \operatorname{tr}(\mathbf{S_{xx}\Theta_{xx}}) - \lambda_1 \|\mathbf{\Theta_{xx}}\|^{p_1}. \tag{28}$$

Step 2: let $\hat{\mathbf{\Theta}}$ maximize

$$\log\{\det(\mathbf{\Theta})\} - \operatorname{tr}(\mathbf{S\Theta}) - \lambda_2 \|\mathbf{\Theta}\|^{p_2}, \tag{29}$$

where the top $p \times p$ submatrix of $\mathbf{\Theta}$ is constrained to equal $\hat{\mathbf{\Theta}}_{\mathbf{xx}}$, obtained in the previous step.

Step 3: compute $\hat{\beta} = -\hat{\mathbf{\Theta}}_{\mathbf{xy}} / \hat{\Theta}_{yy}$.

Step 4: let $\hat{\beta}^* = c\hat{\beta}$, where c is the coefficient of a Cox proportional hazards model fit to $\mathbf{y}$ using $\mathbf{X}\beta$ as a predictor.

$\hat{\beta}^*$ obtained in step 4 is the vector of estimated coefficients for the Cox proportional hazards model. In the procedure above, $\lambda_1, \lambda_2 > 0$ are tuning parameters. In keeping with the notation of previous sections, we shall refer to the resulting coefficient estimates as Scout(p_1, p_2).

5.2. Lymphoma data

We illustrate the effectiveness of the scout method on survival data by using four different data sets, all involving survival times and gene expression measurements for patients with diffuse large B-cell lymphoma. The four data sets are as follows: Rosenwald *et al.* (2002), which consists of 240 patients, Shipp *et al.* (2002), which consists of 58 patients, Hummel *et al.* (2006), which consists of 81 patients, and Monti *et al.* (2005), which consists of 129 patients. For consistency and ease of comparison, we considered only a subset of around 1482 genes that were present in all four data sets.

We randomly split each of the data sets into a training set, a validation set and a test set of equal sizes. For each data set, we fit four models to the training set: the L_1-penalized Cox proportional hazards ('L_1-Cox') method of Park and Hastie (2007), the supervised principal components ('SPC') method of Bair and Tibshirani (2004), Scout(2, 1) and Scout(1, 1). For each data set, we chose the tuning parameter values that resulted in the predictor that gave the highest log-likelihood when used to fit a Cox proportional hazards model on the validation set (this predictor was $\mathbf{X}_{\text{val}}\beta_{\text{train}}$ for the L_1-Cox and scout methods, and it was the first supervised principal component for method SPC). We tested the resulting models on the test set. The mean value of $2\{\log(L) - \log(L_0)\}$ over 10 separate training–test–validation set splits is given in Table 11, where L denotes the likelihood of the Cox proportional hazards model fit on the test set using the predictor that was obtained from the training set (for the L_1-Cox and scout methods, this

Table 11. Mean of $2\{\log(L) - \log(L_0)\}$ on the survival data†

Data set	*Results for the following methods:*			
	L_1-Cox	*SPC*	*Scout(1, 1)*	*Scout(2, 1)*
Hummel *et al.* (2006)	2.640 (0.99)	*3.823* (*0.87*)	*4.245* (*1.07*)	3.293 (0.91)
Monti *et al.* (2005)	1.647 (0.36)	1.231 (0.38)	*2.149* (*0.46*)	*2.606* (*0.47*)
Rosenwald *et al.* (2002)	*4.129* (*0.94*)	3.542 (1.17)	3.987 (0.94)	*4.930* (*1.47*)
Shipp *et al.* (2002)	1.903 (0.48)	1.004 (0.39)	*2.807* (*0.73*)	*2.627* (*0.60*)

†For each data set, the two highest mean values of $2\{\log(L) - \log(L_0)\}$ are shown in italics.

Table 12. Median number of genes used for the survival data.

Data set	*Results for the following methods:*			
	L_1-Cox	*SPC*	*Scout(1, 1)*	*Scout(2, 1)*
Hummel *et al.* (2006)	14	33	78	13
Monti *et al.* (2005)	18.5	17	801.5	144.5
Rosenwald *et al.* (2002)	37.5	32	294	85
Shipp *et al.* (2002)	5.5	10	4.5	5

was $\mathbf{X}_{\text{test}}\beta_{\text{train}}$ and, for method SPC, this was the first supervised principal component), and L_0 denotes the likelihood of the null model. From Tables 11 and 12, it is clear that the scout results in predictors that are on par with, if not better than, the competing methods on all four data sets.

6. Discussion

We have presented covariance-regularized regression, a class of regression procedures (the 'scout' family) that is obtained by estimating the inverse covariance matrix of the data by maximizing the log-likelihood of the data under a multivariate normal model, subject to a penalty. We have shown that three well-known regression methods—ridge regression, the lasso and the elastic net—fall into the covariance-regularized regression framework. In addition, we have explored some new methods within this framework. We have extended the covariance-regularized regression framework to classification and generalized linear model settings, and we have demonstrated the performance of the resulting methods on some gene expression data sets.

A drawback of the scout method is that, when $p_1 = 1$ and the number of features is large, then maximizing the first scout criterion can be quite slow. When more than a few thousand features are present, the scout with $p_1 = 1$ is not a viable option at present. However, the scout with $p_1 = 2$ is very fast, and we are confident that computational and algorithmic improvements will lead to increases in the number of features for which the scout criteria can be maximized with $p_1 = 1$.

Roughly speaking, the method in this paper consists of two steps.

(a) The features $\mathbf{X}$ are used to obtain a regularized estimate of the inverse covariance matrix; this can be thought of as 'preprocessing' the features.

(b) The preprocessed features are combined with the outcome **y** to obtain estimated regression coefficients.

In step (a), the features are preprocessed without using the outcome **y**. Indeed, many methods in the machine learning literature involve preprocessing the features without using the outcome. Principal components regression is a classical example of this; a more recent example with much more extensive preprocessing is in Hinton *et al.* (2006).

It has been shown that, for the lasso to exhibit model selection consistency, certain conditions on the feature matrix **X** must be satisfied (see, for instance, the 'irrepresentability condition' of Zhao and Yu (2006)). A reviewer asked whether the scout can offer a remedy in situations where these conditions are not satisfied. This is an interesting question that seems quite difficult to answer. We hope that it will be addressed in future work.

Covariance-regularized regression represents a new way to understand existing regularization methods for regression, as well as an approach to develop new regularization methods that appear to perform better in many examples.

Acknowledgements

We thank the Joint Editor and two reviewers for helpful comments. We thank Trevor Hastie for showing us the solution to the penalized log-likelihood with an L_2-penalty. We thank both Trevor Hastie and Jerome Friedman for valuable discussions and for providing the code for the L_2-penalized multiclass logistic regression and the elastic net. Daniela Witten was supported by a National Defense Science and Engineering Graduate Fellowship. Robert Tibshirani was partially supported by National Science Foundation grant DMS-9971405 and National Institutes of Health contract N01-HV-28183.

Appendix A

A.1. Proofs of assumptions

A.1.1. Proof of assumption 1

First, suppose that $p_2 = 1$. Consider the penalized log-likelihood

$$\log\{\det(\boldsymbol{\Theta})\} - \operatorname{tr}(\mathbf{S}\boldsymbol{\Theta}) - \frac{\lambda_2}{2}\|\boldsymbol{\Theta}\|^1 \tag{30}$$

with $\boldsymbol{\Theta}_{\mathbf{xx}}$ the top left-hand $p \times p$ submatrix of $\boldsymbol{\Theta}$, fixed to equal the matrix that maximizes the log-likelihood in step 1 of the scout procedure. It is clear that, if $\hat{\boldsymbol{\Theta}}$ maximizes the log-likelihood, then $(\hat{\boldsymbol{\Theta}}^{-1})_{yy} = S_{yy} + \lambda_2/2$. The subgradient equation for maximization of the remaining portion of the log-likelihood is

$$0 = (\boldsymbol{\Theta}^{-1})_{\mathbf{xy}} - \mathbf{S}_{\mathbf{xy}} - \frac{\lambda_2}{2}\Gamma \tag{31}$$

where $\Gamma_i = 1$ if the ith element of $\boldsymbol{\Theta}_{\mathbf{xy}}$ is positive, $\Gamma_i = -1$ if the ith element of $\boldsymbol{\Theta}_{\mathbf{xy}}$ is negative and otherwise Γ_i is between -1 and 1.

Let $\beta = \boldsymbol{\Theta}_{\mathbf{xx}}(\boldsymbol{\Theta}^{-1})_{\mathbf{xy}}$. Therefore, we equivalently wish to find β that solves

$$0 = 2(\boldsymbol{\Theta}_{\mathbf{xx}})^{-1}\beta - 2\mathbf{S}_{\mathbf{xy}} - \lambda_2\Gamma. \tag{32}$$

From the partitioned inverse formula, it is clear that $\operatorname{sgn}(\beta) = -\operatorname{sgn}(\boldsymbol{\Theta}_{\mathbf{xy}})$. Therefore, our task is equivalent to finding β which minimizes

$$\beta^{\mathrm{T}}(\boldsymbol{\Theta}_{\mathbf{xx}})^{-1}\beta - 2\mathbf{S}_{\mathbf{xy}}^{\mathrm{T}}\beta + \lambda_2\|\beta\|^1. \tag{33}$$

Of course, this is equation (11). It is an L_1-penalized regression of **y** onto **X**, using only the inner products, with $\mathbf{S}_{\mathbf{xx}}$ replaced with $(\boldsymbol{\Theta}_{\mathbf{xx}})^{-1}$. In other words, $\hat{\beta}$ that solves equation (11) is given by $\boldsymbol{\Theta}_{\mathbf{xx}}(\boldsymbol{\Theta}^{-1})_{\mathbf{xy}}$, where $\boldsymbol{\Theta}$ solves step 2 of the scout procedure.

Now, the solution to step 3 of the scout procedure is $-\boldsymbol{\Theta}_{\mathbf{xy}}/\boldsymbol{\Theta}_{\mathbf{yy}}$. By the partitioned inverse formula, $\boldsymbol{\Theta}_{\mathbf{xx}}(\boldsymbol{\Theta}^{-1})_{\mathbf{xy}} + \boldsymbol{\Theta}_{\mathbf{xy}}(\boldsymbol{\Theta}^{-1})_{yy} = 0$, so

$$-\frac{\boldsymbol{\Theta}_{\mathbf{xy}}}{\Theta_{yy}} = \frac{\boldsymbol{\Theta}_{\mathbf{xx}}(\boldsymbol{\Theta}^{-1})_{\mathbf{xy}}}{(\boldsymbol{\Theta}^{-1})_{yy}\Theta_{yy}} = \frac{\beta}{(\boldsymbol{\Theta}^{-1})_{yy}\Theta_{yy}}.$$

In other words, the solution to step 3 of the scout procedure and the solution to equation (11) differ by a factor of $(\boldsymbol{\Theta}^{-1})_{yy}\Theta_{yy}$. Since step 4 of the scout procedure involves scaling the solution to step 3 by a constant, it is clear that we can replace step 3 of the scout procedure with the solution to equation (11).

Now, suppose that $p_2 = 2$. To find $\boldsymbol{\Theta}_{\mathbf{xy}}$ that maximizes this penalized log-likelihood, we take the gradient and set it to 0:

$$0 = (\boldsymbol{\Theta}^{-1})_{\mathbf{xy}} - \mathbf{S}_{\mathbf{xy}} - \frac{\lambda_2}{2}\boldsymbol{\Theta}_{\mathbf{xy}}. \tag{34}$$

Again, let $\beta = \boldsymbol{\Theta}_{\mathbf{xx}}(\boldsymbol{\Theta}^{-1})_{\mathbf{xy}}$. Therefore, we equivalently wish to find β that solves

$$0 = 2(\boldsymbol{\Theta}_{\mathbf{xx}})^{-1}\beta - 2\mathbf{S}_{\mathbf{xy}} + 2\lambda_3\beta \tag{35}$$

for some new constant λ_3, using the fact, from the partitioned inverse formula, that $-\beta/(\boldsymbol{\Theta}^{-1})_{yy} = \boldsymbol{\Theta}_{\mathbf{xy}}$. The solution β minimizes

$$\beta^{\mathrm{T}}(\boldsymbol{\Theta}_{\mathbf{xx}})^{-1}\beta - 2\mathbf{S}_{\mathbf{xy}}^{\mathrm{T}}\beta + \lambda_3\beta^{\mathrm{T}}\beta.$$

Of course, this is again equation (11). Therefore, $\hat{\beta}$ that solves equation (11) is given (up to scaling by a constant) by $\boldsymbol{\Theta}_{\mathbf{xx}}(\boldsymbol{\Theta}^{-1})_{\mathbf{xy}}$, where $\boldsymbol{\Theta}$ solves step 2 of the scout procedure. As before, by the partitioned inverse formula, and since step 4 of the scout procedure involves scaling the solution to step 3 by a constant, it is clear that we can replace step 3 of the scout procedure with the solution to equation (11).

A.1.2. Proof of assumption 2

If $\hat{\beta}$ minimizes equation (14), then, since $\hat{\beta}_i\hat{\beta}_j \neq 0$, it follows that

$$\frac{\lambda_2}{2}\{\mathrm{sgn}(\hat{\beta}_i) - \mathrm{sgn}(\hat{\beta}_j)\} + \sqrt{(2\lambda_1)}(\hat{\beta}_i - \hat{\beta}_j) = (\mathbf{x}_i^* - \mathbf{x}_j^*)^{\mathrm{T}}(\mathbf{y}^* - \mathbf{X}^*\hat{\beta}), \tag{36}$$

and hence that

$$\sqrt{(2\lambda_1)}|\hat{\beta}_i - \hat{\beta}_j| \leqslant |(\mathbf{x}_i^* - \mathbf{x}_j^*)^{\mathrm{T}}(\mathbf{y}^* - \mathbf{X}^*\hat{\beta})|. \tag{37}$$

Note that

$$\|\mathbf{y}^* - \mathbf{X}^*\hat{\beta}\|^2 \leqslant \|\mathbf{y}^* - \mathbf{X}^*\hat{\beta}\|^2 + \lambda_2\|\hat{\beta}\|^1 + \sqrt{(2\lambda_1)}\|\hat{\beta}\|^2 \leqslant \|\mathbf{y}^*\|^2 = 2\|\mathbf{y}\|^2. \tag{38}$$

Therefore,

$$|\hat{\beta}_i - \hat{\beta}_j| \leqslant \sqrt{\left(\frac{1}{2\lambda_1}\right)}\|\mathbf{x}_i^* - \mathbf{x}_j^*\|\|\mathbf{y}\|\sqrt{2}. \tag{39}$$

Now,

$$\|\mathbf{x}_i^* - \mathbf{x}_j^*\|^2 = \tfrac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2 + \tfrac{1}{2}\|(\tilde{\mathbf{D}}\mathbf{V}^{\mathrm{T}})_i - (\tilde{\mathbf{D}}\mathbf{V}^{\mathrm{T}})_j\|^2.$$

Since we assumed that the features are standardized, it follows that

$$\|\mathbf{x}_i^* - \mathbf{x}_j^*\|^2 = 1 - \rho + \tfrac{1}{2}\|(\tilde{\mathbf{D}}\mathbf{V}^{\mathrm{T}})_i - (\tilde{\mathbf{D}}\mathbf{V}^{\mathrm{T}})_j\|^2$$

where ρ is the correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$. It also is easy to see that $\|(\tilde{\mathbf{D}}\mathbf{V}^{\mathrm{T}})_i - (\tilde{\mathbf{D}}\mathbf{V}^{\mathrm{T}})_j\|^2 \leqslant 1 - \rho$. Therefore, it follows that

$$|\hat{\beta}_i - \hat{\beta}_j| \leqslant \sqrt{\left\{\frac{2(1-\rho)}{\lambda_1}\right\}}\|\mathbf{y}\|. \tag{40}$$

A.1.3. Proof of assumption 3

Consider the latent variable model that is given in Section 2.6; note that, under this model,

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{41}$$

where $\beta = (1/d_1)\mathbf{v}_1$. In addition,

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = d_1^2\mathbf{v}_1\mathbf{v}_1^{\mathrm{T}} + d_2^2\mathbf{v}_2\mathbf{v}_2^{\mathrm{T}} = \sum_{j=1}^{p} d_j^2\mathbf{v}_j\mathbf{v}_j^{\mathrm{T}} \tag{42}$$

where $d_3 = \ldots = d_p = 0$ and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ are orthonormal. We consider two options for the regression of $\mathbf{y}$ onto $\mathbf{X}$: ridge regression and Scout$(2, \cdot)$. Let $\hat{\beta}^{\mathrm{rr}}$ and $\hat{\beta}^{\mathrm{sc}}$ denote the resulting estimates, and let λ^{rr} and λ^{sc} be the tuning parameters of the two methods respectively.

$$\begin{aligned}\hat{\beta}^{\mathrm{rr}} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda^{\mathrm{rr}}\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \\ &= \left(\sum_{j=1}^{p} \frac{1}{d_j^2 + \lambda^{\mathrm{rr}}}\mathbf{v}_j\mathbf{v}_j^{\mathrm{T}}\right)(d_1\mathbf{v}_1\mathbf{u}_1^{\mathrm{T}} + d_2\mathbf{v}_2\mathbf{u}_2^{\mathrm{T}})(\mathbf{u}_1 + \varepsilon) \\ &= \frac{d_1}{d_1^2 + \lambda^{\mathrm{rr}}}\mathbf{v}_1 + \left(\frac{d_1}{d_1^2 + \lambda^{\mathrm{rr}}}\mathbf{v}_1\mathbf{u}_1^{\mathrm{T}} + \frac{d_2}{d_2^2 + \lambda^{\mathrm{rr}}}\mathbf{v}_2\mathbf{u}_2^{\mathrm{T}}\right)\varepsilon. \end{aligned} \tag{43}$$

Similarly, the solution to Scout$(2, \cdot)$ is

$$\begin{aligned}\hat{\beta}^{\mathrm{sc}} &= \left\{\sum_{j=1}^{p} \frac{2}{d_j^2 + \sqrt{(d_j^4 + 8\lambda^{\mathrm{sc}})}}\mathbf{v}_j\mathbf{v}_j^{\mathrm{T}}\right\}(d_1\mathbf{v}_1\mathbf{u}_1^{\mathrm{T}} + d_2\mathbf{v}_2\mathbf{u}_2^{\mathrm{T}})(\mathbf{u}_1 + \varepsilon) \\ &= \frac{2d_1}{d_1^2 + \sqrt{(d_1^4 + 8\lambda^{\mathrm{sc}})}}\mathbf{v}_1 + \left\{\frac{2d_1}{d_1^2 + \sqrt{(d_1^4 + 8\lambda^{\mathrm{sc}})}}\mathbf{v}_1\mathbf{u}_1^{\mathrm{T}} + \frac{2d_2}{d_2^2 + \sqrt{(d_2^4 + 8\lambda^{\mathrm{sc}})}}\mathbf{v}_2\mathbf{u}_2^{\mathrm{T}}\right\}\varepsilon. \end{aligned} \tag{44}$$

The biases of $\hat{\beta}^{\mathrm{rr}}$ and $\hat{\beta}^{\mathrm{sc}}$ are

$$\begin{aligned}E(\hat{\beta}^{\mathrm{rr}} - \beta) &= \left(\frac{d_1}{d_1^2 + \lambda^{\mathrm{rr}}} - \frac{1}{d_1}\right)\mathbf{v}_1, \\ E(\hat{\beta}^{\mathrm{sc}} - \beta) &= \left\{\frac{2d_1}{d_1^2 + \sqrt{(d_1^4 + 8\lambda^{\mathrm{sc}})}} - \frac{1}{d_1}\right\}\mathbf{v}_1 \end{aligned} \tag{45}$$

and the variances are

$$\begin{aligned}\mathrm{var}(\hat{\beta}^{\mathrm{rr}}) &= \left(\frac{d_1}{d_1^2 + \lambda^{\mathrm{rr}}}\right)^2\mathbf{v}_1\mathbf{v}_1^{\mathrm{T}}\sigma^2 + \left(\frac{d_2}{d_2^2 + \lambda^{\mathrm{rr}}}\right)^2\mathbf{v}_2\mathbf{v}_2^{\mathrm{T}}\sigma^2, \\ \mathrm{var}(\hat{\beta}^{\mathrm{sc}}) &= \left\{\frac{2d_1}{d_1^2 + \sqrt{(d_1^4 + 8\lambda^{\mathrm{sc}})}}\right\}^2\mathbf{v}_1\mathbf{v}_1^{\mathrm{T}}\sigma^2 + \left\{\frac{2d_2}{d_2^2 + \sqrt{(d_2^4 + 8\lambda^{\mathrm{sc}})}}\right\}^2\mathbf{v}_2\mathbf{v}_2^{\mathrm{T}}\sigma^2. \end{aligned} \tag{46}$$

The following relationship between λ^{rr} and λ^{sc} results in equal biases:

$$\lambda^{\mathrm{rr}} = \frac{-d_1^2 + \sqrt{(d_1^4 + 8\lambda^{\mathrm{sc}})}}{2}. \tag{47}$$

From now on, we assume that equation (47) holds. Then, if $d_1 > d_2$, it follows that $\mathrm{var}(\hat{\beta}^{\mathrm{sc}}) < \mathrm{var}(\hat{\beta}^{\mathrm{rr}})$. In other words, if the portion of $\mathbf{X}$ that is correlated with $\mathbf{y}$ has a stronger signal than the portion that is orthogonal to $\mathbf{y}$, then (for a given amount of bias) Scout$(2, \cdot)$ will have lower variance than ridge regression.

A.2. Feature selection for scout linear discriminant analysis

The method that we propose in Section 4.1 can be easily modified to perform built-in feature selection. Using the notation in Section 4.1, we observe that

$$\hat{\mu}_k = \arg\min_{\mu_k}\left\{\sum_{i:C(i)=k}(X_i - \mu_k)^{\mathrm{T}}\hat{\mathbf{\Sigma}}_{\mathrm{wc},\lambda}^{-1}(X_i - \mu_k)\right\} \tag{48}$$

and so we replace $\hat{\mu}_k$ in equation (24) with

$$\hat{\mu}_k^{\lambda,\rho} = \arg\min_{\mu_k}\left\{\sum_{i:C(i)=k}(X_i - \mu_k)^{\mathrm{T}}\hat{\mathbf{\Sigma}}_{\mathrm{wc},\lambda}^{-1}(X_i - \mu_k) + \rho\|\hat{\mathbf{\Sigma}}_{\mathrm{wc},\lambda}^{-1}\mu_k\|^1\right\}. \tag{49}$$

This can be solved via an L_1-regression, and it gives the following classification rule for a test observation X:

$$\delta_k^{\lambda,\rho}(X) = X^{\mathrm{T}}\hat{\mathbf{\Sigma}}_{\mathrm{wc},\lambda}^{-1}\hat{\mu}_k^{\lambda,\rho} - \tfrac{1}{2}(\hat{\mu}_k^{\lambda,\rho})^{\mathrm{T}}\hat{\mathbf{\Sigma}}_{\mathrm{wc},\lambda}^{-1}\hat{\mu}_k^{\lambda,\rho} + \log(\pi_k), \tag{50}$$

and X is assigned to class k' if k' maximizes $\delta_k^{\lambda,\rho}$.

References

Bair, E. and Tibshirani, R. (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLOS Biol.*, **2**, 511–522.

Banerjee, O., El Ghaoui, L. E. and d'Aspremont, A. (2008) Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res.*, **9**, 485–516.

Bickel, P. and Levina, E. (2008) Covariance regularization by thresholding. *Ann. Statist.*, to be published.

Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Dey, D. and Srinivasan, C. (1985) Estimation of a covariance matrix under Stein's loss. *Ann. Statist.*, **13**, 1581–1591.

Frank, I. and Friedman, J. (1993) A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, **35**, 109–148.

Friedman, J. (1989) Regularized discriminant analysis. *J. Am. Statist. Ass.*, **84**, 165–175.

Friedman, J., Hastie, T. and Tibshirani, R. (2007) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

Friedman, J., Hastie, T. and Tibshirani, R. (2008) Regularization paths for generalized linear models via coordinate descent. To be published.

Green, P. J. (1984) Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J. R. Statist. Soc.* B, **46**, 149–192.

Guo, Y., Hastie, T. and Tibshirani, R. (2007) Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, **8**, 86–100.

Haff, L. (1979) Estimation of the inverse covariance matrix: random mixtures of the inverse Wishart matrix and the identity. *Ann. Statist.*, **7**, 1264–1276.

Hinton, G., Osindero, S. and Teh, Y. (2006) A fast learning algorithm for deep belief nets. *Neur. Computn*, **18**, 1527–1553.

Hoerl, A. E. and Kennard, R. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hummel, M., Bentink, S., Berger, H., Klappwe, W., Wessendorf, S., Barth, F. T. E., Bernd, H.-W., Cogliatti, S. B., Dierlamm, J., Feller, A. C., Hansmann, M. L., Haralambieva, E., Harder, L., Hasenclever, D., Kuhn, M., Lenze, D., Lichter, P., Martin-Subero, J. I., Moller, P., Muller-Hermelink, H.-K., Ott, G., Parwaresch, R. M., Pott, C., Rosenwald, A., Rosolowski, M., Schwaenen, C., Sturzenhofecker, B., Szczepanowski, M., Trautmann, H., Wacker, H.-H., Spang, R., Loefler, M., Trumper, L., Stein, H. and Siebert, R. (2006) A biological definition of Burkitt's lymphoma from transcriptional and genomic profiling. *New Engl. J. Med.*, **354**, 2419–2430.

James, W. and Stein, C. (1961) Estimation with quadratic loss. In *Proc. 4th Berkeley Symp. Mathematics and Statistical Probability*, vol. 1, pp. 361–379. Berkeley: University of California Press.

Kalbfleisch, J. and Prentice, R. (1980) *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Liang, F., Mukherjee, S. and West, M. (2007) The use of unlabeled data in predictive modeling. *Statist. Sci.*, **22**, 189–205.

Mardia, K., Kent, J. and Bibby, J. (1979) *Multivariate Analysis*. London: Academic Press.

McLachlan, G. J. (1992) *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.

Meinshausen, N. and Bühlmann, P. (2006) High dimensional graphs and variable selection with the lasso. *Ann. Statist.*, **34**, 1436–1462.

Monti, S., Savage, K. J., Kutok, J. L., Feuerhake, F., Kurtin, P., Mihm, M., Wu, B., Pasqualucci, L., Neuberg, D., Aguiar, R. C. T., Dal Cin, P., Ladd, C., Pinkus, G. S., Salles, G., Harris, N. L., Dalla-Favera, R., Habermann, T. M., Aster, J. C., Golub, T. R. and Shipp, M. A. (2005) Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood*, **105**, 1851–1861.

O'Neill, T. (1978) Normal discrimination with unclassified observations. *J. Am. Statist. Ass.*, **73**, 821–826.

Park, M. Y. and Hastie, T. (2007) L_1-regularization path algorithm for generalized linear models. *J. R. Statist. Soc.* B, **69**, 659–677.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., Poggio, T., Gerald, W., Loda, M., Lander, E. and Golub, T. (2001) Multiclass cancer diagnosis using tumor gene expression signature. *Proc. Natn. Acad. Sci. USA*, **98**, 15149–15154.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B. and Staudt, L. M. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. *New Engl. J. Med.*, **346**, 1937–1947.

Rothman, A., Levina, E. and Zhu, J. (2008) Sparse permutation invariant covariance estimation. *Electr. J. Statist.*, **2**, 494–515.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuberg, D. S., Lander, E. S., Aster, J. C. and Golub, T. R. (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.*, **8**, 68–74.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natn. Acad. Sci. USA*, **99**, 6567–6572.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2003) Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statist. Sci.*, **18**, 104–117.
Tusher, V. G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natn. Acad. Sci. USA*, **98**, 5116–5121.
Zhao, P. and Yu, B. (2006) On model selection consistency of lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.
Zhu, J. and Hastie, T. (2004) Classification of gene microarrays by penalized logistic regression. *Biostatistics*, **5**, 427–443.
Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B, **67**, 301–320.