Covariance-regularized regression and classification for high-dimensional problems

Daniela M. Witten†

Department of Statistics, Stanford University, 390 Serra Mall, Stanford CA 94305, USA. E-mail: dwitten@stanford.edu

Robert Tibshirani

Departments of Statistics and Health Research & Policy, Stanford University, 390 Serra Mall, Stanford CA 94305, USA. E-mail: tibs@stat.stanford.edu

Summary. In recent years, many methods have been developed for regression in high-dimensional settings. We propose covariance-regularized regression, a family of methods that use a shrunken estimate of the inverse covariance matrix of the features in order to achieve superior prediction. An estimate of the inverse covariance matrix is obtained by maximizing its log likelihood, under a multivariate normal model, subject to a constraint on its elements; this estimate 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 analyze gene expression data sets with multiple class and survival outcomes.

Keywords: regression, classification, $n \ll p$, covariance regularization

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 used to estimate the regression coefficients, using penalties that are based on the magnitudes of the coefficients:

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

(Here, the notation $||\beta||^s$ is used to indicate $\sum_{i=1}^{p} |\beta_i|^s$.) 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 & Kennard 1970), and $p_1 = 1$ gives the lasso (Tibshirani 1996). More generally, for $\lambda_2 = 0$ and $p_1 \geq 0$, the above equation defines the bridge estimators (Frank & Friedman 1993). Equation 1 defines the elastic net (up to a scaling) in the case that $p_1 = 1$ and $p_2 = 2$ (Zou & 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 inverse covariance matrix under a multivariate normal model.

†Corresponding author.

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 zero correspond to pairs of variables that have no 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 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 $\hat{\beta}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$. 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 its log likelihood under a multivariate normal model, subject to a constraint on its elements. 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 in order to obtain regularized regression coefficients. We call the class of regression methods 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 relative 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 a number of gene expression data sets.

2. The Scout Method

2.1. The General Scout Family

Let $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ 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 centered. We can create a matrix $\tilde{\mathbf{X}} = \begin{pmatrix}\mathbf{X} & \mathbf{y}\end{pmatrix}$, which has dimension $n \times (p+1)$. If we assume that $\tilde{\mathbf{X}}$ is generated from the model $N(0, \mathbf{\Sigma})$, then we can find the maximum likelihood estimator of the population inverse covariance matrix $\mathbf{\Sigma}^{-1}$ by maximizing

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

where $\mathbf{S} = \begin{pmatrix}\mathbf{S_{xx}} & \mathbf{S_{xy}} \\ \mathbf{S_{xy}}^T & S_{yy}\end{pmatrix}$ is the empirical covariance matrix of $\tilde{\mathbf{X}}$. Assume for a moment that $\mathbf{S}$ is invertible. Then, the maximum likelihood estimator for $\mathbf{\Sigma}^{-1}$ is $\mathbf{S}^{-1}$ (we use the fact that $\frac{d}{d\mathbf{W}} \log\det \mathbf{W} = \mathbf{W}^{-1}$ for a symmetric positive definite matrix $\mathbf{W}$). Let $\mathbf{\Theta} = \begin{pmatrix}\mathbf{\Theta_{xx}} & \mathbf{\Theta_{xy}} \\ \mathbf{\Theta_{xy}}^T & \Theta_{yy}\end{pmatrix}$ denote a symmetric estimate of $\mathbf{\Sigma}^{-1}$. The problem of regressing $\mathbf{y}$ onto $\mathbf{X}$ is closely related to the problem of estimating $\mathbf{\Sigma}^{-1}$, since the least squares coefficients for the regression equal $-\frac{\mathbf{\Theta_{xy}}}{\Theta_{yy}}$ for $\mathbf{\Theta} = \mathbf{S}^{-1}$ (this follows from the partitioned inverse formula). If $p > n$, then some type of regularization is needed in order 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 in order to achieve superior prediction. The connection between estimation of $\mathbf{\Theta}$ 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 e.g. ridge regression or the lasso, we can obtain shrunken β estimates through maximization of the penalized log likelihood of the inverse covariance matrix $\mathbf{\Sigma}^{-1}$.

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

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

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 $\beta = -\frac{\mathbf{\Theta_{xy}}}{\Theta_{yy}}$. However, recall that if $\tilde{\mathbf{X}} \sim N(0, \mathbf{\Sigma})$, then the ij element of $\mathbf{\Theta}$ gives the correlation of $\tilde{\mathbf{x}}_i$ with $\tilde{\mathbf{x}}_j$, conditional on all of the other variables in $\tilde{\mathbf{X}}$. 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:

The Scout Procedure for General Penalty Functions

1. Compute $\hat{\mathbf{\Theta}}_{\mathbf{xx}}$, which maximizes

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

2. Compute $\hat{\mathbf{\Theta}}$, which maximizes

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

   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.

3. Compute $\hat{\beta}$, defined by $\hat{\beta} = -\frac{\hat{\mathbf{\Theta}}_{\mathbf{xy}}}{\hat{\Theta}_{yy}}$.

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 obtained using this new method. Step 1 of the Scout Procedure involves obtaining shrunken estimates of $(\mathbf{\Sigma_{xx}})^{-1}$ in order 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 obtained in Step 1. Thus, we obtain regularized estimates of which predictors are dependent on $\mathbf{y}$, given all of 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

Table 1. Special cases of the scout.

$J_1(\mathbf{\Theta_{xx}})$	$J_2(\mathbf{\Theta})$	*Method*
0	0	Least Squares
$\text{tr}(\mathbf{\Theta_{xx}})$	0	Ridge Regression
$\text{tr}(\mathbf{\Theta_{xx}})$	$\|\|\mathbf{\Theta}\|\|^{\mathbf{1}}$	Elastic Net
0	$\|\|\mathbf{\Theta}\|\|^{\mathbf{1}}$	Lasso
0	$\|\|\mathbf{\Theta}\|\|^{\mathbf{2}}$	Ridge Regression

dependent due only to chance. We are searching, or **scouting**, for variables that truly are correlated with each other, conditional on all of 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. One expects 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 to 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 in order 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 $\hat{\beta}^*$ 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**.

In the rest of the paper, when we discuss properties of the scout, for ease of notation we will 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, we will exclusively be interested in the case that $J_1(\mathbf{\Theta}) = \lambda_1||\mathbf{\Theta}||^{p_1}$ and $J_2(\mathbf{\Theta}) = \frac{\lambda_2}{2}||\mathbf{\Theta}||^{p_2}$, where the norm is taken elementwise over the entries of $\mathbf{\Theta}$, and where $\lambda_1, \lambda_2 \geq 0$. 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 will be as follows:

The Scout Procedure with L_p Penalties

1. Compute $\hat{\mathbf{\Theta}}_{\mathbf{xx}}$, which maximizes

$$\log(\det \mathbf{\Theta_{xx}}) - \text{tr}(\mathbf{S_{xx}\Theta_{xx}}) - \lambda_1||\mathbf{\Theta_{xx}}||^{p_1} \tag{6}$$

2. Compute $\hat{\mathbf{\Theta}}$, which maximizes

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

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. Note that because of this constraint, the penalty really is only being applied to the last row and column of $\hat{\mathbf{\Theta}}$.

3. Compute $\hat{\beta}$, defined by $\hat{\beta} = -\frac{\hat{\mathbf{\Theta}}_{\mathbf{xy}}}{\hat{\Theta}_{yy}}$.

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

(Note that in Step 2, because the top left $p \times p$ submatrix of $\hat{\mathbf{\Theta}}$ is fixed, the penalty on the top left $p \times p$ elements has no effect).

2.3. Simple Example

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

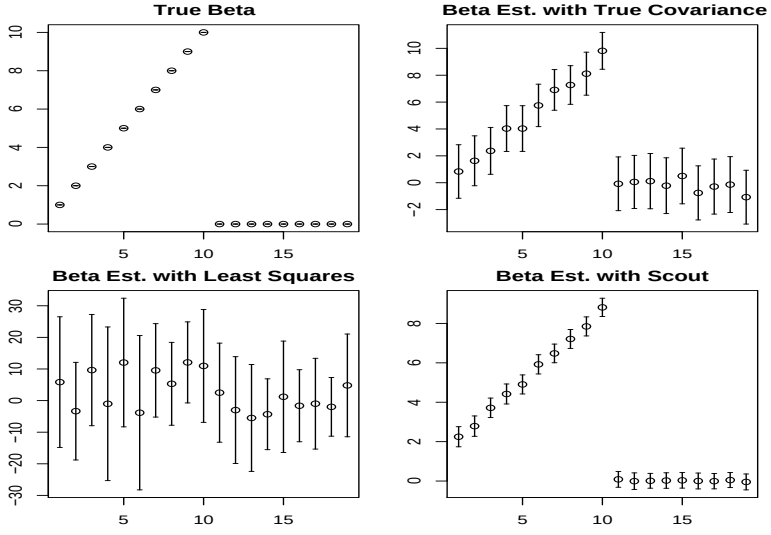

Fig. 1. *Data were generated under a simple model; results shown are averaged over 500 simulations. Standard error bars and coefficient estimates are shown. Clockwise from the top left, panels show the true value of* β, $\mathbf{\Sigma}^{-1}\mathrm{Cov}(\mathbf{X}, \mathbf{y})$, $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, *and* $Scout(1, \cdot)$.

Figure 1 shows the average over 500 simulations of the following four quantities: the true value of β, the least squares regression coefficients, the estimate $\mathbf{\Sigma}^{-1}\mathrm{Cov}(\mathbf{X}, \mathbf{y})$ where $\mathbf{\Sigma}$ is the true covariance matrix of $\mathbf{X}$ in the underlying model, and the $Scout(1, \cdot)$ regression estimate. It is not surprising that least squares performs poorly in this situation, since N is barely larger than p. $Scout(1, \cdot)$ performs extremely well; though

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

2.4. Minimization of the Scout Criteria with L_p Penalties

If $\lambda_1 = 0$, then the minimum 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 $p_1 = 1$, minimization of the first scout criterion has been studied extensively; see e.g. Meinshausen & Bühlmann (2006). The solution can be found via the "graphical lasso", an efficient algorithm given by Banerjee et al. (2008) and Friedman et al. (2008*b*) that involves iteratively regressing one row of the estimated covariance matrix onto the others, subject to an L_1 constraint, in order 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

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

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

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

This equation implies that $\mathbf{\Theta_{xx}}$ and $\mathbf{S_{xx}}$ share the same eigenvectors. Letting θ_i denote the i^{th} eigenvalue of $\mathbf{\Theta_{xx}}$ and letting s_i denote the i^{th} 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 minimize the second scout criterion directly, as there is an easier alternative:

Claim 1. *For $p_2 \in \{1, 2\}$, the solution to Step 3 of the Scout Procedure is equal to the solution to the following, up to scaling by a constant:*

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

where $\hat{\mathbf{\Sigma}}_{\mathbf{xx}}$ is the inverse of the solution to Step 1 of the Scout Procedure.

(The proof of Claim 1 is in Section 8.1.1 in the Appendix.) 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.

In addition, we will show in Section 2.5.1 that if $p_1 = 2$ and $p_2 = 1$, then the scout can be re-written as an elastic net problem with slightly different data; therefore,

Table 2. Minimization of the scout criteria: special cases. The scout criteria can be easily minimized if $\lambda_1 = 0$ or $p_1 \in \{1, 2\}$, and if $\lambda_2 = 0$ or $p_2 \in \{1, 2\}$.

	$\lambda_2 = 0$	$p_2 = 1$	$p_2 = 2$
$\lambda_1 = 0$	Least Squares	L_1 Regression	L_2 Regression
$p_1 = 1$	Graphical Lasso	Graphical Lasso + L_1 Reg.	Graphical Lasso + L_2 Reg.
$p_1 = 2$	Eigenvalue Problem	Elastic Net	Eigenvalue Problem + L_2 Reg.

Table 3. Timing comparisons for minimization of the scout criteria. The numbers of CPU seconds required to run four versions of the scout are shown, for $\lambda_1 = \lambda_2 = 0.2$, $n = 100$, $\mathbf{X}$ dense, and various values of p.

p	$Scout(1, \cdot)$	$Scout(1, 1)$	$Scout(2, \cdot)$	$Scout(2, 1)$
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

fast algorithms for solving the elastic net (Friedman et al. 2008*a*) can be used to solve $Scout(2, 1)$. The methods for minimizing 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 $n = 100$, $\lambda_2 = \lambda_1 = 0.2$, and $\mathbf{X}$ dense. All timings were carried out on a Intel Xeon 2.80 GHz processor. Table 3 shows the number of CPU seconds required for each of these methods for a range of values of p. For all methods, after the scout coefficients have been estimated for a given set of parameter values, estimation for different parameter values is faster due to a warm start (when $p_1 = 1$) or because the eigen decomposition has already been computed (when $p_1 = 2$).

2.5. Properties of the Scout

In this section, for ease of notation, we will consider an equivalent form of the Scout Procedure obtained by replacing $\mathbf{S_{xx}}$ with $\mathbf{X}^T\mathbf{X}$ and $\mathbf{S_{xy}}$ with $\mathbf{X}^T\mathbf{y}$.

2.5.1. Similarities between Scout, Ridge Regression, and the Elastic Net

Let $\mathbf{U}_{n\times p}\mathbf{D}_{p\times p}\mathbf{V}^T_{p\times p}$ denote the singular value decomposition of $\mathbf{X}$ with d_i the i^{th} diagonal element of $\mathbf{D}$ and $d_1 \geq d_2 \geq ... \geq d_r > d_{r+1} = ... = d_p = 0$, where $r = \text{rank}(\mathbf{X}) \leq \min(n, p)$. Consider $Scout(2, p_2)$. As previously discussed, the first step in the Scout Procedure corresponds to finding $\mathbf{\Theta}$ that solves

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

Since $\mathbf{\Theta}$ and $\mathbf{X}^T\mathbf{X}$ therefore share the same eigenvectors, it follows that $\mathbf{\Theta}^{-1} = \mathbf{V}(\mathbf{D}^2+\tilde{\mathbf{D}}^2)\mathbf{V}^T$ where $\tilde{\mathbf{D}}^2$ is a $p\times p$ diagonal matrix with i^{th} 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:

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

Therefore, while 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:

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

In fact, both solutions can be re-written:

$$\begin{aligned}\hat{\beta}_{enet} &= \arg\min_{\beta} \beta^T\mathbf{X}^T\mathbf{X}\beta - 2\beta^T\mathbf{X}^T\mathbf{y} + c||\beta||^2 + \lambda_2||\beta||^1 \\ \hat{\beta}_{scout(2,1)} &= \arg\min_{\beta} \beta^T \mathbf{V}(\hat{\mathbf{D}}^2 + \sqrt{2\lambda_1}\mathbf{I})\mathbf{V}^T\beta - 2\beta^T\mathbf{X}^T\mathbf{y} + \lambda_2||\beta||^1 \\ &= \arg\min_{\beta} \beta^T \mathbf{V}\hat{\mathbf{D}}^2\mathbf{V}^T\beta - 2\beta^T\mathbf{X}^T\mathbf{y} + \sqrt{2\lambda_1}||\beta||^2 + \lambda_2||\beta||^1 \\ &= \arg\min_{\beta} \beta^T \mathbf{X}^{*T}\mathbf{X}^*\beta - 2\beta^T\mathbf{X}^{*T}\mathbf{y}^* + \sqrt{2\lambda_1}||\beta||^2 + \lambda_2||\beta||^1\end{aligned} \tag{15}$$

where $\hat{\mathbf{D}}$ is a diagonal matrix with diagonal elements $\frac{1}{2}(d_i^2 + \sqrt{d_i^4 + 8\lambda_1}) - \sqrt{2\lambda_1}$, $\mathbf{X}^* = \hat{\mathbf{D}}_{(\mathbf{r})}\mathbf{V}_{(\mathbf{r})}{}^T$, $\mathbf{y}^* = \hat{\mathbf{D}}_{(\mathbf{r})}^{-1}\mathbf{D}_{(\mathbf{r})}\mathbf{U}_{(\mathbf{r})}{}^T\mathbf{y}$. $\mathbf{D}_{(\mathbf{r})}$ and $\hat{\mathbf{D}}_{(\mathbf{r})}$ are the $r \times r$ submatrices of $\mathbf{D}$ and $\hat{\mathbf{D}}$ corresponding to non-zero diagonal elements, and $\mathbf{U}_{(\mathbf{r})}$ and $\mathbf{V}_{(\mathbf{r})}$ correspond to the first r columns of $\mathbf{U}$ and $\mathbf{V}$. From Equation 15, it is clear that $Scout(2,1)$ solutions can be obtained using software for the elastic net on data $\mathbf{X}^*$ (which has dimension no greater than the original 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 & Hastie (2005) show 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 $Scout(2,1)$:

Claim 2. *Assume that the predictors are standardized and that* $\mathbf{y}$ *is centered. 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* $\mathrm{sgn}(\hat{\beta}_i) = \mathrm{sgn}(\hat{\beta}_j)$*, then the following holds:*

$$|\hat{\beta}_i - \hat{\beta}_j| \leq \sqrt{\frac{2(1-\rho)}{\lambda_1}}||\mathbf{y}|| \tag{16}$$

The proof of Claim 2 is in Section 8.1.2 in the Appendix. Similar results hold for $Scout(2,\cdot)$ and $Scout(2,2)$, without the assumptions about the signs of $\hat{\beta}_i$ and $\hat{\beta}_j$.

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 $\frac{1}{\lambda_1 + \mathbf{x}_i^T \mathbf{x}_i}$. (More specifically, if $\lambda_1 \geq |\mathbf{x}_i^T \mathbf{x}_j|$ for all $i \neq j$, then 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 i^{th} component of the $Scout(1, \cdot)$ solution, then $\hat{\beta}_i^* = \frac{\mathbf{x}_i^T \mathbf{y}}{\lambda_1 + 1}$. 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:

$$\hat{\beta}_i^* = \frac{1}{\lambda_1 + 1} \text{sgn}(\mathbf{x}_i^T \mathbf{y}) \max(0, |\mathbf{x}_i^T \mathbf{y}| - \frac{\lambda_2}{2}) \tag{17}$$

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

$$\hat{\beta}^* = (1 + \lambda_1 + \lambda_2)^{-1} \mathbf{X}^T \mathbf{y} \tag{18}$$

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

$$\begin{aligned} \mathbf{X} &= d_1 \mathbf{u}_1 \mathbf{v}_1^T + d_2 \mathbf{u}_2 \mathbf{v}_2^T \\ d_1, d_2 &> 0 \\ \mathbf{y} &= \mathbf{u}_1 + \epsilon \end{aligned} \tag{19}$$

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

Claim 3. *Under this model, if $d_1 > d_2$, then $Scout(2, \cdot)$ results in estimates of the regression coefficients that have lower variance than those obtained via ridge regression.*

A more technical explanation of Claim 3, as well as a proof, are given in Section 8.1.3 of the Appendix. Note that a simple example of the above model would be the case of 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. Claim 3 shows that if the signal associated with the relevant gene pathway is sufficiently large, then $Scout(2, \cdot)$ will provide a benefit over ridge.

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 five simulated examples. The first four simulations are based on those used in the original elastic net paper (Zou & Hastie 2005)

and the original lasso paper (Tibshirani 1996). The fifth is of our own invention. All five simulations are based on the model $\mathbf{y} = \mathbf{X}\beta + \sigma\epsilon$ where $\epsilon \sim N(0,1)$. For each simulation, each data set consists of a small training set, a small validation set (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 using the notation $\cdot/\cdot/\cdot$. The five simulations are as follows:

1. Each data set consists of 20/20/200 observations, 8 predictors with coefficients $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$, and $\sigma = 3$. The pairwise correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ is $0.5^{|i-j|}$.

2. This simulation is as in Simulation 1, except that $\beta_i = 0.85$ for all i.

3. Each data set consists of 100/100/400 observations and 40 predictors. $\beta_i = 0$ for $i \in 1, ..., 10$ and for $i \in 21, ..., 30$; for all other i, $\beta_i = 2$. We also set $\sigma = 15$ and the correlation between all pairs of predictors was 0.5.

4. Each data set consists of 50/50/400 observations and 40 predictors. $\beta_i = 3$ for $i \in 1, ..., 15$ and $\beta_i = 0$ for $i \in 16, ..., 40$, and $\sigma = 15$. The predictors are generated as follows:

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

   Also, $\mathbf{x}_i \sim N(0,1)$ are independent and identically distributed for $i = 16, ..., 40$, and $\epsilon_i^x \sim N(0, 0.01)$ are independent and identically distributed for $i = 1, ..., 15$.

5. Each data set consists of 50/50/400 observations and 50 predictors; $\beta_i = 2$ for $i < 9$ and $\beta_i = 0$ for $i \geq 9$. $\sigma = 2$ and $\text{Cor}(\mathbf{x}_i, \mathbf{x}_j) = .5 \times 1_{i,j \leq 9}$.

For each simulation, 200 data sets were generated, and the median mean squared errors (with standard errors given in parentheses) are given in Table 4. For each simulation, the two methods resulting in lowest median mean squared error are shown in bold. 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, and 5; 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 five models.

3.2. 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 $\text{Cov}(\mathbf{X}, \mathbf{y})$. Step 1 of this procedure does not involve the response $\mathbf{y}$ at all.

Now, consider a situation in which one has 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 $\mathbf{X_1}$

Table 4. Median mean squared error over 200 simulated data sets is shown for each simulation. Standard errors are given in parentheses. For each simulation, the two methods with lowest median mean squared errors are shown in bold. Least squares was not performed for Simulation 5, because $p = n$.

Simulation	*Least Squares*	*Lasso*	*ENet*	*Scout*$(1,1)$	*Scout*$(2,1)$
Sim 1	5.83(0.43)	2.30(0.16)	**1.77(0.20)**	**1.71(0.13)**	1.85(0.14)
Sim 2	5.83(0.43)	2.84(0.10)	1.90(0.10)	**0.89(0.08)**	**1.15(0.10)**
Sim 3	147.14(3.63)	42.03(0.91)	30.79(0.61)	**20.11(0.16)**	**18.22(0.27)**
Sim 4	961.57(42.82)	46.44(2.14)	**20.49(1.97)**	**23.15(1.62)**	23.70(1.89)
Sim 5	NA	1.32(0.06)	0.55(0.02)	**0.27(0.02)**	**0.52(0.04)**

Table 5. Median L_2 distance over 200 simulated data sets is shown for each simulation; details are as in Table 4.

Simulation	*Least Squares*	*Lasso*	*ENet*	*Scout*$(1,1)$	*Scout*$(2,1)$
Sim 1	3.05(0.10)	1.74(0.05)	1.65(0.08)	**1.58(0.05)**	**1.62(0.06)**
Sim 2	3.05(0.10)	1.95(0.02)	1.62(0.03)	**0.90(0.03)**	**1.04(0.04)**
Sim 3	17.03(0.22)	8.91(0.09)	7.70(0.06)	**6.15(0.01)**	**5.83(0.03)**
Sim 4	168.40(5.13)	17.40(0.16)	**3.85(0.13)**	5.19(2.3)	**3.80(0.14)**
Sim 5	NA	1.23(0.04)	1.03(0.03)	**0.62(0.03)**	**0.89(0.02)**

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 is available. Then, one could estimate the inverse covariance matrix in Step 1 of the Scout Procedure using both $\mathbf{X_1}$ and $\mathbf{X_2}$, and perform Step 2 using $\mathrm{Cov}(\mathbf{X_1}, \mathbf{y})$. 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 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 only use $\mathbf{X_1}$ in estimating the inverse covariance matrix. However, one can imagine situations in which one can use $\mathbf{X_2}$ to obtain a more accurate estimate of the inverse covariance matrix.

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:

$$\begin{aligned} x_{ij} &= 2u_i + \epsilon_{ij}, j = 1, ..., 5, i = 1, ..., n \\ x_{ij} &= \epsilon_{ij}, j = 6, ..., 10, i = 1, ..., n \\ y_i &= 8u_i + 4\epsilon_i', i = 1, ..., n \end{aligned} \tag{21}$$

In addition, we let $\epsilon_{ij}, \epsilon_i', u_i \sim N(0,1)$ i.i.d. 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 6.

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

Table 6. Making use of observations w/o response values: Set-up. The training set consists of 12 observations and associated responses. We wish to predict responses on a test set of 200 observations. We have access to 36 observations for which responses are not available and not of interest.

	Sample Size	*Response Description*
Training Set	12	Available
Test Set	200	Unavailable - Must be predicted
Additional Obs.	36	Unavailable - Not of interest

Table 7. Making use of observations w/o response values: Results. Mean squared prediction errors are averaged over 500 simulations; standard errors are shown in parentheses.

Method	*Mean Squared Prediction Error*
$Scout(1,\cdot)$ w/Additional Obs.	25.65 (0.38)
$Scout(1,\cdot)$ w/o Additional Obs.	29.92 (0.62)
ENet	32.38 (1.04)
Lasso	47.24 (3.58)
$\mathbf{\Sigma}^{-1}\text{Cov}(\mathbf{X},\mathbf{y})$	86.66 (2.07)
Least Squares	1104.9 (428.84)
Null Model	79.24 (0.3)

inverse covariance matrix, and using also the observations without response values. All tuning parameter values are chosen by 5-fold cross-validation. The results in Table 7 are the average mean squared prediction errors obtained over 500 simulations. From the table, it is clear that both versions of scout outperform all of the other methods. In addition, using observations that do not have response values does result in a significant improvement.

In this example, twelve labeled observations on ten variables do not suffice to reliably estimate the inverse covariance matrix. The scout can make use of the observations that lack response values in order to improve the estimate of the inverse covariance matrix, thereby yielding superior predictions. It is worth noting that in this exaple, the formula $\hat{\beta} = \mathbf{\Sigma}^{-1}\text{Cov}(\mathbf{X}_1,\mathbf{y})$ (where $\mathbf{\Sigma}$ is the true covariance matrix) yields an average prediction error that is higher than that of the null model. Therefore, it is clear that in this example, shrinkage is necessary to achieve good predictions.

4. Classification via the Scout

In classification problems, linear discriminant analysis (LDA) can be used if $n > p$, but 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 its log likelihood, 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 $k \in 1, ..., K$. 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 $\mathbf{S_{wc}} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:C(i)=k} (X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^T$ denote the estimated within-class covariance matrix (based on the training set) that is used for ordinary LDA. Then, the scout procedure for clasification is as follows:

The Scout Procedure for Classification

1. Compute the shrunken within-class inverse covariance matrix $\hat{\mathbf{\Sigma}}^{-1}_{\mathbf{wc},\lambda}$ as follows:

$$\hat{\mathbf{\Sigma}}^{-1}_{\mathbf{wc},\lambda} = \arg\max_{\mathbf{\Sigma}^{-1}} \{\log\det \mathbf{\Sigma}^{-1} - \mathrm{tr}(\mathbf{S_{wc}}\mathbf{\Sigma}^{-1}) - \lambda||\mathbf{\Sigma}^{-1}||^s\} \tag{22}$$

   where λ is a shrinkage parameter.

2. Classify test set observation X to class k' if $k' = \arg\max_k \delta^{\lambda}_k(X)$, where

$$\delta^{\lambda}_k(X) = X^T \hat{\mathbf{\Sigma}}^{-1}_{\mathbf{wc},\lambda} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\mathbf{\Sigma}}^{-1}_{\mathbf{wc},\lambda} \hat{\mu}_k + \log \pi_k \tag{23}$$

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

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

This classification rule performs quite well on real microarray data (as is shown below), but has the drawback that it results in a classification rule that makes use of all of 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 $\hat{\mathbf{\Sigma}}^{-1}_{\mathbf{wc},\lambda} \hat{\mu}_k$; note (from Equation 23) 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 Section 8.2 of the Appendix. We will refer to the method detailed in Equation 23 as $Scout(s, \cdot)$ because the penalized log likelihood that is maximized in Equation 22 is analogous to the first scout criterion in the regression case. The tuning parameter λ in Equations 22 and 23 can be chosen via cross-validation.

4.2. Ramaswamy Data

We assess the performance of this method on the Ramaswamy microarray data set, which is discussed in detail in Ramaswamy et al. (2002) and explored further in Zhu & 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 $Scout(2, \cdot)$ to nearest shrunken centroids (NSC) (Tibshirani et al. (2002) and Tibshirani et al. (2003)), L_2 penalized multinomial (Zhu & Hastie 2004), the support vector machine (SVM) with one-versus-all classification (Ramaswamy et al. 2002), and regularized discriminant analysis (RDA) (Guo et al. 2007). For each method, tuning parameter values were chosen by cross-validation. In addition to running $Scout(2, \cdot)$ on all 16063 genes, we

Table 8. Cross-validation and test set errors for the following methods are compared on the Ramaswamy Data: $Scout(2,\cdot)$, $Scout(2,\cdot)$ using the genes with highest t-statistics on the training set, regularized discriminant analysis, nearest shrunken centroids, support vector machine using one-versus-all classification, and the L_2 penalized multinomial. With the exception of RDA, all methods were performed after cube roots of the data were taken and the patients were standardized. RDA was run on the standardized patients, without taking cube roots, as this led to much better performance.

Method	*CV Err. (of 144)*	*Test Err. (of 54)*	*No. Genes*
NSC	35	17	5217
L_2 Penalized Mult.	29	15	16063
SVM	33	14	16063
RDA	34	10	16063
$Scout(2,\cdot)$	38	11	16063
$Scout(2,\cdot)$ High T-stat.	21	8	4000

also ran it on only the genes with highest univariate t-statistics (Tibshirani et al. 2002) in the training set. In the latter case, cross-validation was performed in order to select the number of genes to include in the model. (The model with 4000 genes had lowest cross-validation error). Note that the selection of genes with highest t-statistics was performed separately in each training fold during cross-validation.

The results can be seen in Table 8. $Scout(2,\cdot)$ performed on the 4000 genes with highest training set t-statistics had the lowest test error rate out of all of the methods that we considered.

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 datasets 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 $i \in 1, ..., n$, 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 $(x_1^i, ..., x_p^i)$ for individual i. Failure times are $t_1 < t_2 < ... < t_k$; there are d_i failures at time t_i. We wish to estimate the parameter $\beta = (\beta_1, ..., \beta_p)^T$ in the proportional hazards model $\lambda(t|x) = \lambda_o(t)\exp(\sum_j x_j\beta_j)$. We assume that censoring is noninformative. 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 e.g. Kalbfleisch & Prentice (1980)):

$$L(\beta) = \Pi_{r\in D} \frac{\exp(\sum_{j\in D_r} \eta_j)}{(\sum_{j\in R_r} \exp(\eta_j))^{d_r}} \tag{24}$$

In order to fit the proportional hazards model, we must find the β that maximizes the likelihood above. Note that $\frac{\partial l}{\partial \beta} = (\frac{\partial \eta}{\partial \beta})^T \frac{\partial l}{\partial \eta} = \mathbf{X}^T \frac{\partial l}{\partial \eta}$ and $\frac{\partial^2 l}{\partial \beta \beta^T} \approx \mathbf{X}^T \frac{\partial^2 l}{\partial \eta \eta^T} \mathbf{X}$. Let $\mathbf{u} = \frac{\partial l}{\partial \eta}$ and $\mathbf{A} = -\frac{\partial^2 l}{\partial \eta \eta^T}$. 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

$$\mathbf{X}^T \mathbf{A} \mathbf{X} (\beta - \beta_0) = \mathbf{X}^T \mathbf{u} \tag{25}$$

This is equivalent to finding β that minimizes

$$||\mathbf{y}^* - \mathbf{X}^* \beta^*||^2 \tag{26}$$

where $\mathbf{X}^* = \mathbf{A}^{1/2}\mathbf{X}$, $\mathbf{y}^* = \mathbf{A}^{-1/2}\mathbf{u}$, $\beta^* = \beta - \beta_o$ (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 attained at the previous iteration. We propose to solve the above equation 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 $\mathbf{S} = \begin{pmatrix} \mathbf{X}^T\mathbf{A}\mathbf{X} & \mathbf{X}^T\mathbf{u} \\ \mathbf{u}^T\mathbf{X} & \mathbf{u}^T\mathbf{A}^{-1}\mathbf{u} \end{pmatrix}$, the Scout Procedure for survival data is almost identical to the regression case, as follows:

The Scout Procedure for the Cox Model

1. Let $\hat{\mathbf{\Theta}}_{\mathbf{xx}}$ maximize

$$\log\det \mathbf{\Theta_{xx}} - \mathrm{tr}(\mathbf{S_{xx}}\mathbf{\Theta_{xx}}) - \lambda_1 ||\mathbf{\Theta_{xx}}||^{p_1} \tag{27}$$

2. Let $\hat{\mathbf{\Theta}}$ maximize

$$\log\det \mathbf{\Theta} - \mathrm{tr}(\mathbf{S}\mathbf{\Theta}) - \lambda_2 ||\mathbf{\Theta}||^{p_2} \tag{28}$$

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

3. Compute $\hat{\beta} = -\frac{\hat{\mathbf{\Theta}}_{\mathbf{xy}}}{\hat{\Theta}_{yy}}$.

4. Let $\hat{\beta}^* = c\hat{\beta}$, where c is the coefficient of a Cox proportional hazards model fit to 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 will refer to the resulting coefficient estimates as $Scout(p_1, p_2)$.

Table 9. Mean of $2(\log(L)-\log(L_o))$ on Survival Data. L_1 Cox, supervised principal components, $Scout(1,1)$, and $Scout(2,1)$ are compared on the Hummel, Monti, Rosenwald, and Shipp data sets over ten random training/validation/test set splits. The predictor obtained from the training set is fit on the test set using a Cox proportional hazards model, and the median value of $2(\log(L)-\log(L_o))$ over the ten repetitions is reported. For each data set, the two highest mean values of $2(\log(L)-\log(L_o))$ are shown in bold.

Data Set	L_1 *Cox*	*SPC*	$Scout(1,1)$	$Scout(2,1)$
Hummel	2.640	**3.823**	**4.245**	3.293
Monti	1.647	1.231	**2.149**	**2.606**
Rosenwald	**4.129**	3.542	3.987	**4.930**
Shipp	1.903	1.004	**2.807**	**2.627**

5.2. Lymphoma Data

We illustrate the effectiveness of the scout method on survival data 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) ("Rosenwald"), which consists of 240 patients, Shipp et al. (2002) ("Shipp"), which consists of 58 patients, Hummel et al. (2006) ("Hummel"), which consists of 81 patients, and Monti et al. (2005) ("Monti"), 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 & Hastie (2007), the supervised principal components (SPC) method of Bair & 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 $\mathbf{X_{val}}\beta_{train}$ for L_1 Cox and scout, and it was the first supervised principal component for SPC). We tested the resulting models on the test set. The mean value of $2(\log(L)-\log(L_o))$ over ten separate training/test/validation set splits is given in Table 9, where L denotes the likelihood of the Cox proportional hazards model fit on the test set using the predictor obtained from the training set (for L_1 Cox and scout, this was $\mathbf{X_{test}}\beta_{train}$, and for SPC, this was the first supervised principal component), and L_o denotes the likelihood of the null model. From Tables 9 and 10, it is clear that the scout results in predictors that are quite competitive 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) obtained by maximizing the log likelihood of the inverse covariance matrix of the data, rather than by minimizing the sum of squared errors, subject to a penalty. We have shown that three well-known regression methods - ridge, 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

Table 10. Median Number of Genes Used for Survival Data. L_1 Cox, supervised principal components, $Scout(1,1)$ and $Scout(2,1)$ are compared on the Hummel, Monti, Rosenwald, and Shipp data sets over ten random training/validation/test set splits; the median number of genes used in each of the resulting models is reported.

Data Set	L_1 *Cox*	*SPC*	$Scout(1,1)$	$Scout(2,1)$
Hummel	14	33	78	13
Monti	18.5	17	801.5	144.5
Rosenwald	37.5	32	294	85
Shipp	5.5	10	4.5	5

model settings, and we have demonstrated the performance of the resulting methods on a number of gene expression data sets.

A drawback of the scout method is that when $p_1 = 1$ and the number of features is large, then minimizing 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, 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 minimized with $p_1 = 1$.

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 practice in many examples.

7. Acknowledgments

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 multinomial 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.

8. Appendix

8.1. Proofs of Claims

8.1.1. Proof of Claim 1

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

$$\log\det\boldsymbol{\Theta} - \mathrm{tr}(\mathbf{S}\boldsymbol{\Theta}) - \frac{\lambda_2}{2}||\boldsymbol{\Theta}||^1 \tag{29}$$

with $\boldsymbol{\Theta}_{\mathbf{xx}}$, the top left $p \times p$ 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 $(\hat{\boldsymbol{\Theta}}^{-1})_{yy} = S_{yy} + \frac{\lambda_2}{2}$. The subgradient equation for maximization of the remaining portion of the log-likelihood is

$$0 = (\boldsymbol{\Theta}^{-1})_{\mathbf{xy}} - \mathbf{S}_{\mathbf{xy}} - \frac{\lambda_2}{2}\Gamma \tag{30}$$

where $\Gamma_i = 1$ if the i^{th} element of $\mathbf{\Theta_{xy}}$ is positive, $\Gamma_i = -1$ if the i^{th} element of $\mathbf{\Theta_{xy}}$ is negative, and otherwise Γ_i is between -1 and 1.

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

$$0 = 2(\mathbf{\Theta_{xx}})^{-1}\beta - 2\mathbf{S_{xy}} - \lambda_2\Gamma \tag{31}$$

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

$$\beta^T(\mathbf{\Theta_{xx}})^{-1}\beta - 2\mathbf{S_{xy}}\beta + \lambda_2||\beta||^1 \tag{32}$$

Of course, this is Equation 11. It is an L_1-penalized regression of $\mathbf{y}$ onto $\mathbf{X}$, using only the inner products, with $\mathbf{S_{xx}}$ replaced with $(\mathbf{\Theta_{xx}})^{-1}$. In other words, $\hat{\beta}$ that solves Equation 11 is given by $\mathbf{\Theta_{xx}}(\mathbf{\Theta}^{-1})_{\mathbf{xy}}$, where $\mathbf{\Theta}$ solves Step 2 of the Scout Procedure.

Now, the solution to Step 3 of the Scout Procedure is $-\frac{\mathbf{\Theta_{xy}}}{\mathbf{\Theta_{yy}}}$. By the partitioned inverse formula, $\mathbf{\Theta_{xx}}(\mathbf{\Theta}^{-1})_{\mathbf{xy}} + \mathbf{\Theta_{xy}}(\mathbf{\Theta}^{-1})_{yy} = 0$, so $-\frac{\mathbf{\Theta_{xy}}}{\Theta_{yy}} = \frac{\mathbf{\Theta_{xx}}(\mathbf{\Theta}^{-1})_{\mathbf{xy}}}{(\mathbf{\Theta}^{-1})_{yy}\Theta_{yy}} = \frac{\beta}{(\mathbf{\Theta}^{-1})_{yy}\Theta_{yy}}$. In other words, the solution to Step 3 of the Scout Procedure and the solution to Equation 11 differ by a factor of $(\mathbf{\Theta}^{-1})_{yy}\Theta_{yy}$. Since Step 4 of the Scout Procedure involves scaling the solution to Step 3 by a constant, it is clear that one can replace Step 3 of the Scout Procedure with the solution to Equation 11.

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

$$0 = (\mathbf{\Theta}^{-1})_{\mathbf{xy}} - \mathbf{S_{xy}} - \frac{\lambda_2}{2}\mathbf{\Theta_{xy}} \tag{33}$$

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

$$0 = 2(\mathbf{\Theta_{xx}})^{-1}\beta - 2\mathbf{S_{xy}} + 2\lambda_3\beta \tag{34}$$

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

$$\beta^T(\mathbf{\Theta_{xx}})^{-1}\beta - \mathbf{S_{xy}}^T\beta + \lambda_3\beta^T\beta$$

Of course, this is again Equation 11. Therefore, $\hat{\beta}$ that solves Equation 11 is given (up to scaling by a constant) by $\mathbf{\Theta_{xx}}(\mathbf{\Theta}^{-1})_{\mathbf{xy}}$, where $\mathbf{\Theta}$ 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 one can replace Step 3 of the Scout Procedure with the solution to Equation 11.

8.1.2. Proof of Claim 2

Recall that the solution to $Scout(2,1)$ minimizes the following:

$$\beta^T(\mathbf{V}(\mathbf{D}^2 + \tilde{\mathbf{D}}^2)\mathbf{V}^T)\beta - 2\beta^T\mathbf{X}^T\mathbf{y} + \lambda_2||\beta||^1 \tag{35}$$

where $\tilde{\mathbf{D}}^2$ is a $p \times p$ diagonal matrix with i^{th} diagonal entry equal to $\frac{1}{2}(-{d_i}^2 + \sqrt{{d_i}^4 + 8\lambda_1})$ and $\mathbf{X} = \mathbf{UDV}^T$. Equivalently, the solution minimizes

$$\begin{aligned} & \beta^T(\mathbf{V}(\frac{1}{2}\mathbf{D}^2 + \frac{1}{2}\bar{\mathbf{D}}^2 + \sqrt{2\lambda_1})\mathbf{V}^T)\beta - 2\beta^T\mathbf{X}^T\mathbf{y} + \lambda_2||\beta||^1 \\ = \quad & \beta^T(\mathbf{V}(\frac{1}{2}\mathbf{D}^2 + \frac{1}{2}\bar{\mathbf{D}}^2)\mathbf{V}^T)\beta - 2\beta^T\mathbf{X}^T\mathbf{y} + \lambda_2||\beta||^1 + \sqrt{2\lambda_1}||\beta||^2 \end{aligned} \tag{36}$$

where $\bar{\mathbf{D}}^2$ is the diagonal matrix with elements $\sqrt{d_i{}^4 + 8\lambda_1} - \sqrt{8\lambda_1}$, because $\mathbf{V}$ is $p \times p$ orthogonal. It is easy to see that the solution also minimizes the following:

$$||\mathbf{y}^* - \mathbf{X}^*\beta||^2 + \lambda_2||\beta||^1 + \sqrt{2\lambda_1}||\beta||^2 \tag{37}$$

where $\mathbf{X}^* = \frac{1}{\sqrt{2}}\begin{pmatrix}\mathbf{X}\\ \bar{\mathbf{D}}\mathbf{V}^T\end{pmatrix}, \mathbf{y}^* = \begin{pmatrix}\sqrt{2}\mathbf{y}\\ 0\end{pmatrix}$. If $\hat{\beta}$ minimizes Equation 37, and if we assume that $\text{sgn}(\hat{\beta}_i) = \text{sgn}(\hat{\beta}_j)$, then it follows that

$$\sqrt{2\lambda_1}|\hat{\beta}_i - \hat{\beta}_j| = |(\mathbf{x}_i^* - \mathbf{x}_j^*)^T(\mathbf{y}^* - \mathbf{X}^*\hat{\beta})| \tag{38}$$

Note that

$$||\mathbf{y}^* - \mathbf{X}^*\hat{\beta}||^2 \leq ||\mathbf{y}^* - \mathbf{X}^*\hat{\beta}||^2 + \lambda_2||\hat{\beta}||^1 + \sqrt{2\lambda_1}||\hat{\beta}||^2 \leq ||\mathbf{y}^*||^2 = 2||\mathbf{y}||^2 \tag{39}$$

Therefore,

$$|\hat{\beta}_i - \hat{\beta}_j| \leq \sqrt{\frac{1}{2\lambda_1}}||\mathbf{x}_i^* - \mathbf{x}_j^*||||\mathbf{y}||\sqrt{2} \tag{40}$$

Now, $||\mathbf{x}_i^* - \mathbf{x}_j^*||^2 = \frac{1}{2}||\mathbf{x}_i - \mathbf{x}_j||^2 + \frac{1}{2}||(\bar{\mathbf{D}}\mathbf{V}^T)_i - (\bar{\mathbf{D}}\mathbf{V}^T)_j||^2$. Since we assumed that the features are standardized, it follows that $||\mathbf{x}_i^* - \mathbf{x}_j^*||^2 = 1 - \rho + \frac{1}{2}||(\bar{\mathbf{D}}\mathbf{V}^T)_i - (\bar{\mathbf{D}}\mathbf{V}^T)_j||^2$ where ρ is the correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$. It also is easy to see that $||(\bar{\mathbf{D}}\mathbf{V}^T)_i - (\bar{\mathbf{D}}\mathbf{V}^T)_j||^2 \leq 1 - \rho$. Therefore, it follows that

$$|\hat{\beta}_i - \hat{\beta}_j| \leq \sqrt{\frac{2(1-\rho)}{\lambda_1}}||\mathbf{y}|| \tag{41}$$

8.1.3. Proof of Claim 3

Consider data generated under the latent variable model given in Section 2.6. Note that it follows that

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

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

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

Similarly, using the fact that $Scout(2,\cdot)$ results in replacing d_j^2 with $\frac{1}{2}(d_j^2 + \sqrt{d_j^4 + 8\lambda^{sc}})$:

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

It is clear that the signal part of $\hat{\beta}^{rr}$ is $\frac{d_1}{d_1^2+\lambda^{rr}}\mathbf{v}_1$, and that the signal part of $\hat{\beta}^{sc}$ is $\frac{2d_1}{d_1^2+\sqrt{d_1^4+8\lambda^{sc}}}\mathbf{v}_1$. The following relationship between λ^{rr} and λ^{sc} results in signals that are equal:

$$\lambda^{rr} = \frac{-d_1^2 + \sqrt{d_1^4 + 8\lambda^{sc}}}{2} \tag{45}$$

From now on, we assume that Equation 45 holds. It is clear that the noise parts of $\hat{\beta}^{rr}$ and $\hat{\beta}^{sc}$ are $(\frac{d_1}{d_1^2+\lambda^{rr}}\mathbf{v}_1\mathbf{u}_1^T + \frac{d_2}{d_2^2+\lambda^{rr}}\mathbf{v}_2\mathbf{u}_2^T)\epsilon$ and $(\frac{2d_1}{d_1^2+\sqrt{d_1^4+8\lambda^{sc}}}\mathbf{v}_1\mathbf{u}_1^T + \frac{2d_2}{d_2^2+\sqrt{d_2^4+8\lambda^{sc}}}\mathbf{v}_2\mathbf{u}_2^T)\epsilon$. Using Equation 45, we know that $\frac{d_1}{d_1^2+\lambda^{rr}}\mathbf{v}_1\mathbf{u}_1^T = \frac{2d_1}{d_1^2+\sqrt{d_1^4+8\lambda^{sc}}}\mathbf{v}_1\mathbf{u}_1^T$. So it suffices to compare $\frac{d_2}{d_2^2+\lambda^{rr}}$ and $\frac{2d_2}{d_2^2+\sqrt{d_2^4+8\lambda^{sc}}}$. Recall that $d_2 > 0$. So,

$$\begin{aligned}(d_2^2 + \lambda^{rr}) - (\frac{d_2^2 + \sqrt{d_2^4 + 8\lambda^{sc}}}{2}) &= (d_2^2 + \frac{-d_1^2 + \sqrt{d_1^4 + 8\lambda^{sc}}}{2}) - (\frac{d_2^2 + \sqrt{d_2^4 + 8\lambda^{sc}}}{2}) \\ &= \frac{1}{2}\{(\sqrt{d_1^4 + 8\lambda^{sc}} - d_1^2) - (\sqrt{d_2^4 + 8\lambda^{sc}} - d_2^2)\}\end{aligned} \tag{46}$$

If $d_1 > d_2$, then the above quantity is negative; if $d_1 < d_2$, then it is positive. Therefore, the scout solution has a smaller noise term than the ridge solution if and only if $d_1 > d_2$. 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 $Scout(2, \cdot)$ will perform better than ridge, because it will shrink the parts of the inverse covariance matrix that correspond to variables that are uncorrelated with the response.

8.2. Feature Selection for Scout LDA

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

$$\hat{\mu}_k = \arg\min_{\mu_k}\{\sum_{i:C(i)=k}(X_i - \mu_k)^T\hat{\mathbf{\Sigma}}_{\mathbf{wc},\lambda}^{-1}(X_i - \mu_k)\} \tag{47}$$

and so we replace $\hat{\mu}_k$ in Equation 23 with

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

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

$$\delta_k^{\lambda,\rho}(X) = X^T\hat{\mathbf{\Sigma}}_{\mathbf{wc},\lambda}^{-1}\hat{\mu}_k^{\lambda,\rho} - \frac{1}{2}(\hat{\mu}_k^{\lambda,\rho})^T\hat{\mathbf{\Sigma}}_{\mathbf{wc},\lambda}^{-1}\hat{\mu}_k^{\lambda,\rho} + \log\pi_k \tag{49}$$

References

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

Banerjee, O., El Ghaoui, L. E. & d'Aspremont, A. (2008), 'Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data', *Journal of Machine Learning Research* .

Frank, I. & Friedman, J. (1993), 'A statistical view of some chemometrics regression tools (with discussion)', *Technometrics* **35**(2), 109–148.

Friedman, J. (1989), 'Regularized discriminant analysis', *Journal of the American Statistical Association* **84**, 165–175.

Friedman, J., Hastie, T. & Tibshirani, R. (2008*a*), 'Regularization paths for generalized linear models via coordinate descent', *In preparation* .

Friedman, J., Hastie, T. & Tibshirani, R. (2008*b*), 'Sparse inverse covariance estimation with the graphical lasso', *Biostatistics* .

Green, P. J. (1984), 'Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives', *Journal of the Royal Statistical Society, Series B* **46**, 149–192.

Guo, Y., Hastie, T. & Tibshirani, R. (2007), 'Regularized linear discriminant analysis and its application in microarrays', *Biostatistics* **8**, 86–100.

Hoerl, A. E. & 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. & Siebert, R. (2006), 'A biological definition of Burkitt's lymphoma from transcriptional and genomic profiling', *New England Journal of Medicine* **354**, 2419–2430.

Kalbfleisch, J. & Prentice, R. (1980), *The statistical analysis of failure time data*, Wiley, New York.

Meinshausen, N. & Bühlmann, P. (2006), 'High dimensional graphs and variable selection with the lasso', *Annals of Statistics* **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. & 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.

Park, M. Y. & Hastie, T. (2007), 'An L_1 regularization path algorithm for generalized linear models', *Journal of the Royal Statistical Society Series B* **69**(4), 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. & Golub, T. (2002), 'Multiclass cancer diagnosis using tumor gene expression signature', *PNAS* **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. & Staudt, L. M. (2002), 'The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma', *The New England Journal of Medicine* **346**, 1937–1947.

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. & Golub, T. R. (2002), 'Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning', *Nature Medicine* **8**, 68–74.

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

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2002), 'Diagnosis of multiple cancer types by shrunken centroids of gene expression', *Proc. Natl. Acad. Sci.* **99**, 6567–6572.

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2003), 'Class prediction by nearest shrunken centroids, with applications to DNA microarrays', *Statistical Science* pp. 104–117.

Zhu, J. & Hastie, T. (2004), 'Classification of gene microarrays by penalized logistic regression', *Biostatistics* **5**(2), 427–443.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *J. Royal. Stat. Soc. B.* **67**, 301–320.