Published in final edited form as:

Biometrics. 2015 December ; 71(4): 1185–1194. doi:10.1111/biom.12359.

A Permutation Approach for Selecting the Penalty Parameter in Penalized Model Selection

Jeremy A Sabourin^{1,2}, William Valdar^{1,4}, and Andrew B Nobel^{3,4,5,*}

¹Department of Genetics, University of North Carolina at Chapel Hill, North Carolina

²Genometrics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD

³Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, North Carolina

⁴Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, North Carolina

⁵Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina

Summary

We describe a simple, computationally efficient, permutation-based procedure for selecting the penalty parameter in LASSO penalized regression. The procedure, permutation selection, is intended for applications where variable selection is the primary focus, and can be applied in a variety of structural settings, including that of generalized linear models. We briefly discuss connections between permutation selection and existing theory for the LASSO. In addition, we present a simulation study and an analysis of real biomedical data sets in which permutation selection is compared with selection based on the following: cross-validation (CV), the Bayesian information criterion (BIC), Scaled Sparse Linear Regression, and a selection method based on recently developed testing procedures for the LASSO.

Keywords

LASSO; Penalized regression; Variable selection

1 Introduction

The analysis of high dimensional data, for which the number of measured predictors is large and can exceed the number of samples, is a common feature of many statistical applications. Whether samples are accompanied by a real or categorical response, analysis of high dimensional data typically includes model fitting with the aim of doing prediction, or

*nobel@email.unc.edu.

Supplementary Materials

Sample code and application guidelines for permutation selection are available in Web Appendix D, which along with Web Appendices A, B and C (as referenced in Sections 3, 4.5 and 5 respectively), are available with this paper at the Biometrics website on Wiley Online Library.

variable selection, or both. The goal of prediction is to derive a rule capable of accurately predicting the response of a new, unlabeled sample. The goal of variable selection is to select a (small) subset of the measured predictors whose individual or coordinated activity is significantly associated with the response. In either case, it is common to act as if the observed data arose from an underlying model that is sparse, in the sense that only a small subset of the predictors are related to the response. Whether this sparsity is assumed, or merely viewed as desirable, analysis of high dimensional data is often carried out using methods that restrict the inclusion of predictors via penalization. Popular penalization methods include the LASSO (Tibshirani, 1996) and SCAD (Fan and Li, 2001), and variations of these methods. In what follows, we focus our attention on the LASSO.

The LASSO and its variants require specification of a penalty (or tuning) parameter, λ , that controls the tradeoff between model fit and model size. Specification of this parameter is necessary to fully determine the selected model (varying λ yields different coefficient estimates), so selecting a suitable value of the penalty parameter is an important part of model fitting. At present, the most widely used procedures for selecting the penalty parameter in the LASSO are cross validation (CV) and the Bayesian Information Criterion (BIC). Cross validation, which is based on out-of-sample prediction, is a natural choice when the goal of model fitting is prediction. Results of (Shao, 1993) show that CV is, in general, inconsistent for model selection. Nonetheless, there is some support for the use of CV with LASSO (see, for example, Bühlmann's discussion of Tibshirani, 2011) and modified CV procedures have been shown to be model selection consistent (see, for example Feng and Yu, 2013). BIC, though designed primarily for variable selection, is based in part on how well the selected model fits the data, making it a reasonable choice for either prediction or variable selection.

In recent work, Lockhart et al. (2014) proposed a covariance test that yields a p-value for each change in the LASSO path that results in the inclusion of a variable. For simple linear models, the test requires that the error variance be known or estimated from the data, limiting its general use to data for which the number of samples n is larger than the number of predictors p . Estimation of the error variance in high dimensional settings, and application of the test to selection of the LASSO penalty parameter are currently under investigation by these authors.

For linear regression, Ye and Zhang (2010) studied the rate minimaxity under ℓ_q risk of the LASSO using universal and related minimax penalties. These penalties are independent of the data, but they depend in a direct way on the variance of the noise, estimation of which can be difficult in high dimensional problems. Belloni et al. (2011) and Sun and Zhang (2012) addressed estimation of the noise variance by studying the solutions of optimization problems related to, but different from, the ℓ_1 penalized least squares that defines the ordinary LASSO. The resulting methods, known respectively as the square-root LASSO and scaled LASSO, have desirable theoretical properties, including near oracle rates of convergence for prediction.

In this paper we investigate a permutation-based procedure ("permutation selection") that selects a penalty parameter for the LASSO when the primary goal of model fitting is

variable selection. Permutation selection was introduced in a more limited context in Valdar et al. (2012) and Sabourin et al. (2015), based on an earlier suggestion of Ayers and Cordell (2010). The procedure is simple and produces a model whose meaning is readily interpretable: for multiple (random) permutations of the response, we identify the minimal penalty required to remove all predictors from the LASSO model, and then output the median of these penalties. Permutation of the response provides a baseline under which there are no true associations with the predictors. Applying a comparable level of penalization to the true (un-permuted) response ensures that the variables included in the model have a joint relationship with the response that is stronger than joint relationships arising by chance.

The permutation selection procedure applies to standard linear regression as well as generalized linear models. Its principal assumption is that individual samples are exchangeable. The method is fast and straightforward to compute; in particular, its computation time is not affected by the correlation structure of the data. On simulated and real data (see below) the performance of permutation selection is comparable to that of competing methods. Analysis of a closely related procedure (see Theorem 1 below) provides theoretical support for permutation selection and establishes a connection with universal penalties for the LASSO.

The remainder of the paper is organized as follows. In the next section, we outline the general data setting, and describe the permutation selection procedure. Section 3 is devoted to theoretical analysis. Section 4 contains a simulation study in which permutation selection is compared to CV, BIC, and two newer selection procedures. In Section 5, we examine the application of these same selection methods to two real data sets. After summarizing the simulation and real data results in Section 6, we conclude with a brief discussion in Section 7.

2 Permutation Selection of the Penalty Parameter

2.1 Data Setting and Model

We assume that the available data is in the form of an $n \times p$ data matrix $\mathbf{X}$ and an $n \times 1$ response vector $\mathbf{y}$. The rows $\mathbf{x}_1, \dots, \mathbf{x}_n$ of $\mathbf{X}$ correspond to samples; the columns $x_{\cdot 1}, \dots, x_{\cdot p}$ of $\mathbf{X}$ correspond to measured variables of interest, and are assumed to be standardized. We describe permutation selection in a setting where a generalized linear model (GLM) is augmented by use of a 1-norm penalty to derive a sparse model from the available data.

A GLM consists of a random component associated with the response, a systematic component equal to a linear function of the data, and a link function connecting the random and systematic components. Let $\mathbf{y} = y_1, \dots, y_n$ be independent observations of a response variable $\mathbf{Y}$ whose distribution belongs to an exponential family. Here we focus on the Gaussian for continuous responses and on the binomial for binary, case/control type responses. Let $\boldsymbol{\mu} = E(\mathbf{Y})$ and let $g(\cdot)$ be a known link function. The response and data matrix are related by the GLM

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}^*. \quad (1)$$

The LASSO procedure (Tibshirani, 1996) fits a sequence, or path, of models parameterized by a non-negative number λ that balances the overall fit of the model and its complexity. Let

$\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$ be the 1-norm of $\boldsymbol{\beta}$, and let

$$\ell(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) = \sum_{i=1}^n \log L(\boldsymbol{\beta}; y_i, \mathbf{x}_i) \quad (2)$$

be the log-likelihood of $\boldsymbol{\beta}$ given $\mathbf{X}$ and $\mathbf{y}$, where $L(\cdot)$ is the likelihood function of a single observation under the GLM (1). For a specified value of $\lambda \geq 0$ the LASSO procedure identifies the (unique) model defined by the coefficient vector

$$\hat{\boldsymbol{\beta}}(\lambda; \mathbf{y}, \mathbf{X}) = \arg \max_{\boldsymbol{\beta}} \{ \ell(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) - \lambda \|\boldsymbol{\beta}\|_1 \}. \quad (3)$$

For each $\lambda \geq 0$ let

$$S_0(\lambda) = \{j: \hat{\beta}_j(\lambda; \mathbf{y}, \mathbf{X}) \neq 0\} \quad (4)$$

be the set of variables included in the model associated with the coefficient vector $\hat{\boldsymbol{\beta}}(\lambda; \mathbf{y}, \mathbf{X})$. As noted in the introduction, the 1-norm penalty ensures that when λ is large the coefficient vector $\hat{\boldsymbol{\beta}}(\lambda; \mathbf{y}, \mathbf{X})$ will be sparse, or equivalently, the set $S_0(\lambda)$ of selected variables will be small. It is this ability of the 1-norm penalty to induce sparsity that makes LASSO and related procedures attractive in high dimensional settings.

2.2 Permutation Selection Procedure

For any permutation π of $[n] := \{1, \dots, n\}$ let $\mathbf{y}_\pi = (y_{\pi(1)}, \dots, y_{\pi(n)})^T$ be a corresponding, permuted version of the response $\mathbf{y}$. Let $\pi_1, \dots, \pi_N$ be obtained by sampling uniformly at random from the set of permutations on $[n]$. For each $1 \leq l \leq N$ and each $\lambda \geq 0$ let

$$S_l(\lambda) = \{j: \hat{\beta}_j(\lambda; \mathbf{y}_{\pi_l}, \mathbf{X}) \neq 0\}$$

be the set of variables selected by the LASSO procedure (3) with response $\mathbf{y}_{\pi_l}$ and penalty parameter λ . In particular, let

$$\lambda_0(\mathbf{y}_{\pi_l}) = \min \{ \lambda: |S_l(\lambda)| = 0 \} \quad (5)$$

be the smallest value of the penalty parameter such that no variables are selected.

Under permutation of the response, $\mathbf{y}_{\pi_l}$ and the measured predictors in $\mathbf{X}$ have on average no systematic relation. This lack of relation most naturally corresponds to a fitted model in which all variables are excluded, and the quantity $\lambda_0(\mathbf{y}_{\pi_l})$ is the smallest amount of

penalization required to ensure this correspondence is maintained. Permutation selection outputs the median of these observed minimal penalties, namely,

$$\hat{\lambda}_{perm} = \text{median} \left(\lambda_0(y_{\pi_1}), \dots, \lambda_0(y_{\pi_N}) \right). \quad (6)$$

As can be seen from its definition, $\hat{\lambda}_{perm}$ is intended to address the goal of variable selection; its predictive performance is therefore only considered briefly in the real data analyses below.

Variability of $\hat{\lambda}_{perm}$ arising from the use of random permutations, can lead to variability in the selected model $S_0(\lambda_{perm})$. Ayers and Cordell (2010) proposed use of a penalty equal to the *maximum* of $\lambda_0(y_{\pi_l})$ over 25 permutations of the response. They showed, empirically, that this penalty controls the family-wise error rate of falsely including a predictor unrelated to the response. Nonetheless, estimation based on a maximum (rather than an average or median) implies high estimator variance and high variability of the selected model. This variability could be reduced by increasing the number of permutations, but doing so would change the implied family wise error rate (for instance, choosing the maximum $\lambda_0(y_{\pi_l})$ from all $n!$ permutations would ensure no inclusions on the real data). By contrast, when estimating $\hat{\lambda}_{perm}$, more permutations reduce variability without changing meaning. In our experience with simulated and real data, $N = 100$ permutations is sufficient to control variability and provide reasonable values of $\hat{\lambda}_{perm}$; in some cases, a lower N can suffice (*e.g.*, in Valdar et al. (2012) and Sabourin et al. (2015) the median of $N = 20$ permutations was enough to select a stable model $S_0(\hat{\lambda}_{perm})$ in the setting of fine-mapping a GWAS locus).

3 Some Connections with Universal Thresholds

In this section we investigate, in the setting of linear regression, a connection between permutation selection and existing theory on penalty selection for the LASSO. In particular, we establish the limiting behavior of a parametric variant of permutation selection in which independent permutations of the response y are replaced by independent, uniformly distributed vectors with norm $\|y\|$. Our analysis, which assumes that the design matrix itself is random, is intended to be illustrative; a more detailed theoretical analysis is left to future work.

Let $S^{n-1} = \{x \in \mathbb{R}^n: \|x\| = 1\}$ denote the unit sphere in $\mathbb{R}^n$. In the remainder of this section we assume that the design matrix $\mathbf{X} = \{X_{ij}\}$ is random and equal to $n^{1/2} \mathbf{V}$, where $\mathbf{V}$ is an $n \times p$ matrix whose columns $\mathbf{v}_1, \dots, \mathbf{v}_p$ are independent and uniformly distributed on S^{n-1} . (Equivalently, one may regard $\mathbf{X}$ as being generated by rescaling the columns of a sample matrix with independent $\mathcal{N}_p(0, \mathbf{I})$ rows.) We assume that the design matrix $\mathbf{X}$ and response y are related by the standard linear regression model

$$y = \mathbf{X}\beta^* + \varepsilon, \quad (7)$$

where $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ is a vector of independent $\mathcal{N}(0, \sigma^2)$ random variables that is independent of $\mathbf{X}$. Under the linear model (7), the LASSO coefficient vector is defined by

$$\hat{\beta}(\lambda; \mathbf{y}, \mathbf{X}) = \arg \min_{\beta} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_1 \right\}. \quad (8)$$

Note that each of the permuted responses $\mathbf{y}_\pi$ lies on the scaled unit sphere $\|\mathbf{y}\| S^{n-1}$. Although they are conditionally dependent, it is easy to see that when no signal is present (i.e., $\beta^* = 0$) the marginal distribution of the permuted responses is uniform on $\|\mathbf{y}\| S^{n-1}$. When selecting the penalty parameter, permutations of the response can behave like independent, uniformly distributed points even when the level of signal present is moderate (see simulation results in Web Appendix A). With this in mind, the next theorem considers a variant $\tilde{\lambda}_n = \text{median}(\lambda_0(\mathbf{Y}_1), \dots, \lambda_0(\mathbf{Y}_N))$ of permutation selection in which the responses $\mathbf{y}_\pi$ are replaced by independent random vectors $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ distributed as $\|\mathbf{y}\| \mathbf{U}$, where $\mathbf{U}$ is uniform on S^{n-1} . (We expect a similar result to hold for $\hat{\lambda}_{perm}$ but the analysis is substantially more complicated, as the permuted responses depend in a more direct way on the design matrix $\mathbf{X}$.)

THEOREM 1: Let $\mathbf{X}$ and $\mathbf{y}$ be defined as above. If n and p tend to infinity in such a way that $\log p = o(n)$ while $\|\beta^*\|$ remains constant, then

$$\tilde{\lambda}_n = \sqrt{\frac{2(\|\beta^*\|^2 + \sigma^2) \log p}{n}} (1 + o_p(1)). \quad (9)$$

Proof: As the number of vectors $\mathbf{Y}_i$ used to define $\tilde{\lambda}$ is fixed, it suffices to show that $\lambda_0(\|\mathbf{y}\| \mathbf{U})$ has the stated asymptotic behavior. It is known (cf. Friedman et al. 2010) that

$$\lambda_0(\|\mathbf{y}\| \mathbf{U}) = n^{-1} \max_{1 \leq j \leq p} |(X_{\cdot j})^T (\|\mathbf{y}\| \mathbf{U})| = \|\mathbf{y}\| n^{-1/2} \max_{1 \leq j \leq p} |\mathbf{v}_j^T \mathbf{U}|.$$

By a result of Zhang (2013), our assumptions concerning $\{\mathbf{v}_j\}$ ensure that

$$\frac{1}{\sqrt{1 - p^{-2/(n+1)}}} \cdot \max_{1 \leq j \leq p} |\mathbf{v}_j^T \mathbf{U}| \rightarrow 1 \quad \text{in probability}$$

as n, p tend to infinity. As $\log p = o(n)$, the ratio of $1 - p^{-2/(n+1)}$ and $2n^{-1} \log p$ tends to one, and we conclude that

$$\max_{1 \leq j \leq p} |\mathbf{v}_j^T \mathbf{U}| = \sqrt{\frac{2 \log p}{n}} (1 + o_p(1)). \quad (10)$$

It therefore suffices to show that $\|\mathbf{y}\|^2 = n(\|\beta^*\| + \sigma^2)(1 + o_p(1))$. By virtue of (7),

$$\|y\|^2 = \|\mathbf{X}\beta^*\|^2 + \|\varepsilon\|^2 + 2\varepsilon^T \mathbf{X}\beta^*.$$

A straightforward variance calculation shows that $\|\varepsilon\|^2 = \sigma^2 n(1 + O_p(n^{-1/2}))$. Analysis of the remaining terms is facilitated in a routine way by representing the entries of the design matrix X in terms of standard normals. In particular, we may write

$$X_{ij} = n^{1/2} \frac{\xi_{ij}}{\sqrt{\sum_{i=1}^n \xi_{ij}^2}} \text{ where } \{\xi_{ij}\} \text{ are independent } \mathcal{N}(0, 1). \quad (11)$$

Note that for each fixed row index i the variables $X_{i1}, \dots, X_{ip}$ are independent. The cross-product $\varepsilon^T \mathbf{X}\beta^*$ can be written as $\sum_{i=1}^n \varepsilon_i W_i$ where $W_i = \sum_{j=1}^p X_{ij} \beta_j^*$. It follows from (11) and elementary symmetry arguments that $\mathbb{E}X_{ij} = 0$ and $\text{Var}(X_{ij}) = 1$. Thus $\mathbb{E}W_i = 0$, and

$$\mathbb{E}(W_i^2) = \text{Var}\left(\sum_{j=1}^p X_{ij} \beta_j^*\right) = \sum_{j=1}^p \text{Var}(X_{ij} \beta_j^*) = \sum_{j=1}^p (\beta_j^*)^2 \text{Var}(X_{ij}) = \|\beta^*\|^2.$$

We wish to calculate the variance of $\varepsilon^T \mathbf{X}\beta^*$. To this end, let W_1^n denote $W_1, \dots, W_n$ and note that

$$\text{Var}\left(\sum_{i=1}^n \varepsilon_i W_i | W_1^n\right) = \sum_{i=1}^n \text{Var}(\varepsilon_i W_i | W_1^n) = \sum_{i=1}^n W_i^2 \text{Var}(\varepsilon_i) = \sigma^2 \sum_{i=1}^n W_i^2.$$

As the conditional expectation of $\varepsilon^T \mathbf{X}\beta^*$ given W_1^n is zero, it follows from the last display and the conditional variance decomposition that

$$\text{Var}(\varepsilon^T \mathbf{X}\beta^*) = \mathbb{E}\left[\sigma^2 \sum_{i=1}^n W_i^2\right] = \sigma^2 \sum_{i=1}^n \mathbb{E}(W_i^2) = n\sigma^2 \|\beta^*\|^2.$$

Since $\mathbb{E}\varepsilon^T \mathbf{X}\beta^* = 0$, we conclude that $\varepsilon^T \mathbf{X}\beta^* = O_p(n^{1/2})$.

We now turn our attention to the initial term in the expression for $\|y\|^2$. Expanding the square, we find $\|\mathbf{X}\beta^*\|^2 = n \|\beta^*\|^2 + nZ$ where $Z = \sum_{j \neq k} \beta_j^* \beta_k^* (\mathbf{v}_j^T \mathbf{v}_k)$. For each $j \neq k$, the representation (11) and symmetry arguments imply that

$$\mathbb{E}(\mathbf{v}_j^T \mathbf{v}_k) = 0 \quad \text{and} \quad \text{Var}(\mathbf{v}_j^T \mathbf{v}_k) = \mathbb{E}\left[(\mathbf{v}_j^T \mathbf{v}_k)^2\right] = n^{-1}.$$

In particular, $\mathbb{E}Z = 0$. It is easy to see that $\mathbb{E}\left[(\mathbf{v}_j^T \mathbf{v}_k)(\mathbf{v}_s^T \mathbf{v}_t)\right]$ is equal to n^{-1} if $\{j, k\} = \{s, t\}$ and equal to zero otherwise. It then follows from a routine calculation that

$$\text{Var}(Z) = n^{-1} \left(\|\beta^*\|^4 - \sum_{i=1}^n (\beta_i^*)^4 \right) \leq n^{-1} \|\beta^*\|^4$$

and we conclude that $\|\mathbf{X}\beta^*\|^2 = n\|\beta^*\|^2 + O_p(n^{1/2})$. The desired bound on $\|y\|$ follows by combining the bounds derived above.

Remarks: Theorem 1 does not make any assumptions about the sparsity of the true coefficient vectors β^* . An examination of the proof shows that the relation (9) continues to hold if $\|\beta^*\|^2 = o(n^{1/2})$. As (9) makes clear, the asymptotic behavior of $\tilde{\lambda}_n$ depends both on the noise variance σ^2 and the signal strength $\|\beta^*\|^2$. In particular, $\tilde{\lambda}_n$ is within a constant factor of the universal threshold $\lambda_n^* = \sigma \sqrt{n^{-1} \log p}$ for the Gaussian case when $\|\beta^*\|^2$ is bounded. More on universal thresholds and their relation to the asymptotic properties of the LASSO for prediction and variable selection, can be found in the book of Bühlmann and van de Geer 2011.

Recent work of Hebiri and Lederer (2013), Dalalyan et al. (2014), and others has examined in some detail how correlations among the columns x_j of the design matrix $\mathbf{X}$ affect the performance of the LASSO and the selection of the penalty parameter λ . The broad conclusion of this work is that smaller values of λ are appropriate when the columns of $\mathbf{X}$ are more (positively) correlated, a relationship that holds for the permutation selection procedure. Indeed, the permutation selection procedure adapts to linear dependence among the columns of the design matrix in a direct way, by assessing their maximum correlation with a permuted response.

4 Simulations

We performed a simulation study in which permutation selection and competing methods were used to select the penalty parameter of the LASSO. The competing methods were BIC, 10-fold CV, a selection procedure based on the covariance test of Lockhart et al. (2014), and Scaled Sparse Linear Regression (Sun and Zhang, 2012) (hereafter referred to as scalreg). Simulated data sets differed in their dependence and level of sparsity (number of truly active predictors). Both linear (Gaussian) and logistic (binomial) settings were considered.

4.1 Predictor Matrix Generation

We simulated predictor matrices with independent, multivariate Gaussian rows (samples). In particular, the n rows of $\mathbf{X}$ were generated as independent samples from a $\mathcal{N}_p(\mathbf{0}, \Sigma)$ distribution. Several structures for the covariance matrix Σ were considered.

- (A) Independent: $\Sigma = \mathbf{I}_p$;
- (B) Block correlation structure: $\sigma_{ij} = 0.5$ if $i \bmod 10 = j \bmod 10$ and $\sigma_{ij} = 0$ otherwise;
- (C) Fast decaying AR(1) correlations: $\sigma_{ij} = 0.9^{|i-j|}$;
- (D) Slowly decaying AR(1) correlations: $\sigma_{ij} = 0.99^{|i-j|}$.

4.2 Response Generation

Given the predictor matrix $\mathbf{X}$ and a simulated effects vector β (details below), we modeled the response using the generalized linear model (GLM) in equation 1. Two GLM settings were considered: Gaussian (standard linear model) and Bernoulli (logistic regression).

4.2.1 Linear Regression Model—Gaussian regression responses were simulated based on the standard linear regression version of equation (1), in which the link function $g(\cdot)$ is the identity. Given a set of s true variables selected uniformly at random, we generated the components of the associated effect vector β_s independently from a $U(0.25, 1)$ distribution. Letting $\mathbf{X}_s$ denote the restriction of the predictor matrix to the columns of the s selected variables, the response vector y was generated as

$$y = \mathbf{X}_s \beta_s + \epsilon \quad (12)$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ is a vector of independent Gaussian errors. The variance σ^2 was selected to achieve a desired signal to noise ratio (SNR) given by

$$SNR = \sigma^{-2} \beta_s^T \text{Var}(\mathbf{X}_s) \beta_s. \quad (13)$$

One hundred simulations were performed for each choice of the following simulation parameters (64 in total):

- Sample correlation structures (A)-(D);
- $SNR = 0.5, 2$;
- Number of subjects $n = 200, 1000$;
- Number of true variables $s = 1, 5, 10, 20$.

In each case, the number of variables p was equal to 500.

4.2.2 Logistic Regression Model—Binary responses were generated from a logistic GLM. For each sample i , a response was generated as $y_i \sim \text{Bernoulli}(q_i)$ where q_i was specified via the linear predictor with a logit link,

$$\text{logit}(q_i) = \log\left(\frac{q_i}{1 - q_i}\right) = \mu + \mathbf{x}_{i,s}^T \beta_s, \quad (14)$$

where $\mu = \mathbf{1}^T(\mathbf{X}_s \beta_s)$ is the intercept needed to obtain expected balance in the number of 1s (cases) and 0s (controls), $\mathbf{x}_{i,s}$ are the s predictors in sample i having truly non-zero effects, and β_s are the effects, generated independently according to the rule $\exp\{\beta_{sj}\} \sim N(\mu_\beta, 0.02^2)$. The parameter μ_β was used to control the signal level in each of the high and low dimensional data settings. For the low dimensional setting, low and high signal levels were set using $\mu_\beta = \log(1.35)$ and $\mu_\beta = \log(1.75)$ respectively. In the high dimensional setting, low and high signal levels were set using $\mu_\beta = \log(2.5)$ and $\mu_\beta = \log(3)$ respectively. For each signal level, we performed 100 simulations for each combination of conditions (correlation structure, sample size n , and number of true variables s) used in the Gaussian setting.

4.3 Competing methods

4.3.1 LASSO with λ -selection by Bayesian Information Criterion (BIC)—We implemented the standard version of BIC, selecting penalty parameter λ_{BIC} as

$$\lambda_{\text{BIC}} = \underset{\lambda}{\operatorname{argmin}} \left\{ -2\ell \left(\hat{\beta}(\lambda; \mathbf{y}, \mathbf{X}) : \mathbf{y}, \mathbf{X} \right) + \operatorname{df}(\lambda) \log(n) \right\}, \quad (15)$$

where $\operatorname{df}(\lambda) = |S_0(\lambda)|$ is the number of non-zero coefficients in the coefficient vector $\hat{\beta}(\lambda; \mathbf{y}, \mathbf{X})$. The resulting variable set is $S_0(\lambda_{\text{BIC}})$.

4.3.2 LASSO with λ -selection by Cross Validation (CV)—Cross-validation was implemented using the `cv.glmnet()` function from the R-package R/glmnet, as described in Friedman et al. (2010). In the linear regression case, the CV error for a value of λ under K -fold cross validation is defined as

$$CV(\lambda) = \sum_{k=1}^K \sum_{i \in V_k} \left[y_i - \mathbf{x}_i^T \hat{\beta}^{(-k)}(\lambda; \mathbf{y}^{(-k)}, \mathbf{X}^{(-k)}) \right]^2. \quad (16)$$

where $V_1, \dots, V_K$ is a randomly chosen partition of $\{1, \dots, n\}$ into groups of size $\lfloor n/K \rfloor$ or $\lceil n/K \rceil$, and the superscript $(-k)$ indicates that the subjects in V_k have been excluded. We report results based on the `cv.glmnet` default value of $K = 10$, using the value for λ_{1SE} (the largest value of λ having $CV(\lambda)$ within one standard error of the minimum CV error). The resulting variable set is $S_0(\lambda_{\text{CV}})$. An alternative choice, λ_{min} , is the value of λ having the minimum CV error. This latter option will produce a larger model than λ_{1SE} and is likely to have a substantially larger FDR. Also considered were $K = 3$ and $K = n$ (“jackknife”), which were found to give similar results (data not shown).

4.3.3 LASSO with λ -selection by Covariance Test—The covariance test of Lockhart et al. (2014) provides a list of p-values corresponding to points in the LASSO path where a variable is added to the model. In principle, these p-values can be used to select a value of the penalty parameter; Lockhart et al. left the specification of such a procedure for future work. Here we propose a simple procedure to select a penalty parameter λ_{CT} using the p-values from the covariance test. Let p_r be the p-value produced by the covariance test for the r th change in the LASSO path, and let

$$r^* = \max \{ r : p_r \leq \alpha \}, \quad (17)$$

where α is a fixed significance level, which we take to be 0.05 in what follows. Define λ_{CT} to be the value of λ at which the r^* th change in the LASSO path takes place; the resulting variable set is then $S_0(\lambda_{\text{CT}})$. In the case that no p-values are less than α , we define $S_0(\lambda_{\text{CT}}) = \emptyset$. In recent work, Grazier G’Sell et al. (2013) studied the FDR control methods for sequential selection procedures, and investigated the application of these procedures to the sequential p-values derived from the covariance test. As noted there, variable dependence and intermingling of signal and noise variables present practical difficulties for sequential selection procedures, and further investigation is ongoing.

Use of λ_{CT} is limited to settings in which the covariance test can be applied. The covariance test in linear regression settings relies on the error variance being known or estimable. Estimation of the error variance in high dimensional ($p > n$) settings is an active research area. For an overview of some possible methods to estimate the error variance see, for example, Reid et al. (2013). For simplicity, we provide the covariance test with the true error variance in order to assess its performance independently of any variance estimation procedure. Results based on a reasonable estimator should be similar to those with the true variance.

4.4 Scaled Sparse Linear Regression (Scalreg)

To evaluate the performance of our method against a method based on the universal threshold for the LASSO (Ye and Zhang, 2010), we consider Scaled sparse linear regression (scalreg) (Sun and Zhang, 2012). Procedures based on the universal threshold effectively eliminates the need to select a λ from the model path by selecting a penalty based on data dimension and model noise variance. Scalreg simplifies the universal threshold by eliminating the dependence on either knowing or having an appropriate estimate of the regression noise variance. It accomplishes this by an iterative procedure that jointly estimates the model coefficients and noise level in the linear model. We consider the use of the corresponding R-package, R/scalreg, under the default settings on all linear regression models.

4.5 Simulation Results

Across our simulations, the relationships between penalty selection methods were relatively consistent, with overall performance depending primarily on the complexity of the chosen parameters (stronger dependence, lower SNR, fewer samples, and more active variables). Below we describe the simulation results using the block correlation structure (covariance structure B), which are representative of our overall findings. Figures showing simulation results under other settings can be found in Web Appendix B.

4.5.1 Gaussian Simulation Results—Figure 1 shows the average power (number of true discoveries divided by total number of active variables) and one minus the average false discovery rate (number of false discoveries divided by total number of discoveries) for high and low SNR regimes, and for different numbers of true variables s . Examining the results from the low dimensional setting (left) and high dimensional setting (right), it is difficult to discern a clear winner. If we regard power plus $1 - \text{FDR}$ as a rough measure of performance, then permutation selection is comparable to, and in some cases better than, the other methods under consideration. As the number s of active variables increases, most methods exhibit a decrease in power accompanied by a simultaneous reduction in FDR, this trend being most apparent for the covariance test.

4.5.2 Logistic Simulation Results—Figure 2 summarizes the results of the logistic simulations for all applicable methods. Excluded from comparison is Scalreg, which is not applicable to the logistic setting. As in the linear case, there is no clear winner. Permutation selection is again comparable to the other methods with regards to power plus $1 - \text{FDR}$. The performance of the covariance test degrades as the number s of active variables increases.

Again, there is a trend towards decreasing power and lower FDR with increasing s , this trend being strongest for the covariance test.

4.6 CPU Time Comparison

Table 1 summarizes the CPU time (in seconds) for each method. CPU times are segregated based on the primary parameters that appeared to have the most significant effect on their value: model (Gaussian or Logistic), data dimension, and the covariance structure of the data. The only method for which covariance structure had no effect on CPU times across all settings was permutation selection. For data with dependent variables, with the exception of high dimensional logistic regression, the CPU times of permutation selection were comparable or less than those of other methods.

5 Real Data Analysis

Here we apply all applicable penalty parameter selection methods on several real data sets. An additional example is available in Web Appendix C. The data sets examined are:

- NMRI outbred mice GWAS data (Zhang et al., 2012): A genome-wide association study (GWAS) of HDL cholesterol levels in a population of outbred laboratory mice (the NMRI population). For each of 288 mice, the dataset includes HDL measurements and genotypes for 44,428 single nucleotide polymorphisms (SNPs). (These data are publicly available at <http://cgd.jax.org/datasets/phenotype/nmri.shtml>)
- The Cancer Genome Atlas Network (TCGA) breast cancer data (TCGA, 2012): Tumor and germline DNA samples were obtained from 825 patients. Gene expression values quantified by RSEM were normalized to the upper quartile of the total counts. The data was then $\log_2$ transformed and genes were median centered. Genes were filtered such that at least 70% of samples had a value, and then imputation was performed. (These data are available from the TCGA data portal at <https://tcga-data.nci.nih.gov/tcga/>). We examine the problem of distinguishing 371 Luminal A cancer samples from 170 Luminal B cancer samples based on the expression values of the 17,007 genes that remained after quality control.

For each fitting procedure we recorded the CPU time and, although our primary focus is variable selection, examined the predictive ability of the fitted model. In each assessment, the data was randomly split into training and test sets. The training set comprised two thirds of the data and was used by each method to select the LASSO penalty parameter; the test set comprised the remaining third and was used to evaluate predictive performance. Results are based on the aggregation over 100 random splittings of each data set.

5.1 NMRI outbred mice GWAS

Table 2 summarizes the sizes of the selected models, the mean squared prediction error, and the computational time for BIC, CV, and permutation selection. (The covariance test and scalreg were excluded from this comparison due to computational errors in the underlying LARS fitting procedure used by both.) Permutation selection consistently selected the

smallest model, a result consistent with the reported findings of three significant SNPs in Zhang et al. (2012). For all methods, the mean squared prediction errors were relatively high (the intercept only model had a prediction error of 0.869).

5.2 TCGA Data Analysis

Table 3 summarizes the sizes of the selected models, the percent of the test set misclassified, and the computation times for each penalty selection method (with the exception of scalreg, which is not applicable to logistic regression). The best predictive performance was given by cross-validation, followed by permutation selection, BIC, and the covariance test. CPU times varied widely: the fastest method was BIC, followed by CV and permutation selection, with the covariance test last, being significantly slower than the other methods.

6 Summary

The permutation selection method is straightforward in application, implementation, and interpretation. In our simulations, permutation selection was comparable, and in some cases superior, to the other methods under study.

As expected, the choosing λ using (the prediction-based) cross validation tends to select larger models from the LASSO sample path; these models have high power, but also relatively large false discovery rates.

By contrast, the covariance test tends to select smaller models, with lower power and lower false discover rates. Permutation selection, scalreg, and BIC lie somewhere between cross validation and the covariance test, favoring moderate sized models and a more balanced tradeoff between power and false discovery rate.

Computational errors in the underlying LARS fitting procedure used by the R packages for scalreg and the covariance test prevented us from evaluating those methods in some settings. Currently, application of the covariance test to high dimensional ($p > n$) linear models is limited to scenarios in which the error variance is known or can be estimated accurately. Thus, in our simulations under this high dimensional linear setting, as there is no default method in the R/covtest package to estimate the error variance, we supplied the covariance test with true error variance.

Cross validation, BIC, and the covariance test all require computation of the full LASSO path. By contrast, permutation selection requires only the initial point of the LASSO path, albeit under multiple permutations of the response, followed by evaluation of the LASSO at the single chosen value of λ . Since the initial point of the LASSO path can be identified analytically, determination of λ_0 for each permutation is rapid, permutation selection is often faster, sometimes by an order of magnitude, than competing methods. Nonetheless, if λ_{perm} lies far into the LASSO path of the unpermuted response, then the CPU time for the LASSO fit will be comparable to fitting the full path, and (when combined with the CPU time required to find λ_{perm}) permutation selection can be somewhat slower than competing methods. A related point is that computational optimizations for fitting the full path (eg, as might be implemented in glmnet) benefit the CPU time of permutation selection less than

other methods owing to fact that the initial part of the permutation selection procedure, estimation of perm, is fixed.

We note that all LASSO fits could have improved performance if we refitted the unpenalized model with just the variables selected; but as prediction is not our primary goal we have not done any refitting. In cases where the number of true variables is known to be very small, simple predictor-by-predictor selection methods may outperform the LASSO, regardless of how the penalty parameter is selected.

7 Discussion

When selecting variables for inclusion in a model using penalized regression methods such as the LASSO, an important part of the analysis process is selection of the penalty parameter. Selection is often accomplished using cross validation (CV) or the Bayesian information criterion (BIC). In this paper, we introduce a simple, permutation-based procedure that is intended for analyses where the primary goal of model fitting is variable selection. Permutation selection chooses a model whose associated predictors have a joint relationship with the response that is stronger than joint relationships arising by chance, under permutation of the response. Permutation selection is simple, fast, and yields interpretable models. In our simulations and real data analyses, permutation selection was competitive with CV and BIC, as well as two other selection procedures derived from recent work on regularized fitting procedures. Although we have focused here on the LASSO, permutation selection is applicable in principle to other penalized regression methods such as SCAD.

We have considered settings in which all variables are subject to penalization. In some cases it is natural to consider models in which some variables (for example, predictors known to affect the response; hereafter, “covariates”) are unpenalized. Although CV and BIC can be extended to this situation, permutation selection as described here does not immediately apply, since permutation of the response will nullify its relationship between penalized and unpenalized variables alike.

One option would be to simultaneously permute the (unpenalized) covariates with the response, assuming that there are no relationships between the them and the penalized variables; in this case, the response’s relationship with the covariates will be preseved while its relationship with the penalized variables will be broken. Note, however, that under this scheme, when relationships do exist between the covariates and penalized variables (which in real data would be common), the covariate effects will be altered as soon as any of the penalized variables are included. If that is deemed undesirable, an alternative might be to replace the response with its residuals after fitting the covariates; although since this amounts to conditioning on estimates of the covariates’ unknown effects, there is clearly scope for more work in this area.

Both permutation selection and cross validation make use of randomization, leading to variability of the selected penalty parameter. For permutation selection this variability is controlled by considering the median of the null penalties across a moderate number (100)

of permutations. Nonetheless, in most applications with moderate or small sample sizes, a more significant source of variability is the randomness of the sample itself — one that is often not accounted for in applications of the LASSO, regardless of how the penalty parameter is selected. To address the stability of models with respect to variability of the sample, a number of resampling based methods have been recently proposed; we refer the interested reader to Meinshausen and Bühlmann (2010), Valdar et al. (2012) and Sabourin et al. (2015) for more details.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank Kai Zhang and Alexandre Belloni for helpful discussions concerning their work and the theoretical properties of the permutation selection procedure. Research reported in this manuscript was supported by the following: the National Institute of General Medical Sciences of the National Institutes of Health (NIH) under award number R01GM104125 (partial support for J.S. and W.V.); the National Institute of Mental Health of the NIH under award number R01MH101819 (A.N.); the National Science Foundation (NSF) under grant number DMS-1310002 (A.N.); the University of North Carolina Lineberger Comprehensive Cancer Center (partial support for J.S. and W.V.); and by the Division of Intramural Research at the National Human Genome Research Institute, NIH (partial support for J.S.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NSF.

References

- Ayers KL, Cordell HJ. SNP Selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology*. 2010; 34:879–91. [PubMed: 21104890]
- Belloni A, Chernozhukov V, Wang L. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*. 2011; 98:791–806.
- Bühlmann, P.; van de Geer, S. *Statistics for High-Dimensional Data - Methods, Theory and Applications*. Springer; Berlin, Heidelberg: 2011. Springer Series in Statistics 0172-7397
- Dalalyan, AS.; Hebiri, M.; Lederer, J. On the Prediction Performance of the Lasso. 2014. ArXiv e-prints
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96:1348–1360.
- Feng, Y.; Yu, Y. Consistent Cross-Validation for Tuning Parameter Selection in High-Dimensional Variable Selection. 2013. ArXiv e-prints
- Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*. 2010; 33:1. [PubMed: 20808728]
- Grazier G'Sell, M.; Wager, S.; Chouldechova, A.; Tibshirani, R. False Discovery Rate Control for Sequential Selection Procedures, with Application to the Lasso. 2013. ArXiv e-prints
- Hebiri M, Lederer J. How correlations influence lasso prediction. *IEEE Transactions on Information Theory*. 2013; 59:1846–1854.
- Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A significance test for the lasso. *The Annals of Statistics*. 2014; 42:413–468. [PubMed: 25574062]
- Meinshausen N, Bühlmann P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2010; 72:417–473.
- Reid, S.; Tibshirani, R.; Friedman, J. A Study of Error Variance Estimation in Lasso Regression. 2013. ArXiv e-prints
- Sabourin J, Nobel AB, Valdar W. Fine-Mapping Additive and Dominant SNP Effects Using Group-LASSO and Fractional Resample Model Averaging. *Genetic Epidemiology*. 2015; 39:77–88. [PubMed: 25417853]

- Shao J. Linear model selection by cross-validation. *Journal of the American Statistical Association*. 1993; 88:486–494.
- Sun T, Zhang C-H. Scaled sparse linear regression. *Biometrika*. 2012
- TCGA. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012; 490:61–70. [PubMed: 23000897]
- Tibshirani R. Regression shrinkage and selection via lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1996; 58:267–288.
- Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2011; 73:273–282.
- Valdar W, Sabourin J, Nobel A, Holmes CC. Reprioritizing genetic associations in hit regions using lasso-based resample model averaging. *Genetic Epidemiology*. 2012; 36:451–62. [PubMed: 22549815]
- Ye F, Zhang C-H. Rate minimaxity of the lasso and dantzig selector for the q loss in r balls. *JOURNAL OF MACHINE LEARNING RESEARCH*. 2010; 11:3519–3540.
- Zhang, K. Rank-Extreme Association of Gaussian Vectors and Low-Rank Detection. 2013. ArXiv e-prints
- Zhang W, Korstanje R, Thaisz J, Staedtler F, Harttman N, Xu L, Feng M, Yanas L, Yang H, Valdar W, Churchill GA, DiPetrillo K. Genome-wide association mapping of quantitative traits in outbred mice. *G3: Genes—Genomes—Genetics*. 2012; 2:167–174. [PubMed: 22384395]

Figure 1.

Bar plots of power and 1-false discovery rate (FDR) for selected gaussian regression simulation settings. s indicates the number of true signals.

Figure 2.

Bar plots of power and 1–false discovery rate (FDR) for selected logistic regression simulation setting. s indicates the number of true signals.

Table 1
Mean CPU time (seconds) for Gaussian and Logistic regression simulation settings

Gaussian Regression Simulation CPU times								
Method	Low Dimensional Data; Covariance:				High Dimensional Data; Covariance:			
	A	B	C	D	A	B	C	D
Permutation	0.548	0.581	0.605	0.654	0.113	0.106	0.107	0.104
BIC	0.290	0.492	1.117	3.925	0.057	0.072	0.105	0.130
CV	3.412	5.677	12.856	43.315	0.645	0.763	1.085	1.419
covTest	11.320	11.614	13.497	12.479	5.628	5.309	5.825	5.069
scalreg	16.021	16.596	20.034	19.630	3.340	3.412	4.192	5.231

Logistic Regression Simulation CPU times								
Method	Low Dimensional Data; Covariance:				High Dimensional Data; Covariance:			
	A	B	C	D	A	B	C	D
Permutation	1.208	1.278	1.200	1.397	0.492	0.542	0.500	0.496
BIC	11.426	10.586	18.642	26.148	0.085	0.092	0.101	0.175
CV	50.823	61.335	120.690	279.152	0.909	0.992	1.065	1.783
covTest	619.399	984.847	993.310	2866.871	27.846	30.675	29.598	35.452

Table 2
Means and standard errors of the model sizes, mean squared prediction error, and computation times from 10 random splits of the NMRI outbred mice GWAS data

Method	Model Size	Mean squared prediction error	CPU seconds
BIC	4.67 (0.211)	0.698 (0.01)	1.94 (0.015)
CV	8.14 (1.137)	0.709 (0.012)	21.686 (0.215)
Permutation	3.03 (0.174)	0.809 (0.014)	18.065 (0.163)

Table 3
Means and standard errors of the model sizes, percent of test set misclassified, and computation times from 10 random splits of the TCGA luminal subsets data

Method	Model Size	Percent Misclassified	CPU seconds
BIC	17.85 (0.628)	15.1 (0.374)	1.257 (0.035)
covTest	4.33 (0.844)	28.2 (0.650)	1365.096 (14.674)
CV	58.15 (0.926)	8.8 (0.199)	15.135 (0.288)
Permutation Selection	22.31 (0.311)	13.8 (0.282)	12.717 (0.236)