2011-03-10

# Variable Selection and Parameter Estimation Using a Continuous and Differentiable Approximation to the L0 Penalty Function

Douglas Nielsen VanDerwerken
*Brigham Young University - Provo*

# Variable Selection and Parameter Estimation using a Continuous and Differentiable Approximation to the $L_0$ Penalty Function

Douglas N. VanDerwerken

A project submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

H. Dennis Tolley, Chair
David A. Engler
William F. Christensen

Department of Statistics

Brigham Young University

April 2011

ABSTRACT

Variable Selection and Parameter Estimation using a Continuous and Differentiable Approximation to the $L_0$ Penalty Function

Douglas N. VanDerwerken
Department of Statistics, BYU
Master of Science

$L_0$ penalized likelihood procedures like Mallows' $C_p$, AIC, and BIC directly penalize for the number of variables included in a regression model. This is a straightforward approach to the problem of overfitting, and these methods are now part of every statistician's repertoire. However, these procedures have been shown to sometimes result in unstable parameter estimates as a result on the $L_0$ penalty's discontinuity at zero. One proposed alternative, seamless-$L_0$ (SELO), utilizes a continuous penalty function that mimics $L_0$ and allows for stable estimates. Like other similar methods (*e.g.* LASSO and SCAD), SELO produces sparse solutions because the penalty function is non-differentiable at the origin. Because these penalized likelihoods are singular (non-differentiable) at zero, there is no closed-form solution for the extremum of the objective function. We propose a continuous and everywhere-differentiable penalty function that can have arbitrarily steep slope in a neighborhood near zero, thus mimicking the $L_0$ penalty, but allowing for a nearly closed-form solution for the $\hat{\boldsymbol{\beta}}$ vector. Because our function is not singular at zero, $\hat{\boldsymbol{\beta}}$ will have no zero-valued components, although some will have been shrunk arbitrarily close thereto. We employ a BIC-selected tuning parameter used in the shrinkage step to perform zero-thresholding as well. We call the resulting vector of coefficients the ShrinkSet estimator. It is comparable to SELO in terms of model performance (selecting the truly nonzero coefficients, overall MSE, etc.), but we believe it to be more intuitive and simpler to compute. We provide strong evidence that the estimator enjoys favorable asymptotic properties, including the oracle property.

Keywords: Penalized likelihood; variable selection; oracle property; large $p$, small $n$

# ACKNOWLEDGMENTS

I would like to thank my advisor Dr. H. Dennis Tolley for proposing the ShrinkSet penalty function and for his unrelenting insistence on rigor. In addition, I would like to thank the other faculty members on my committee, especially Dr. David Engler, for their helpful revisions. Many thanks also to Dr. Xihong Lin for her helpful comments and to Dr. Valen Johnson for lending me R-code for the salary data analysis. Finally, I would like to thank my wife, Charisse, for being supportive of me throughout this process, which included many late nights in the Talmage building.

CONTENTS

CHAPTER 1

# INTRODUCTION

$L_0$ penalized likelihood procedures like Mallows' $C_p$, AIC, and BIC directly penalize for the number of variables included in a regression model. This is a straightforward approach to the problem of overfitting, and these methods are now part of every statistician's repertoire. However, Breiman (1996) demonstrates that these procedures can result in unstable parameter estimates. According to Dicker (2010), this instability is the result of the $L_0$ penalty's discontinuity at zero. In addition, $L_0$ likelihood procedures are generally NP-hard. In order to improve computational efficiency over other methods and alleviate the instability problems identified by Breiman (1996), Dicker (2010) recommends "continuous penalty functions designed to mimic the $L_0$ penalty" and proposes one such penalty called the seamless-$L_0$ (SELO). We go one step further and propose a continuous and differentiable penalty function that can have arbitrarily steep slope in a neighborhood near zero, thus mimicking the $L_0$ penalty, but allowing for a nearly closed-form solution for the $\hat{\boldsymbol{\beta}}$ vector. The resulting estimator is comparable to SELO in terms of model performance (selecting the truly nonzero coefficients, overall MSE, etc.), but we believe it to be more intuitive and simpler to compute.

It has been shown (see Fan and Li (2001), Antoniadis and Fan (2001)) that a penalty function cannot produce sparse solutions (solutions where many coefficients are estimated to be zero) unless the penalty is singular at the origin. Because the proposed penalty is differentiable at zero, it cannot produce sparse solutions on its own. However, due to its arbitrarily steep slope in a neighborhood of the origin, the proposed penalty is able to shrink estimates arbitrarily close to zero. We employ the same BIC-selected tuning parameter to both shrink estimates towards zero and then threshold the shrunken estimates to exactly zero. While our method at this point ceases to remain fully continuous, the

amount of discontinuity is so small (because of the shrinkage step) that the major problems with discontinuity identified by Breiman are averted. Importantly, the proposed estimator appears to also enjoy asymptotic normality with the same asymptotic variance as the least squares estimator. These properties (sparsity and asymptotic normality) comprise what Fan and Li (2001) call the "oracle properties." In other words, an oracle procedure performs as well as if the set of truly nonzero coefficients were known in advance.

## 1.1 Outline

We first give a thorough discussion of the relevant literature on penalized likelihood methods. We then introduce the proposed penalty function and present an iterative closed-form solution for $\hat{\boldsymbol{\beta}}$. We approximate the finite-sample distribution of the estimator, and establish (through heuristic argument and simulation) its asymptotic oracle properties. We present the results of various simulations and comment on various other applications of this method, including where $n \ll p$. We conclude by analyzing a medical data set and a data set used in a discrimination lawsuit. Throughout, we compare our method to standards in the field.

CHAPTER 2

# LITERATURE REVIEW

Let $\mathbf{X}$ be an $n \times p$ data matrix where rows $1, \ldots, n$ represent observations and columns $1, \ldots, p$ represent variables. If an estimate for the intercept is desired, a column of 1's may be appended to $\mathbf{X}$. The vector $\mathbf{y}$ is of length $n$ and represents the response associated with each observation. It is typically assumed that $\mathbf{y}$ follows a normal distribution with expectation $\mathbf{X}\boldsymbol{\beta}$ and variance $\sigma^2 I$ when errors are independent, or more generally $\boldsymbol{\Sigma}$. The maximum likelihood estimator for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. It is straightforward to show that $\hat{\boldsymbol{\beta}}$ is distributed normally with mean $\boldsymbol{\beta}$ and variance $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$. Thus, the ordinary least squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ is unbiased. Other advantages include that it has the minimum variance of all unbiased estimators, that it is not computationally burdensome to calculate, and that it is widely recognized and understood among other disciplines. In addition, it is relatively easy to derive the OLS estimator's finite-sample and asymptotic properties. However, Hastie, Tibshirani, and Friedman (2009) point out two important drawbacks of $\hat{\boldsymbol{\beta}}$:

1. *Prediction inaccuracy.* Although $\hat{\boldsymbol{\beta}}$ has zero bias, its variance can be high relative to biased estimators. Shrinking or setting some coefficients to zero can improve overall mean square error and prediction accuracy.

2. *Interpretation.* When $p$ is high, it can be difficult to understand all effects simultaneously. Sometimes it may be advantageous to lose a little accuracy in exchange for a more parsimonious model.

The following methods are meant to improve upon $\hat{\boldsymbol{\beta}}$ in one or both of these areas.

## 2.1 Subset selection

The idea of subset selection is to select the best subset (not necessarily proper) of the $p$ predictors of $\mathbf{y}$ in terms of some optimality criterion. The model chosen is the OLS estimator fit on the $\mathbf{X}$ matrix of reduced dimensionality determined by those predictors which were retained. Most optimality criteria directly penalize the number of terms in the model, thus making use of the $L_0$ penalty, and reward terms that are effective at reducing model error. According to Dicker (2010), $L_0$ penalties are of the form

$$p_\lambda(\beta_j) = \lambda I\{\beta_j \neq 0\},$$

where $I$ is the indicator function. Examples of $L_0$-based optimality criteria include Mallows' $C_p$ (Mallows 1973), AIC, (Akaike 1974), and BIC (Schwarz 1978). Breiman (1996) argues that subset selection is unstable in that changing just one observation can drastically alter the minimizer of the residual sum of squares. This is largely due to the all-or-nothing nature of the $L_0$ penalty. Less sophisticated methods such as choosing predictors so as to maximize adjusted $R^2$ or zero-thresholding coefficients with nonsignificant $t$-statistics are also meant to reduce model size, but do not formally incorporate the $L_0$ penalty.

## 2.2 Ridge regression

One of the first alternatives to ordinary least squares was ridge regression, which was developed by Hoerl and Kennard (1970) in order to deal with multicollinearity. The ridge estimator is found by minimizing

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left( y_i - \mathbf{x}_i'\boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

with respect to $\boldsymbol{\beta}$. In essence, the solution $\hat{\boldsymbol{\beta}}$ minimizes the sum of squared residuals subject to

$$\sum_{j=1}^{p} \beta_j^2 \leq f(\lambda),$$

where $f$ is a one-to-one function (Hastie et al. 2009). For $\lambda > 0$, $Q(\boldsymbol{\beta})$ penalizes coefficients based on their magnitude, where the amount of shrinkage is directly related to the choice of $\lambda$. It is easily shown that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda I)^{-1}\mathbf{X}'\mathbf{y}$$

minimizes $Q(\boldsymbol{\beta})$. Hoerl and Kennard (1970) introduced a graphic known as the "ridge trace" to help the user determine the optimal value for $\lambda$. An important aspect of the ridge shrinkage is that, in general, no coefficients are shrunk to exactly zero. Because ridge regression does not shrink coefficients to zero, it is not really a variable selection technique. We discuss it because it was one of the first attempts to gain popularity using an objective function other than the sum of squared residuals. The other procedures presented below all perform shrinkage of some coefficients to zero.

### 2.3 LASSO

LASSO, which stands for "least absolute shrinkage and selection operator," was proposed by Tibshirani (1996) as an alternative to subset selection and ridge regression that retains the good features of both. Minimizing

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left( y_i - \mathbf{x}_i'\boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

for a fixed value of $\lambda$ with respect to $\boldsymbol{\beta}$ yields the LASSO estimate $\hat{\boldsymbol{\beta}}$. This is equivalent to minimizing the sum of squared residuals subject to

$$\sum_{j=1}^{p} |\beta_j| \leq t,$$

where $t \geq 0$ (or $\lambda$) is a tuning parameter that can be chosen by cross-validation, generalized cross-validation, or "an analytical unbiased estimate of risk" (Tibshirani 1996). As in the ridge case, increasing the tuning parameter leads to increased shrinkage. Unlike before, no truly closed-form solution exists, though the iterative ridge regression algorithm

$$\hat{\boldsymbol{\beta}}^{(k+1)} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{W})^{-1}\mathbf{X}'\mathbf{y},$$

where $\mathbf{W}$ is a diagonal $p \times p$ matrix with diagonal elements $|\hat{\beta}^{(k)}|^{-1}$, generally guarantees convergence to the minimum of $Q(\boldsymbol{\beta})$. Tibshirani (1996) offers several other algorithms for finding $\hat{\boldsymbol{\beta}}$, but these were supplanted by the more efficient least angular regression (LARS) algorithm in 2004 (Efron, Hastie, Johnstone, and Tibshirani 2004).

### 2.4 Bridge regression

Developed by Frank and Friedman (1993), bridge regression encompasses a large class of estimators, of which subset selection, ridge, and LASSO are special cases. The bridge estimator is found by minimizing:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left( y_i - \mathbf{x}_i' \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^\gamma$$

with respect to $\boldsymbol{\beta}$ where $\lambda > 0$ (as before) determines the strength of the penalty and $\gamma > 0$ is an additional meta parameter that "controls the degree of preference for the true coefficient vector $\boldsymbol{\beta}$ to align with the original variable axis directions in the predictor space" (Frank and Friedman 1993). This is equivalent to minimizing the sum of squared residuals subject to

$$\sum_{j=1}^{p} |\beta_j|^\gamma \leq t$$

for some $t$, a function of $\lambda$. According to Fu (1998), Frank and Friedman (1993) did not solve for the estimator of bridge regression for an arbitrary $\gamma > 0$, but they did recommend optimizing the $\gamma$ parameter. Fan and Li (2001) explain that the bridge regression solution is only continuous for $\gamma \geq 1$, but does not threshold (produce sparse solutions) when $\gamma > 1$. Thus, only when $\gamma = 1$ (which is the LASSO penalty) is the solution continuous and sparse; but then the solution is biased by a constant $\lambda$.

### 2.5 SCAD

The SCAD or "smoothly clipped absolute deviation" penalty was proposed by Fan and Li (2001) in their seminal paper *Variable Selection via Nonconcave Penalized Likelihood and*

*its Oracle Properties*. The SCAD penalty is best defined in terms of its first derivative,

$$p'_\lambda(\boldsymbol{\beta}) = \lambda\left\{ I\{\boldsymbol{\beta} \leq \lambda\} + \frac{(a\lambda - \boldsymbol{\beta})_+}{(a-1)\lambda} I\{\boldsymbol{\beta} > \lambda\} \right\},$$

where $I$ is the indicator function. An important improvement of SCAD over LASSO is that large values of $\boldsymbol{\beta}$ are penalized less than small values of $\boldsymbol{\beta}$. Also, unlike traditional variable selection procedures, the SCAD estimator's sampling properties can be precisely established. For example, Fan and Li (2001) demonstrated that as $n$ increases, the SCAD procedure selects the true set of nonzero coefficients with probability tending to one. In addition, the SCAD estimator exhibits asymptotic normality with mean $\boldsymbol{\beta}$ and variance $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$, the variance with the true submodel known. These two properties — consistency in variable selection (sometimes called "sparsity") and asymptotic normality with the minimum possible variance of an unbiased estimator — constitute the "oracle" properties. In other words, an oracle procedure performs as well asymptotically as if the set of truly nonzero coefficients were known in advance. Asymptotic oracle properties have become the gold standard in the penalized likelihood literature. Finally, Fan and Li (2001) show that the SCAD penalty can be effectively implemented in robust linear and generalized linear models.

## 2.6 Adaptive LASSO

LASSO was designed to retain the good features of subset selection (sparsity), and ridge regression (continuity at zero, which leads to stable estimates). This it does, but at the cost of producing asymptotically biased estimates because the penalty function is unbounded for large $\beta_j$. In addition, Zou (2006) proves that in nontrivial instances LASSO is inconsistent. That is, in some cases it does not select, even asymptotically, the true set of nonzero coefficients with probability tending to one. Therefore, LASSO was shown to not possess the oracle properties. To remedy this problem, Zou (2006) proposes a weighted LASSO which minimizes

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left( y_i - \mathbf{x}_i'\boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|.$$

Zou (2006) suggests using $\hat{w}_j = 1/|\beta_j|^\gamma$, and calls the resulting estimator the adpative LASSO. Optimal values of $\gamma$ and $\lambda$ can be found using two-dimensional cross-validation. The minimum of $Q(\boldsymbol{\beta})$ can be found efficiently using the LARS (Efron et al. 2004) algorithm since the penalty function is convex. Importantly, for fixed $p$, as $n$ increases, the adaptive LASSO selects the true set of nonzero coefficients with probability tending to one. In addition, the adaptive LASSO estimate exhibits asymptotic normality with mean $\boldsymbol{\beta}$ and variance $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$. Adaptive LASSO theory and methodology have been extended to generalized linear models, where the oracle properties also hold.

## 2.7 SELO

The SELO or "seamless-$L_0$" penalty was designed to mimic the $L_0$ penalty without the jump discontinuity. Proposed by Dicker (2010),

$$p_{SELO}(\beta_j, \lambda, \tau) = \frac{\lambda}{log(2)} log\left(\frac{|\beta_j|}{|\beta_j| + \tau} + 1\right)$$

utlizes a tuning parameter $\tau$ in addition to the $\lambda$ parameter common in most penalized likelihood procedures. When $\tau$ is small, $p_{SELO}(\beta_j, \lambda, \tau) \approx \lambda I\{\beta_j \neq 0\}$, yet because the penalty is continuous, the $L_0$ penalty's inherent instability problems identified by Breiman (1996) are mitigated. A recommended implementation fixes $\tau = 0.01$, then maximizes the penalized likelihood by a coordinate descent algorithm for a collection of $\lambda$s, selecting the optimal $\lambda$ in terms of minimal BIC (Dicker 2010). (The LARS algorithm (Efron et al. 2004) could not be employed for optimization, because the SELO penalty is non-convex.) Dicker (2010) shows through various simulations that in finite samples the SELO procedure is superior to existing methods in terms of proportion of the time that the correct model is selected, MSE, and model error (defined as $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\Sigma(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$). We shall use the same simulation setup and criteria as Dicker (2010) to examine the finite-sample performance of our estimator. Asymptotically, the SELO estimator enjoys the oracle properties described by Fan and Li (2001).

## 2.8 Other methods

The foregoing list is by no means exhaustive. It does, however, represent the most common variable selection techniques based on penalized likelihood. Other variable selection methods include the nonnegative garrote (Breiman 1995), the Dantzig selector (Candès and Tao 2007), nonparametric methods (Doksum, Tang, and Tsui 2008), Bayesian methods (Mitchell and Beauchamp 1988), the elastic net (Zou and Hastie 2005), neural networks (Wikel and Dow 1993), and random forests used for determining variable importance (Breiman 2001). In addition, several authors have extended results for specific penalty functions to a broader class of functions (see, for example, Fessler (1996), Antoniadis and Fan (2001), Fan and Li (2001), or Antoniadis, Gijbels, and Nikolova (2009)).

CHAPTER 3

# METHODS

## 3.1 Proposed penalty function

The penalty function that we propose is of the form:

$$p(\beta_k) = \lambda\Big(1 - exp\big(-\frac{\beta_k^2}{\delta^2}\big)\Big).$$

By way of illustration, for $\lambda = 1$ and fixed $\delta$, the penalty $p(\beta_k)$ is essentially unity for all $\beta_k$ that deviate more than $3\delta$ from zero. For $\beta_k$ near zero, the penalty function is essentially zero. This penalty is similar to the SELO penalty (Fig. 1), and can be made arbitrarily similar to the $L_0$ penalty by setting $\delta \approx 0$ (Fig. 2).

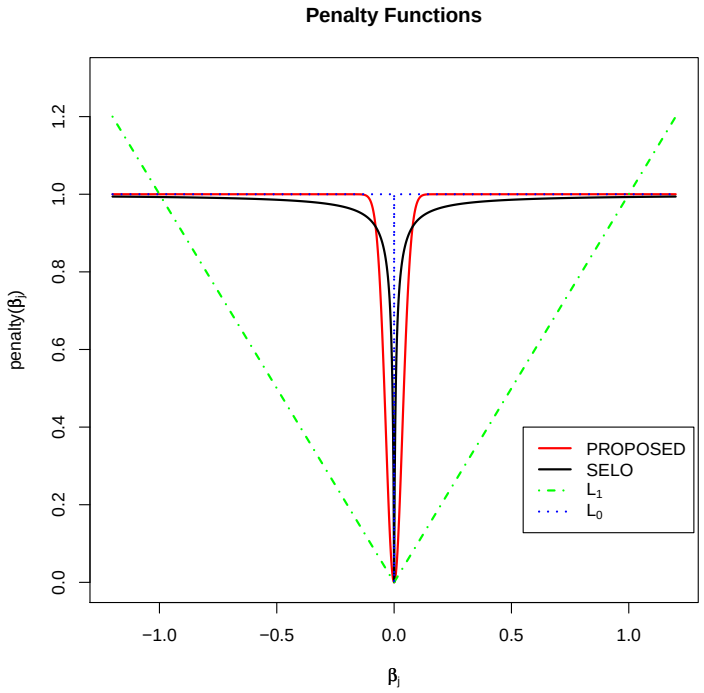


Figure 3.1: SELO: $\lambda = 1, \tau = 0.01$; $L_1, L_0 : \lambda = 1$; PROPOSED: $\lambda = 1, \delta = .05$

In fact, it can be easily shown that the limit of $p(\beta_k)$ as $\delta^2 \to 0$ is the $L_0$ penalty. We state this result as Theorem 1.

*Theorem 1.*

$$\lim_{\delta^2 \to 0} \quad p_\lambda(\beta_k) = \lambda I\{\beta_k \neq 0\},$$

where $I$ is the indicator function.

*Proof:* Suppose $\beta_k \neq 0$. Then $\lim_{\delta^2 \to 0} \frac{\beta_k^2}{\delta^2} = \infty$. Since $\lim_{x \to \infty} e^{-x} = 0$, we have $\lim_{\delta^2 \to 0} p_\lambda(\beta_k) = \lambda$. Now suppose $\beta_k = 0$. By L'Hopital's rule, we have $\lim_{\delta^2 \to 0} \frac{\beta_k^2}{\delta^2} = 0$. So $\lim_{\delta^2 \to 0} p_\lambda(\beta_k) = \lambda(1 - e^0) = 0$.

∎

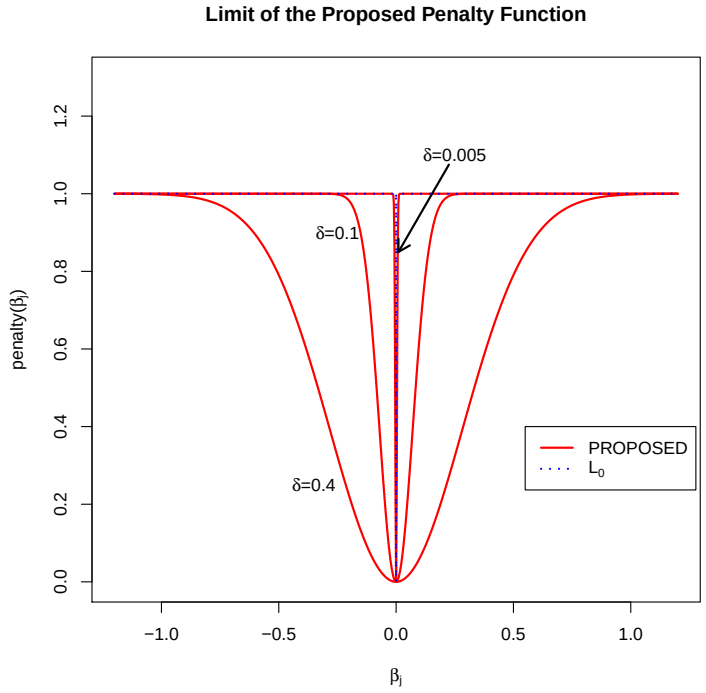


Figure 3.2: $L_0 : \lambda = 1$; PROPOSED: $\lambda = 1$ and $\delta$ varies. The limit of the proposed penalty function as $\delta$ goes to zero is $L_0$.

We note that in their paper on categorizing classes of penalized likelihood penalties, Antoniadis et al. (2009) refer to a slightly different form of $p(\beta_k)$ as an example of a penalty function that is smooth at zero and non-convex. In addition, several papers on image processing have cited this penalty function (sometimes called the Leclerc penalty) in connection with shrinkage methods (e.g. Nikolova (2005), Corepetti, Heitz, Arroyo, Mémin, and Santa-Cruz (2006)). However, at the time of this writing, no attempt has been made in the literature either to propose a method for obtaining an optimal value of $\delta$, or to use this penalty as the first step of a variable selection procedure, or to compare the associated sparse estimator to established methods in terms of finite-sample or asymptotic performance, as we do here.

### 3.2 Iterative closed-form solution for our estimator

An important distinction between this penalty function and those of LASSO, SELO, and other variable selection methods is that it is everywhere-differentiable. This allows for an essentially closed-form solution for $\hat{\boldsymbol{\beta}}$, which is obtained by minimizing

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big( y_i - \mathbf{x}_i' \boldsymbol{\beta} \Big)^2 + \lambda \sum_{j=1}^{p} \Big( 1 - exp\big( - \frac{\beta_j^2}{\delta^2} \big) \Big).$$

with respect to $\boldsymbol{\beta}$. This simultaneously shrinks the near-zero entries of the $\boldsymbol{\beta}$ vector towards zero and gives accurate estimates for the truly nonzero $\boldsymbol{\beta}$s. Note that the first component of $Q(\boldsymbol{\beta})$ is the residual sum of squares and the second is the penalty function with some user-specified constant $\lambda$. Heuristically, the penalty term forces coefficients that are not contributing to the reduction of squared error to decrease in magnitude. We proceed with the analytical solution for the minimum of $Q(\boldsymbol{\beta})$.

For fixed $\lambda$, $\delta$, and $p = ncol(\mathbf{X})$, let

$$f(\boldsymbol{\beta}) = \lambda p - \lambda \Big( exp\big( - \frac{\beta_1^2}{\delta^2} \big) + \ldots + exp\big( - \frac{\beta_p^2}{\delta^2} \big) \Big).$$

Then

$$\frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \left[ 2\beta_1 \frac{\lambda}{\delta^2} exp\big( - \frac{\beta_1^2}{\delta^2} \big) \qquad \ldots \qquad 2\beta_p \frac{\lambda}{\delta^2} exp\big( - \frac{\beta_p^2}{\delta^2} \big) \right]^T.$$

Now define the $p \times p$ weight matrix $\mathbf{W}$ in terms of some known estimate of $\boldsymbol{\beta}$ (e.g., the OLS estimator) called $\tilde{\boldsymbol{\beta}}$:

$$\mathbf{W} = diag\left(\left[\frac{\lambda}{\delta^2}exp\big(-\frac{\tilde{\beta}_1^2}{\delta^2}\big) \qquad \ldots \qquad \frac{\lambda}{\delta^2}exp\big(-\frac{\tilde{\beta}_p^2}{\delta^2}\big)\right]\right).$$

With $\mathbf{W}$ known, we can differentiate $Q(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$,

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{y} + 2(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} + 2\mathbf{W}\boldsymbol{\beta},$$

yielding the closed-form solution $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \mathbf{W})^{-1}\mathbf{X}'\mathbf{y}$ for known initial $\tilde{\boldsymbol{\beta}}$.

Now, recall that $\mathbf{W}$ was estimated using $\tilde{\boldsymbol{\beta}}$. In order to account for this, we repeat the calculation of $\mathbf{W}$ using our new estimate $\hat{\boldsymbol{\beta}}$ and then solve for $\hat{\boldsymbol{\beta}}$ again. This is repeated until sufficient convergence is achieved. This process is analogous to what Fan and Li (2001) call "iterative ridge regression." Our studies indicated that fewer than five or six iterations were usually satisfactory (as compared to at least ten iterations for SCAD – see Fan and Li (2001)), but if one prefers a more exact criterion for convergence, the process can be continued until the sum of absolute deviations between successive estimators is less than a user-specified $\epsilon$, that is, until

$$\sum_{j=1}^{p} |\hat{\beta}_j^{(k)} - \hat{\beta}_j^{(k+1)}| < \epsilon.$$

As we show below, the one-step estimator, for which there is an exact closed-form solution, behaves well asymptotically.

One difference between our process and those employed by adaptive LASSO and SCAD is that the proposed penalty function and its derivatives can be exactly calculated at the previous estimate of $\hat{\boldsymbol{\beta}}$ while the other methods rely on a local quadratic approximation (LQA) of their penalty function and derivatives. Recall that a penalty function cannot produce sparse solutions unless the penalty is singular at the origin (Antoniadis and Fan 2001). Thus, after implementing LQA, the approximate penalty functions can no longer produce sparse solutions. Fan and Li (2001) admit that for SCAD, near-zero values for $\beta_k$ must be set to be exactly zero at each iteration of the algorithm for minimizing the objective

function in order to "significantly reduce the computational burden." Zou (2006) cites Fan and Li's LQA method when introducing adaptive LASSO, but fails to mention the implicit zero-thresholding step. However, Zou and Li (2008), the (co-)authors of the adaptive LASSO and SCAD papers, respectively, jointly wrote an article on one-step penalized likelihood estimators, in which they are explicit about the fact that the zero-thresholding inherent in LQA requires an additional tuning parameter (Zou and Li 2008).

Our method, on the other hand, can calculate the exact values of the penalty function and its derivatives at the previous estimate of $\hat{\boldsymbol{\beta}}$, because it is already everywhere-differentiable. Because it is not singular at zero, it too must employ a zero-thresholding tuning parameter. Recall that the proposed penalty is a function of $\lambda$ and $\delta$. These parameters must either be specified by the user or selected in terms of some optimality criterion. The same tuning parameter $\delta$ can serve not only as an argument for the penalty function, but also as a zero-thresholding cutoff to be used after our estimator converges to the minimum of $Q(\boldsymbol{\beta})$. Finding the minimum of $Q(\boldsymbol{\beta})$ shrinks unimportant coefficients close to zero, and then these coefficients can be set to be exactly zero, as in SCAD or adaptive LASSO. We call it the ShrinkSet estimator to reflect this two-step process and because its function as a variable selection technique is to reduce the dimensionality of our predictor space. The method for choosing optimal values of $\lambda$ and $\delta$ will be explained in more detail hereinafter.

### 3.3 Standard error formula

Letting $\mathcal{A}$ denote the set of indices corresponding to nonzero components of $\hat{\boldsymbol{\beta}}$, and for a fixed matrix $\mathbf{W}$, the vector $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ is a linear combination of $\mathbf{y}$. Thus, the exact variance-covariance of $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$, given $\mathbf{W}$, is $\sigma^2(\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}} + \mathbf{W}_{\mathcal{A},\mathcal{A}})^{-1}\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}(\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}} + \mathbf{W}_{\mathcal{A},\mathcal{A}})^{-1}$. Therefore, we estimate the variance using the most recent estimate of $\mathbf{W}$, called $\widehat{\mathbf{W}}$:

$$\widehat{Var}\left[\hat{\boldsymbol{\beta}}_{\mathcal{A}}\right] = \widehat{Var}\left[\hat{\boldsymbol{\beta}}_{\mathcal{A}}|\widehat{\mathbf{W}}\right] = \hat{\sigma}^2(\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}} + \widehat{\mathbf{W}}_{\mathcal{A},\mathcal{A}})^{-1}\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}(\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}} + \widehat{\mathbf{W}}_{\mathcal{A},\mathcal{A}})^{-1},$$

where $\hat{\sigma}^2 = \widehat{Var}\left[\mathbf{y}\right] = \dfrac{(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}})'(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}})}{(n - |\mathcal{A}|)}$ and $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$. This corresponds nicely with the standard error estimates for adaptive LASSO, SCAD, and SELO

obtained using a sandwich formula (see, for example, Tibshirani (1996)). Note that the limit of $\widehat{Var}\left[\hat{\boldsymbol{\beta}}_{\mathcal{A}}\right]$ as $\mathbf{W}_{\mathcal{A},\mathcal{A}}$ approaches the zero matrix is $\hat{\sigma}^2(\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-1}$.

### 3.4 Selection of $\lambda$

The user specifies $\lambda$ based on the perceived danger of overfitting. We would hope that a low $\lambda$ value would result in many nonzero $\hat{\beta}_k$s and that a high $\lambda$ value would allow only the most extreme nonzero $\hat{\beta}_k$s, as in LASSO. Because the ShrinkSet penalty is differentiable at zero, however, it cannot shrink estimates to exactly zero, and so regardless of $\lambda$, the minimizer of $Q(\boldsymbol{\beta})$ retains all $p$ coefficients. A reasonable surrogate, then, for the number of nonzero parameters in the model, is

$$df(\lambda) = tr\Big(\mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{W}(\lambda))^{-1}\mathbf{X}\Big),$$

which is monotone decreasing in $\lambda$ (Hastie et al. 2009). If $\lambda = 0$, then $df(\lambda) = p$, and the minimizer of $Q(\boldsymbol{\beta})$ is the OLS estimator. Note that this corresponds nicely to the interpretation of $\lambda$ in the LASSO and ridge regression cases. Our studies suggest that after zero-thresholding (discussed below), the resulting estimates $\hat{\boldsymbol{\beta}}(\lambda)$ are similar across large ranges of $\lambda$. Throughout this paper we use $\lambda = 50,000$.

### 3.5 Selection of $\delta$

Dicker (2010) cites several papers establishing the superiority of BIC tuning parameter selection over GCV and AIC. Our simulations confirmed that in general BIC was a better surrogate for MSE (prediction error) than was AIC or GCV. Following Dicker (2010), we select the tuning parameter $\delta$ (for a fixed $\lambda$) so as to minimize

$$\text{BIC}(\hat{\boldsymbol{\beta}}) = log\left[\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{(n-d)}\right] + \frac{d \times log(n)}{n},$$

where $d$ is the number of variables (nonzero coefficients) in the model.

Recall that everywhere-differentiable functions do not set any coefficients to exactly zero. To produce sparse solutions using the ShrinkSet penalty, it is therefore necessary to

employ a zero-thresholding function, which sets all components (of a vector-valued estimator) to zero that do not exceed some threshold. A convenient choice for this threshold is $\delta$. Thus $\delta$ serves as a tuning parameter for the iterative ridge regression, which shrinks unimportant $\beta_k$ towards zero, and as a zero-thresholding cutoff to be applied after said shrinkage. It turns out that this thresholding after shrinkage is an important feature of our proposed variable selection method. That the maximum penalized likelihood estimator (or its computationally superior one-step surrogate) shrinks certain $\beta_k$ towards zero before employing the zero-thresholding makes it superior to simply setting different sets of OLS coefficients to zero and selecting the estimator with minimum BIC (see simulation studies section). This is equivalent to running the ShrinkSet algorithm using $\lambda = 0$. The discontinuity discouraged by Breiman (1996) is not so pronounced for relatively large $\lambda$s, because the estimates are esentially zero before they were set to be exactly zero. Also, unlike SCAD and adaptive LASSO, it is not necessary to specify an additional tuning parameter, which can be computationally burdensome.

In order to find the optimal $\delta$, it was necessary to search over a dense grid of possible values. The nature of the penalty function is such that $\delta \in \big(0,\ \max(|\tilde{\boldsymbol{\beta}}|)\big)$, but usually closer to 0. We therefore begin our search over the range $[\epsilon,\ \max(|\tilde{\boldsymbol{\beta}}|)]$ for some user-specifed $\epsilon > 0$. A reasonable choice in many applications is $\epsilon = 0.005$, or $\epsilon = \frac{1}{2}\min(|\tilde{\boldsymbol{\beta}}|)$. At each of 30 equally spaced $\delta$ values on this interval (endpoints included in the 30 values), we employ the shrinking step by iteratively calculating $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \mathbf{W}(\delta))^{-1}\mathbf{X}'\mathbf{y}$ until convergence. We then employ the zero-thresholding step by setting all $\hat{\boldsymbol{\beta}}_k < \delta$ equal to zero. We calculate BIC for each zero-thresholded $\hat{\boldsymbol{\beta}}$ and find the $\delta$ associated with the minimum of the 30 BIC values. Call this value $\delta_{min}$, and the $\delta$ values directly left and right of $\delta_{min}$ call $\delta_L$ and $\delta_R$. We then narrow the search grid to $[\delta_L,\ \delta_R]$ and calculate the thresholded estimate and associated BIC at 30 equally spaced $\delta$ values in this range, including the endpoints. We select as our optimal $\delta$ the value in the finer grid corresponding to the minimum BIC. We found this approach to be more precise and more efficient than a dense search of the original

range $[\epsilon,\ \max(|\tilde{\boldsymbol{\beta}}|)]$. This is due to the relatively smooth, convex nature of the BIC curve over this range. BIC($\delta$) will often have the same value for several values of $\delta$ in a small range. This is because different choices for $\delta$ result in the same model. In these cases, we select the smallest $\delta$ associated with the minimum BIC.

This procedure is one possible method for searching over a reasonably dense range of $\delta$ values, but it can be easily adapted if the user desires more (or less) precision. For example, in the $p \gg n$ simulation studies, we elected to do a one-stage search over 30 equally spaced $\delta$ values on $\left[0.005,\ \frac{1}{4}\max(|\tilde{\boldsymbol{\beta}}|)\right]$ in order to save computation time. We found that regardless of the setting the resulting estimator was largely invariant to the denseness of the $\delta$-grid, and the two-stage method described above was suitable for most applications.

### 3.6 Oracle properties

First recommended by Fan and Li (2001), procedures with the oracle properties have since become the gold standard in the field of variable selection. The oracle properties are:

1. the procedure selects the truly nonzero $\boldsymbol{\beta}$s and excludes the truly zero $\boldsymbol{\beta}$s with probability tending to one as $n$ increases, and

2. the nonzero components of the estimator are asymptotically normal about the true nonzero $\boldsymbol{\beta}$ components with variance equal to the variance using OLS on the true submodel.

More formally, let $\mathcal{A}$ denote the set of indices corresponding to truly nonzero components of $\boldsymbol{\beta}$, and let $\hat{\boldsymbol{\beta}}$ be an estimator exhibiting the oracle properties. Then:

1. $P(\hat{\boldsymbol{\beta}}_{\mathcal{A}} \neq 0) \longrightarrow 1$ and $P(\hat{\boldsymbol{\beta}}_{\mathcal{A}^c} = 0) \longrightarrow 1$ as $n$ increases, and

2. $\hat{\boldsymbol{\beta}}_{\mathcal{A}} \xrightarrow{d} N\big(\boldsymbol{\beta}_{\mathcal{A}}, (\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-1}\sigma^2\big)$.

Let $z(\hat{\boldsymbol{\beta}}, t)$ be a thresholding function that sets all components of $\hat{\boldsymbol{\beta}}$ smaller than $t$ in (absolute value) to zero. That is, let

$$z(\hat{\boldsymbol{\beta}}, t) = \begin{cases} \hat{\beta}_k & \text{if } |\hat{\beta}_k| \geq t, \\ 0 & \text{if } |\hat{\beta}_k| < t. \end{cases}$$

*Theorem 2.*

For any threshold $t$ such that $0 < t < \min\big(|\boldsymbol{\beta}_{\mathcal{A}}|\big)$, there exists a pair $(\delta, \lambda)$ such that $z(\hat{\boldsymbol{\beta}}(\delta, \lambda, \tilde{\boldsymbol{\beta}}), t)$ selects the truly nonzero $\boldsymbol{\beta}$s and excludes the truly zero-valued $\boldsymbol{\beta}$s with probability tending to one. That is, there exists a pair such that:

$$\lim_{n\to\infty} P\bigg(\max\big(|\hat{\boldsymbol{\beta}}^{(1)}_{\mathcal{A}^c}|\big) < t < \min\big(|\hat{\boldsymbol{\beta}}^{(1)}_{\mathcal{A}}|\big)\bigg) = 1.$$

*Proof:* Let $t$ be such that $0 < t < \min\big(|\boldsymbol{\beta}_{\mathcal{A}}|\big)$. Let $i$ be the index corresponding to $\max\big(|\tilde{\boldsymbol{\beta}}_{\mathcal{A}^c}|\big)$ and let $j$ be the index corresponding to $\min\big(|\tilde{\boldsymbol{\beta}}_{\mathcal{A}}|\big)$. Consider the quotient

$$\frac{\frac{\lambda}{\delta^2} exp(-\tilde{\beta}_j^2/\delta^2)}{\frac{\lambda}{\delta^2} exp(-\tilde{\beta}_i^2/\delta^2)} = exp\bigg(\frac{\tilde{\beta}_i^2 - \tilde{\beta}_j^2}{\delta^2}\bigg).$$

Note that:

$$\lim_{n\to\infty} P\bigg(\lim_{\delta^2\to 0} exp\bigg(\frac{\tilde{\beta}_i^2 - \tilde{\beta}_j^2}{\delta^2}\bigg) = \infty\bigg) = 1,$$

because $\tilde{\boldsymbol{\beta}}$ is consistent which implies that the numerator of the exponent is asymptotically positive in probability. However, for fixed $\lambda$,

$$\lim_{n\to\infty} P\bigg(\lim_{\delta^2\to 0} \frac{\lambda}{\delta^2} exp(-\tilde{\beta}_j^2/\delta^2) = 0\bigg) = 1$$

and

$$\lim_{n\to\infty} P\bigg(\lim_{\delta^2\to 0} \frac{\lambda}{\delta^2} exp(-\tilde{\beta}_i^2/\delta^2) = 0\bigg) = 1,$$

so

$$\lim_{n\to\infty} P\bigg(\lim_{\delta^2\to 0} \mathbf{W}(\delta, \lambda, \tilde{\boldsymbol{\beta}}) = [\mathbf{0}]\bigg) = 1.$$

Instead of $\mathbf{W} = [\mathbf{0}]$ (yielding $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \mathbf{W})^{-1}\mathbf{X}'\mathbf{y} = \tilde{\boldsymbol{\beta}}$ and consequently no shrinkage), we would prefer $\mathbf{W}$ to have high values on diagonal elements corresponding to zero-valued $\boldsymbol{\beta}$ components, and near-zero values on diagonal elements corresponding to nonzero $\boldsymbol{\beta}$ components. Note that in the ideal (but impossible) scenario, we would have:

$$W_{p\times p} = \begin{pmatrix} w_{1,1} & 0 & \cdots & 0 \\ 0 & w_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{p,p} \end{pmatrix},$$

where $w_{\mathcal{A},\mathcal{A}} = 0$ and $w_{\mathcal{A}^c,\mathcal{A}^c} = \infty$. This would yield the solution

$$\hat{\boldsymbol{\beta}} = \begin{cases} \tilde{\beta}_j & \text{if } \ j \in \mathcal{A}, \\ 0 & \text{if } \ j \in \mathcal{A}^c, \end{cases}$$

which has the oracle properties. We show that $\min(\mathbf{W}_{\mathcal{A},\mathcal{A}})$ can be arbitrarily high in probability, while $\max(\mathbf{W}_{\mathcal{A}^c,\mathcal{A}^c})$ can be arbitrarily small (close to zero) in probability.

Choose $M$ large enough and $\epsilon > 0$ small enough that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \mathbf{W})^{-1}\mathbf{X}'\mathbf{y} \approx \begin{cases} \tilde{\beta}_j & \text{if } \ j \in \mathcal{A}, \\ 0 & \text{if } \ j \in \mathcal{A}^c \end{cases}$$

for

$$W_{p\times p} = \begin{pmatrix} w_{1,1} & 0 & \cdots & 0 \\ 0 & w_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{p,p} \end{pmatrix},$$

where $\max(w_{\mathcal{A},\mathcal{A}}) < \epsilon$ and $\min(w_{\mathcal{A}^c,\mathcal{A}^c}) > M$. Fix $\epsilon^* > 0$. From above, it is possible to choose an $n$ large enough and a $\delta^2$ small enough that

$$P\left(exp\big(\frac{\tilde{\beta}_i^2 - \tilde{\beta}_j^2}{\delta^2}\big) > \frac{M}{\epsilon}\right) > 1 - \epsilon^*.$$

For these values of $n$ and $\delta^2$, we shall define $\lambda = \dfrac{M+1}{\frac{1}{\delta^2}exp(-\tilde{\beta}_j^2/\delta^2)}$. Note that

$$\frac{\frac{\lambda}{\delta^2}exp(-\tilde{\beta}_j^2/\delta^2)}{\frac{\lambda}{\delta^2}exp(-\tilde{\beta}_i^2/\delta^2)} = exp\left(\frac{\tilde{\beta}_i^2 - \tilde{\beta}_j^2}{\delta^2}\right) > \frac{M}{\epsilon}$$

(with arbitrarily high probability), while $\frac{\lambda}{\delta^2}exp(-\tilde{\beta}_j^2/\delta^2) > M$, and $\frac{\lambda}{\delta^2}exp(-\tilde{\beta}_i^2/\delta^2) < \epsilon$ (again, with arbitrarily high probability). Therefore,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \mathbf{W})^{-1}\mathbf{X}'\mathbf{y} \approx \begin{cases} \tilde{\beta}_j & \text{if } j \in \mathcal{A}, \\ 0 & \text{if } j \in \mathcal{A}^c \end{cases}$$

and the approximation can be as exact as desired (with arbitrarily high probability). More specifically, the approximation can be exact enough that for any $t$ such that $0 < t < \min(|\boldsymbol{\beta}_{\mathcal{A}}|)$,

$$\max\big(|\hat{\boldsymbol{\beta}}^{(1)}_{\mathcal{A}^c}|\big) < t < \min\big(|\hat{\boldsymbol{\beta}}^{(1)}_{\mathcal{A}}|\big)$$

with arbitrarily high probability. Therefore, for any $t > 0$, there exists a pair $(\delta, \lambda)$ such that $z(\hat{\boldsymbol{\beta}}(\delta, \lambda, \tilde{\boldsymbol{\beta}}), t)$ selects the truly nonzero $\boldsymbol{\beta}$s and excludes the truly zero-valued $\boldsymbol{\beta}$s with probability tending to one. ∎

We have proved that it is possible to shrink estimates in one step such that the largest near-zero coefficient is arbitrarily small, and the smallest "nonzero" coefficient is arbitrarily close to the corresponding the OLS-estimated coefficient. It is important however, that we extend this theory to $k$-step estimator.

*Conjecture 1.*

Given a data set with $n$ sufficiently large and fixed $p$, assume we have a pair $(\delta, \lambda)$ meeting the criteria of Theorem 2. Then $z(\hat{\boldsymbol{\beta}}^{(1)}, \delta) \xrightarrow{p} \boldsymbol{\beta}$.

*Proof:* (Outline) We know that $\tilde{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$, by the consistency of the OLS estimator. This allowed us to establish that there exists a pair $(\delta, \lambda)$ such that $z(\hat{\boldsymbol{\beta}}^{(1)}, \delta)_{\mathcal{A}^c} = 0$ and

$z(\hat{\boldsymbol{\beta}}^{(1)}, \delta)_{\mathcal{A}} - \tilde{\boldsymbol{\beta}}_{\mathcal{A}}$ was arbitrarily small, both with probability one (in the limit). Assume that we have such a pair $(\delta, \lambda)$. Then

$$z(\hat{\boldsymbol{\beta}}^{(1)}, \delta) \xrightarrow{p} \begin{cases} \text{plim} \;\; \tilde{\beta}_j & \text{if} \;\; j \in \mathcal{A}, \\ 0 & \text{if} \;\; j \in \mathcal{A}^c \end{cases} = \boldsymbol{\beta}.$$

∎

*Conjecture 2.*

Let $\mathbf{X}$ be a data set with $n$ sufficiently large and fixed $p$. We conjecture that for any $k \in \mathbb{N}$, there exists a pair $(\delta, \lambda)$ such that $z(\hat{\boldsymbol{\beta}}^{(k)}, \delta) \xrightarrow{p} \boldsymbol{\beta}$, where the zero-thresholding is applied after the $k$-step shrinkage.

*Proof:* (Outline) Conjecture 1 takes care of the $k = 1$ case. For $k = 2$, recall that by Theorem 2, there exists a pair $(\delta, \lambda)$ such that $z(\hat{\boldsymbol{\beta}}^{(1)}, \delta)_{\mathcal{A}^c} = 0$ and $z(\hat{\boldsymbol{\beta}}^{(1)}, \delta)_{\mathcal{A}} - \tilde{\boldsymbol{\beta}}_{\mathcal{A}}$ was arbitrarily small, both with probability one (in the limit). In fact, even before the zero-thresholding, $\hat{\boldsymbol{\beta}}^{(1)}_{\mathcal{A}^c}$ was arbitrarily close to zero (in probability) and $\hat{\boldsymbol{\beta}}^{(1)}_{\mathcal{A}}$ was arbitrarily close to the $\tilde{\boldsymbol{\beta}}_{\mathcal{A}}$ (in probability). Recall that we pass $\hat{\boldsymbol{\beta}}^{(1)}$ as an argument to the $\mathbf{W}$-matrix of $\hat{\boldsymbol{\beta}}^{(2)}$. It can be shown (in a way similar to the proof of Theorem 2) that $\hat{\boldsymbol{\beta}}^{(2)}_{\mathcal{A}}$ is very close to $\hat{\boldsymbol{\beta}}^{(1)}_{\mathcal{A}}$ (in probability) and $\hat{\boldsymbol{\beta}}^{(2)}_{\mathcal{A}^c}$ is very close to zero (in probability). By an induction-like argument, we may reasonably expect the same thing for any $k$. Thus, there will exist some $t$ between $\max\big(|\hat{\boldsymbol{\beta}}^{(k)}_{\mathcal{A}^c}|\big)$ and $\min\big(|\hat{\boldsymbol{\beta}}^{(k)}_{\mathcal{A}}|\big)$, and thresholding using this $t$ results in an estimator arbitrarily close to the OLS estimates in its nonzero components and exactly equal to zero for the remaining components (with arbitrarily high probability). This estimator will thus be arbitrarily close to $\boldsymbol{\beta}$ in probability. ∎

*Conjecture 3.*

The proposed method of fixing $\lambda$ and finding $\delta^2$ using $t = \delta$ and BIC will obtain a pair $(\lambda, \delta)$ meeting the criteria of Theorem 2 and Conjecture 2 with probability tending to one as $n$ increases.

*Proof:* (Outline) We provide a heuristic argument for this conjecture. BIC has been shown to be asymptotically consistent in a number of penalized likelihood model selection settings as long as the true sparse model is among the candidate models (e.g. Dicker (2010), Zou, Hastie, and Tibshirani (2007), Wang, Li, and Tsai (2007), and their references). For large enough $n$, by searching over a range of zero-thresholding $\delta$ values, we essentially guarantee that the true model will be among the candidate models. Thus, it is reasonable to suspect that BIC will be consistent for model selection in this setting. Note that BIC being consistent in this setting is equivalent to the pair $(\delta_{min(BIC)}, \lambda)$ meeting the criteria of Theorem 2 and Conjecture 2. ∎

The implication of Theorem 2 and Conjectures 1-3 is that the estimator obtained using the described thresholding rule, $z(\hat{\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}}, \delta, \lambda)^{(k)}, \delta)$, will possess the oracle properties for any $k$. This is due to the two-step nature of the ShrinkSet estimator, which shrinks the near-zero estimates very close to zero and then sets them to zero. The coefficients far from zero experience little shrinkage and mimic their corresponding OLS coefficients. We state the oracle property implication formally in Theorem 3.

*Theorem 3.*

If Conjectures 2-3 hold, then the proposed $k$-step ShrinkSet estimator possess the oracle properties for any $k \in \mathcal{N}$.

*Proof:* By Conjecture 3, the proposed method of fixing $\lambda$ and finding $\delta^2$ using $t = \delta$ and BIC will obtain a pair $(\lambda, \delta)$ meeting the criteria of Conjecture 2. By Conjecture 2, $z(\hat{\boldsymbol{\beta}}^{(k)}, \delta) \xrightarrow{p} \boldsymbol{\beta}$ for any $k \in \mathcal{N}$.

In the nonzero components of $\boldsymbol{\beta}$, $z(\hat{\boldsymbol{\beta}}^{(k)}, \delta)$ converges in probability to $\boldsymbol{\beta}$ by virtue of $\hat{\boldsymbol{\beta}}^{(k)}$'s arbitrary nearness to $\tilde{\boldsymbol{\beta}}$. So, asymptotically, $\hat{\boldsymbol{\beta}}^{(k)}_{\mathcal{A}}$ will have the same distribution as the OLS estimator: it will be normally distributed with mean $\boldsymbol{\beta}_{\mathcal{A}}$ and variance $(\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-1}\sigma^2$.

In the zero-valued components of $\boldsymbol{\beta}$, $z(\hat{\boldsymbol{\beta}}^{(k)}, \delta)$ converges in probability to $\boldsymbol{\beta}$ by virtue of the thresholding rule. Since $P\big(|\hat{\beta}_i^{(k)}| < \delta\big)$, where $i$ is the index corresponding to $\max\big(|\hat{\boldsymbol{\beta}}^{(k)}_{\mathcal{A}^c}|\big)$, approaches one asymptotically, the ShrinkSet estimator selects the truly nonzero $\boldsymbol{\beta}$s and excludes the truly zero-valued $\boldsymbol{\beta}$s with probability tending to one. ∎

In the simulation section, we show that $\delta_{min(BIC)}$ tends to zero in probability, but not as fast as the largest should-be-zero coefficients. Furthermore, we demonstrate that as $n$ increases, $P\big(\max(|\hat{\boldsymbol{\beta}}^{(1)}_{\mathcal{A}^c}|) < \delta\big)$ approaches 1, thus showing that the requirements of Theorem 2 have been met. We then demonstrate that for high $n$, the one-step ShrinkSet procedure results in approximately normal estimators, with mean equal to the true $\boldsymbol{\beta}$ value and variance approximately equal to that of the OLS estimator. This further establishes the tenability of Conjectures 1-3, which, when combined with Theorem 2, imply that the ShrinkSet estimators will possess the oracle properties.

### 3.7 Application when $p \gg n$

An important application of variable selection procedures is when there are many more variables measured on each individual than there are individuals in the study, the so-called "large $p$, small $n$" problem. For example, genomic data can have thousands of variables collected on a few dozen patients. We describe two possible implementations of our approach and compare these to two LASSO-based models built using `glmnet` (Friedman, Hastie, and Tibshirani 2010). We do not include SELO in our comparison, because Dicker (2010) has not extended the results to the $p \gg n$ setting. In both ShrinkSet implementations, and for the first LASSO model, we select the estimator which minimizes a BIC-type parameter (called "extended BIC" (EBIC)), adapted for high-dimensional analysis. Following Lian

(2010) closely, we minimize

$$\text{EBIC}(\hat{\boldsymbol{\beta}}) = log\left[\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{n}\right] + \frac{d \cdot log(n)\sqrt{p}}{n}.$$

Although Lian (2010) employs EBIC for high-dimensional varying coefficient models based on B-spline basis expansions, our results suggest that their method can be used effectively in simpler settings. For comparison, we also include a LASSO model selected by minimum cross-validation error. We shall include two simulation scenarios: one with 7 nonzero coefficients out of $p = 10,000$ with moderate covariance, and one with 16 nonzero coefficients out of $p = 1,000$ with very high covariance.

*Implementation 1: Reducing* $\mathbf{X}$ *matrix in terms of correlation with* $\mathbf{y}$

Because our estimate is based on the $\mathbf{X}'\mathbf{X}$ matrix, it is impossible to reliably estimate $p > n$ truly nonzero $\boldsymbol{\beta}$s. However, if we constrain some of these estimates to be zero (which our method does automatically), then we can produce reliable estimates of some subset of less than $n$ coefficients. We observed that choosing $p < .85n$ columns of $\mathbf{X}$ was typically more effective than choosing $p < n$. Therefore, we begin by selecting one at a time the $.85n$ columns of $\mathbf{X}$ most highly correlated with $\mathbf{y}$. We redefine $\mathbf{X}$ as only those columns of $\mathbf{X}$ and run our algorithm. This is very similar to the sure independence screening procedure recommended by Fan and Lv (2008). In the results section, we call this the correlation method.

*Implementation 2: Reducing* $\mathbf{X}$ *matrix based on LASSO output*

A dimension-reduction alternative is to use output from LASSO itself. We overfit a LASSO model using glmnet with small $\lambda$ and keep only those columns of $\mathbf{X}$ corresponding to nonzero $\hat{\boldsymbol{\beta}}$s. If this is larger than $.85n$, we reduce further in terms of correlation with $\mathbf{y}$, until we have less than $p \leq .85n$ We then run our algorithm to shrink "nonsignificant" $\hat{\boldsymbol{\beta}}$s to zero. We call this the hybrid method.

*LASSO with $\lambda$ selected by EBIC*

Using `glmnet` (Friedman et al. 2010), we obtain 100 different $\hat{\boldsymbol{\beta}}$s, each vector corresponding to a different value of $\lambda$. We choose the model with the minimum EBIC.

*LASSO with $\lambda$ selected by cross-validation*

Using `cv.glmnet` (Friedman et al. 2010), we find the value of $\lambda$ associated with the minimum cross-validation error. We then fit a LASSO model by passing this $\lambda$ value as an argument to `glmnet` (Friedman et al. 2010). We use leave-one-out cross-validation in the $p = 1,000$ case and 10-fold cross-validation in the $p = 10,000$ case.

### 3.8 Other applications

*Robust regression*

The least squares estimate is not robust against outliers, and nor is our proposed estimator, because it relies on penalized least squares. However, if we use an outlier-resistant loss function, such as Huber's $\psi$ function (Huber 1964), and minimize

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \psi(|y_i - \mathbf{x}_i'\boldsymbol{\beta}|) + \lambda \sum_{j=1}^{p} \Big(1 - exp\big(-\frac{\beta_j^2}{\delta^2}\big)\Big),$$

as suggested by Fan and Li (2001) (but with a different penalty), the result will be a robust estimator, with nonsignificant terms shrunk to zero after the zero-thresholding step. We suspect that a solution for this optimization is possible using iteratively reweighted least squares, as shown by Fan and Li (2001), but with appropriate modification of the weights.

*Link functions*

Our method can also be generalized so as to accommodate different link functions such as log, logit, and probit. For example, modifying a result by Zou (2006), the penalized logistic

maximum likelihood estimate would be

$$\hat{\boldsymbol{\beta}}_{logistic} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \Big( y_i(\mathbf{x}_i'\boldsymbol{\beta}) - log(1 + exp(\mathbf{x}_i'\boldsymbol{\beta})) \Big) - \lambda \sum_{j=1}^{p} \Big( 1 - exp\big( - \frac{\beta_j^2}{\delta^2} \big) \Big),$$

which is solvable using the Newton-Raphson algorithm. Following Hastie et al. (2009), we express the first derivative of the penalized log-likelihood as

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{X}'(\mathbf{y} - \mathbf{p}) - 2\mathbf{W}(\boldsymbol{\beta})\boldsymbol{\beta},$$

where $\mathbf{p}$ is the vector of fitted probabilities and $\mathbf{W}(\boldsymbol{\beta})$ is defined as was $\mathbf{W}$ before, except $\mathbf{W}(\boldsymbol{\beta})$ is in terms of the actual parameter $\boldsymbol{\beta}$ instead of an estimate thereof. Similarly,

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -\mathbf{X}'(\boldsymbol{V} + \boldsymbol{Z})\mathbf{X},$$

where $\boldsymbol{V}$ corresponds to a diagonal matrix with entries $\mathbf{p}(1 - \mathbf{p})$ and

$$\boldsymbol{Z} = 2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \frac{\partial(\mathbf{W}(\boldsymbol{\beta})\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

For a given estimate of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, we can substitute $\hat{\boldsymbol{Z}} = 2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{W}(\hat{\boldsymbol{\beta}}))(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ for $\boldsymbol{Z}$. Starting with the regular logistic regression estimate or even a vector of zeros (call it $\boldsymbol{\beta}^{old}$), we iterate

$$\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^{old} - \left( \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right)^{-1} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

until convergence, with derivatives being evaluated at $\boldsymbol{\beta}^{old}$ and $\hat{\boldsymbol{Z}}$ being substituted for $\boldsymbol{Z}$. As Hastie et al. (2009) have shown, this is equivalent to iteratively reweighted least squares (IRWLS). Because $\widehat{\mathbf{W}}$ is fully specified by $\boldsymbol{\beta}^{old}$, for fixed $\lambda$ and $\delta$ our method requires only the iterative step inherent to logistic regression/IRWLS. This is significant because other methods (e.g. SELO) require additional iteration (in the form of the coordinate descent algorithm) for calculating $\boldsymbol{\beta}^{new}$ given $\boldsymbol{\beta}^{old}$. Future research may focus on establishing the oracle properties for these generalized methods.

CHAPTER 4

# SIMULATIONS AND RESULTS

## 4.1 Estimator performance

*Setup*

We use the same simulation parameters as in Dicker (2010). That is, we set $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{X} \sim N(0, \boldsymbol{\Sigma})$, $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$; we took $\boldsymbol{\Sigma} = (\sigma_{ij})$, where $\sigma_{ij} = .5^{|i-j|}$; each simulation consisted of the generation of 1,000 independent data sets and parameter estimation for each data set; four simulations were conducted: all combinations of $n \in \{50, 100\}$ and $\sigma^2 \in \{1, 9\}$. Metrics observed at each iteration of the simulations include:

1. whether the correct set of nonzero coefficients was selected,
2. the average model size (number of nonzero coefficients),
3. the mean squared error, $\dfrac{||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2}{p}$, and
4. the model error, $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \boldsymbol{\Sigma} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

All simulations in this paper were performed on a Mac OS X (Version 10.5.8) with a 2.66 GHz Intel Core 2 Duo processor.

One drawback we see in some of the literature on variable selection is a neglect to fairly compare the proposed method to OLS. For example, Dicker (2010) asserts that OLS does not perform variable selection and proceeds to compare its performance in model size to those of variable selection methods without any attempt to pare down the OLS model. A fairer approach would be to include some automated OLS-based variable selection method, like a stepwise selection procedure or even regular OLS with all nonsignificant ($\alpha = 0.05$,

for example) estimates set to zero. In the next section we present our proposed method's results, along with those of Dicker (2010) for reference. We also include OLS results where nonsignificant estimates were estimated to be zero.

*Results*

We report the performance of the $k$-step ShrinkSet estimators, for $k \in \{1, 5, 10\}$, compared with the SELO estimator and the OLS estimator. SELO results in Tables 1-4 were obtained from Table 2 in Dicker (2010).

Table 4.1: Simulation study results ($\sigma^2 = 1$, $n = 50$)

| | Correct model | Model size | MSE | Model error |
|---|---|---|---|---|
| | Mean | Mean | Mean | Mean |
| $\text{ShrinkSet}_1$ | 0.896 | 3.119 | 0.013 | 0.081 |
| $\text{ShrinkSet}_5$ | 0.860 | 3.158 | 0.014 | 0.086 |
| $\text{ShrinkSet}_{10}$ | 0.855 | 3.163 | 0.014 | 0.087 |
| $\text{ShrinkSet}_{\lambda=0}$ | 0.785 | 3.263 | 0.020 | 0.118 |
| OLS | 0.965 | 3.038 | 0.015 | 0.103 |
| SELO | 0.866 | 3.157 | 0.014 | 0.087 |

Table 4.2: Simulation study results ($\sigma^2 = 1$, $n = 100$)

| | Correct model | Model size | MSE | Model error |
|---|---|---|---|---|
| | Mean | Mean | Mean | Mean |
| $\text{ShrinkSet}_1$ | 0.972 | 3.035 | 0.005 | 0.035 |
| $\text{ShrinkSet}_5$ | 0.912 | 3.097 | 0.006 | 0.039 |
| $\text{ShrinkSet}_{10}$ | 0.911 | 3.098 | 0.006 | 0.039 |
| $\text{ShrinkSet}_{\lambda=0}$ | 0.864 | 3.153 | 0.008 | 0.051 |
| OLS | 0.975 | 3.026 | 0.007 | 0.047 |
| SELO | 0.914 | 3.091 | 0.006 | 0.037 |

These results agree with our earlier assertion that the five-step estimator is very close to the true minimum of $Q(\boldsymbol{\beta})$. Specifically, we see very little change in performance between the $k$-step estimators for $k = 5$ and $k = 10$. However, we see that the one-step ShrinkSet estimator is the best in low-variance settings and the worst in high variance settings. We attribute

Table 4.3: Simulation study results ($\sigma^2 = 9$, $n = 50$)

| | Correct model | Model size | MSE | Model error |
|---|---|---|---|---|
| | Mean | Mean | Mean | Mean |
| $\text{ShrinkSet}_1$ | 0.560 | 2.891 | 0.252 | 1.378 |
| $\text{ShrinkSet}_5$ | 0.597 | 2.897 | 0.242 | 1.305 |
| $\text{ShrinkSet}_{10}$ | 0.602 | 2.898 | 0.241 | 1.301 |
| $\text{ShrinkSet}_{\lambda=0}$ | 0.565 | 3.025 | 0.254 | 1.524 |
| OLS | 0.287 | 2.132 | 0.419 | 3.188 |
| SELO | 0.605 | 2.913 | 0.240 | 1.310 |

Table 4.4: Simulation study results ($\sigma^2 = 9$, $n = 100$)

| | Correct model | Model size | MSE | Model error |
|---|---|---|---|---|
| | Mean | Mean | Mean | Mean |
| $\text{ShrinkSet}_1$ | 0.852 | 3.059 | 0.075 | 0.436 |
| $\text{ShrinkSet}_5$ | 0.864 | 3.066 | 0.071 | 0.419 |
| $\text{ShrinkSet}_{10}$ | 0.867 | 3.065 | 0.070 | 0.416 |
| $\text{ShrinkSet}_{\lambda=0}$ | 0.820 | 3.148 | 0.084 | 0.522 |
| OLS | 0.780 | 2.839 | 0.116 | 0.804 |
| SELO | 0.879 | 3.061 | 0.070 | 0.408 |

the varying success of the one-step ShrinkSet estimator to its direct reliance on the OLS estimator. We find it noteworthy that Dicker (2010) observed a similar phenomenon for the SELO one-step estimator. In all scenarios, the five-step and ten-step ShrinkSet estimators tend to be very similar to the SELO estimator. Importantly, these ShrinkSet estimators outperform the $\lambda = 0$ estimator across the board, which shows that the shrinkage step is crucial to obtaining good results.

## 4.2 Validation of Conjecture 1

*Setup*

We are also interested in validating Conjecture 1 through simulation. Particularly, we show that as $n$ increases, the $\delta$ selected by BIC tends to zero in probability, but no faster than the largest $\hat{\boldsymbol{\beta}}_{\mathcal{A}^c}$. We generated data and found the associated one-step estimator as in the

previous simulation setup for all pairwise combinations of $n \in \{50, 500, 5000, 50000\}$ and $\sigma^2 \in \{0.01, 0.1, 1, 10\}$. Each row in Table 5 corresponds to the results across 100 iterations. The value $\bar{\delta}$ refers to the mean of the 100 values for $\delta_{min(BIC)}$, while $P(\delta > |\hat{\boldsymbol{\beta}}_{\mathcal{A}^c}|)$ refers to the proportion (out of 100) of $\delta_{min(BIC)}$ values which exceeded the largest $|\hat{\boldsymbol{\beta}}_{\mathcal{A}^c}|$ after one-step shrinkage.

*Results*

Table 4.5: Simulation study results: Responses of $\bar{\delta}$ and $P(\delta > |\hat{\boldsymbol{\beta}}_{\mathcal{A}^c}|)$ to sample size and variance. Let $E$ be the event that $\delta > |\hat{\boldsymbol{\beta}}_{\mathcal{A}^c}|$.

| | $\sigma^2 = .01$ | | $\sigma^2 = .1$ | | $\sigma^2 = 1$ | | $\sigma^2 = 10$ | |
|---|---|---|---|---|---|---|---|---|
| | $\bar{\delta}$ | $P(E)$ | $\bar{\delta}$ | $P(E)$ | $\bar{\delta}$ | $P(E)$ | $\bar{\delta}$ | $P(E)$ |
| $n = 50$ | 0.026 | 0.92 | 0.080 | 0.92 | 0.222 | 0.92 | 0.459 | 0.87 |
| $n = 500$ | 0.010 | 1.00 | 0.027 | 0.97 | 0.079 | 0.94 | 0.246 | 0.97 |
| $n = 5000$ | 0.007 | 1.00 | 0.009 | 1.00 | 0.027 | 0.97 | 0.081 | 1.00 |
| $n = 50000$ | 0.002 | 1.00 | 0.007 | 1.00 | 0.009 | 1.00 | 0.024 | 1.00 |

We see that in every case, the probability that $\delta$ exceeds $\hat{\boldsymbol{\beta}}_{\mathcal{A}^c}$ tends to one as $n$ increases. Furthermore, $\delta$ gets smaller as $n$ increases, as demonstrated graphically below (Fig. 4.1). In addition, we see that $\bar{\delta}$ remains relatively constant across each diagonal, indicating that a ten-fold increase in $n$ is approximately equivalent to a ten-fold decrease in $\sigma^2$ in terms of its effect on $\bar{\delta}$. Altogether, these results give strong support to Conjecture 1.

*Asymptotic normality*

We now demonstrate that for nonzero coefficients, the one-step ShrinkSet estimator has approximately the same asymptotic distribution as the OLS estimator. From Figure 4.1, we see that at $(n = 50, \sigma^2 = 10)$, we have not yet reached what Casella and Berger (2002) have dubbed "asymptopia," that is, the asymptotic properties of our estimator have not yet kicked in. We expect, therefore, that our estimates for $\boldsymbol{\beta}$ will be nonnormal. By contrast, at $(n = 5000, \sigma^2 = 0.1)$ or $(n = 500, \sigma^2 = 0.01)$, the asymptotic properties appear to be in near-full effect, and $\hat{\boldsymbol{\beta}}$ should be nearly normal. Figure 4.2 contains the empirical marginal

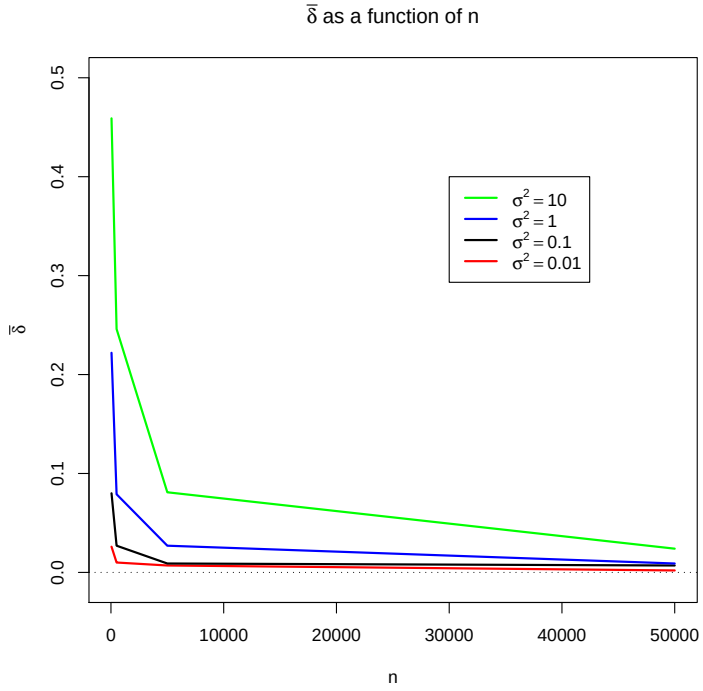


Figure 4.1: For all values $\sigma^2$, the average of $\delta$ values associated with the minimum BIC tends to zero.

distributions (based on 1000 iterations) of the second coefficient of the OLS and one-step ShrinkSet estimators for the $(n = 50, \sigma^2 = 10)$ and $(n = 5000, \sigma^2 = 0.01)$ situations. We chose to look at $\tilde{\beta}_2$ and $\hat{\beta}_2$ because $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and the second element is the smallest nonzero element, so it would be the hardest to distinguish from zero. In addition, we performed Anderson-Darling and Shapiro-Wilks tests for normality on the ShrinkSet estimators in the $(n = 5000, \sigma^2 = 0.01)$ case. Both $p$-values were nonsignificant, indicating that the normality conclusion is reasonable for the ShrinkSet estimator.

## 4.3 Simulations for large $p$, small $n$ setting

*Simulation 1 setup*

We shall investigate model performance for $n \in \{50, 100, 200\}$ with $p = 10,000$ so as to be consistent with the size of typical genomic data sets. Using the same underlying covariance

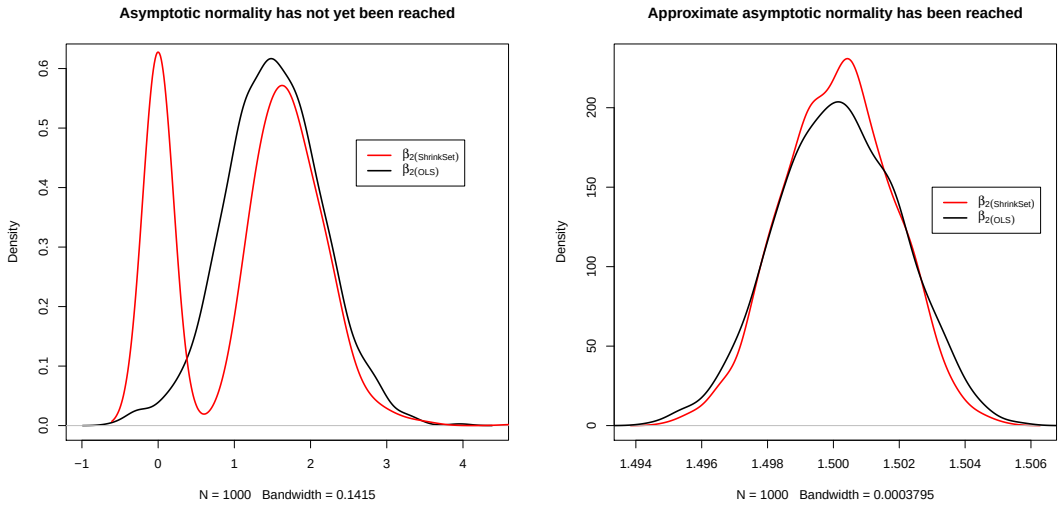


Figure 4.2: As expected, for sufficiently high $n$, the distribution of the ShrinkSet estimator is similar to that of the OLS estimator (left panel: $n = 50$, $\sigma^2 = 10$; right panel: $n = 5000$, $\sigma^2 = 0.01$.)

structure as in the previous simulations, we define

$$\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 2.5, 2, 0, 0, 0, \ldots, 0, 0, 0, 3, 4)^T \in \mathbb{R}^{10000}.$$

(Note that the true model contains 7 nonzero coefficients.) We use $\sigma^2 = 1$ and generate $\mathbf{y}$ and $\mathbf{X}$ with appropriate dimensions. All simulations are the result of 100 iterations.

*Results for Simulation 1*

Our results indicate that the hybrid method outperforms LASSO for all sample sizes, and that the correlation method outperforms LASSO in relatively high-$n$ settings. More specifically, for all $n$ studied, the hybrid method outperforms both LASSOs in every category. And by $n = 100$, the correlation method is clearly better than both LASSO implementations. For $n = 200$, only the correlation method outperforms the hybrid method, but the difference is marginal. Overall, the hybrid method is probably best, because it is vastly superior to LASSO in all settings, and only slightly worse than the correlation method in the high-$n$ setting but clearly better everywhere else.

Table 4.6: Simulation study results ($\sigma^2 = 1$, $n = 50$, $p = 10,000$)

| | Correct model | Model size | MSE | Model error |
|---|---|---|---|---|
| | Mean | Median | Mean | Mean |
| Correlation | 0.03 | 6 | 0.0040 | 47.7 |
| Hybrid | 0.32 | 7 | 0.0012 | 12.4 |
| $\text{LASSO}_{EBIC}$ | 0.02 | 0 | 0.0041 | 62.1 |
| $\text{LASSO}_{CV}$ | 0.00 | 39 | 0.0014 | 18.5 |

Table 4.7: Simulation study results ($\sigma^2 = 1$, $n = 100$, $p = 10,000$)

| | Correct model | Model size | MSE | Model error |
|---|---|---|---|---|
| | Mean | Median | Mean | Mean |
| Correlation | 0.71 | 7 | 0.00018 | 1.42 |
| Hybrid | 0.90 | 7 | 0.00001 | 0.10 |
| $\text{LASSO}_{EBIC}$ | 0.21 | 8 | 0.00014 | 1.95 |
| $\text{LASSO}_{CV}$ | 0.00 | 46 | 0.00008 | 1.02 |

Table 4.8: Simulation study results ($\sigma^2 = 1$, $n = 200$, $p = 10,000$)

| | Correct model | Model size | MSE | Model error |
|---|---|---|---|---|
| | Mean | Median | Mean | Mean |
| Correlation | 0.99 | 7 | 0.0000051 | 0.039 |
| Hybrid | 0.97 | 7 | 0.0000054 | 0.042 |
| $\text{LASSO}_{EBIC}$ | 0.52 | 7 | 0.0000383 | 0.541 |
| $\text{LASSO}_{CV}$ | 0.00 | 46 | 0.0000271 | 0.335 |

*Simulation 2 setup*

It is also of interest to assess performance of our methods under a more complicated data scenario. For the second simulation study, we introduce a more complicated covariance structure in which the 16 nonzero effects are grouped in clusters of 3-4 with high within-cluster correlation. We define

$$\boldsymbol{\beta} = (3, 4, 5, \ldots, 4.5, 6, 3, \ldots, 2, 3, 2, \ldots, 2, 1.5, 2, \ldots, 3.5, 4, 4, 5)^T \in \mathbb{R}^{1000},$$

wherein each ellipsis represents 246 zeros. To induce high covariance, we used $\boldsymbol{\Sigma} = (\sigma_{ij})$, where $\sigma_{ij} = .9^{|i-j|}$. We investigate model performance for $n \in \{25, 50, 100, 200\}$. Again,

we use $\sigma^2 = 1$ and generate $\mathbf{y}$ and $\mathbf{X}$ with appropriate dimensions. All simulations are the result of 100 iterations.

*Results for Simulation 2*

These results indicate that the hybrid method outperforms all others for all but the $n = 25$ scenario, where no method does well. The hybrid method is exceptionally adept at uncovering the correct sparse model. The correlation method has poor MSE and model error at all levels of $n$ included, but by $n = 200$, it outperforms both LASSOs in terms of selecting the true model. Both implementations of LASSO always achieve low MSE, but in general the hybrid method's MSE is still lower. When compared to each other, $\text{LASSO}_{CV}$ tends to have lower model error and $\text{LASSO}_{EBIC}$ tends select the correct model more frequently. For the most difficult data scenario ($n = 25$), no method could uncover the true model, and all suffered from severe over- or under-fitting.

Table 4.9: Simulation study results ($\sigma^2 = 1$, $n = 25$, $p = 1,000$)

| | Correct model | Model size | MSE | Model error |
|---|---|---|---|---|
| | Mean | Median | Mean | Mean |
| Correlation | 0.00 | 2 | 0.42 | 494 |
| Hybrid | 0.00 | 8 | 0.31 | 260 |
| $\text{LASSO}_{EBIC}$ | 0.00 | 0 | 0.21 | 576 |
| $\text{LASSO}_{CV}$ | 0.00 | 21 | 0.22 | 265 |

Table 4.10: Simulation study results ($\sigma^2 = 1$, $n = 50$, $p = 1,000$)

| | Correct model | Model size | MSE | Model error |
|---|---|---|---|---|
| | Mean | Median | Mean | Mean |
| Correlation | 0.00 | 4 | 0.339 | 247.1 |
| Hybrid | 0.17 | 15 | 0.037 | 12.1 |
| $\text{LASSO}_{EBIC}$ | 0.00 | 17 | 0.095 | 182.5 |
| $\text{LASSO}_{CV}$ | 0.00 | 36 | 0.032 | 12.8 |

Table 4.11: Simulation study results ($\sigma^2 = 1$, $n = 100$, $p = 1,000$)

| | Correct model | Model size | MSE | Model error |
|---|---|---|---|---|
| | Mean | Median | Mean | Mean |
| Correlation | 0.02 | 11 | 0.168 | 59.87 |
| Hybrid | 0.89 | 16 | 0.002 | 0.23 |
| $\text{LASSO}_{EBIC}$ | 0.02 | 21 | 0.004 | 1.61 |
| $\text{LASSO}_{CV}$ | 0.00 | 36 | 0.003 | 0.83 |

Table 4.12: Simulation study results ($\sigma^2 = 1$, $n = 200$, $p = 1,000$)

| | Correct model | Model size | MSE | Model error |
|---|---|---|---|---|
| | Mean | Median | Mean | Mean |
| Correlation | 0.35 | 16 | 0.0170 | 3.725 |
| Hybrid | 1.00 | 16 | 0.0006 | 0.089 |
| $\text{LASSO}_{EBIC}$ | 0.03 | 19 | 0.0014 | 0.574 |
| $\text{LASSO}_{CV}$ | 0.00 | 30 | 0.0011 | 0.326 |

CHAPTER 5

# DATA ANALYSIS

## Analysis of HIV data set

Rhee, Taylor, Wadhera, Ben-Hur, Brutlag, and Shafer (2006) describe a publicly available data set on HIV-1 drug resistance and codon mutation. The data consist of mutation information for some 100 protease codons and a continuous response $IC_{50}$, a measure of HIV-1 drug resistance. We were interested in learning which codon mutations were related to resistance to the drug Amprenavir, a protease inhibitor. This data set was also analyzed by Dicker (2010), and we follow their example of removing codons with fewer than three observed mutations and taking the log-transform of $IC_{50}$. We present the results of our method in connection with those of SELO, OLS (as used in the simulation setting), and several other variable selection methods in terms of model size and $R^2$ as a percentage of the $R^2$ value for complete OLS. One additional method presented is $\text{ShrinkSet}_2$, which was was obtained by removing the nonsignificant terms from the ShrinkSet estimator.

| | Model size | $R^2/R^2_{complete}$ |
|---|---|---|
| OLS ($\alpha = .10$) | 30 | 0.985 |
| OLS ($\alpha = .05$) | 22 | 0.959 |
| OLS ($\alpha = .01$) | 17 | 0.949 |
| ShrinkSet | 28 | 0.979 |
| $\text{ShrinkSet}_2$ | 22 | 0.972 |
| SELO | 16 | 0.958 |
| LASSO | 32 | 0.959 |
| Adaptive LASSO | 20 | 0.956 |
| SCAD | 33 | 0.972 |

Table 5.1: Results of various variable selection procedures on HIV data set from Rhee et al.; $\text{ShrinkSet}_2$ was obtained by removing the nonsignificant terms from the model in Table 5.2.

Table 5.2: Codons selected by ShrinkSet method

| Codon | Point estimate | Std. error | Approx. $p$-value |
|---|---|---|---|
| P10 | 0.6605 | 0.0786 | 0.0000 |
| P11 | 0.4192 | 0.2258 | 0.0636 |
| P22 | 0.2738 | 0.4053 | 0.4990 |
| P24 | 0.3252 | 0.1471 | 0.0270 |
| P30 | 0.8879 | 0.1384 | 0.0000 |
| P32 | 0.9217 | 0.1786 | 0.0000 |
| P33 | 0.6724 | 0.0995 | 0.0000 |
| P34 | 0.4874 | 0.2220 | 0.0281 |
| P37 | -0.2800 | 0.0644 | 0.0000 |
| P38 | -0.3012 | 0.4083 | 0.4603 |
| P45 | -0.3657 | 0.1705 | 0.0314 |
| P46 | 0.4853 | 0.0743 | 0.0000 |
| P47 | 1.0053 | 0.2222 | 0.0000 |
| P48 | 0.5978 | 0.1605 | 0.0002 |
| P50 | 0.5147 | 0.1546 | 0.0009 |
| P54 | 0.5387 | 0.0860 | 0.0000 |
| P64 | -0.3951 | 0.0745 | 0.0000 |
| P65 | -0.6255 | 0.2787 | 0.0250 |
| P66 | -0.2858 | 0.1691 | 0.0908 |
| P67 | 0.2835 | 0.1954 | 0.1460 |
| P71 | -0.3067 | 0.0765 | 0.0001 |
| P76 | 1.2616 | 0.1596 | 0.0000 |
| P83 | -0.8397 | 0.4863 | 0.0840 |
| P84 | 0.8928 | 0.0893 | 0.0000 |
| P88 | -1.3109 | 0.1151 | 0.0000 |
| P89 | 0.3378 | 0.1300 | 0.0094 |
| P90 | 0.6779 | 0.0790 | 0.0000 |
| P93 | -0.3147 | 0.0622 | 0.0000 |

The ShrinkSet procedure explained more variation in $\mathbf{y}$ than all but OLS ($\alpha = .10$), yet it used fewer terms than OLS ($\alpha = .10$), LASSO, and SCAD. The reasonable but slightly ad hoc approach ShrinkSet$_2$ was arguably the best of all, explaining the same amount of variation as SCAD, but with 11 fewer terms. The following summary table (5.2) gives the 28 codons selected by the ShrinkSet method. Interestingly, 16 of the 17 most significant $p$-values correspond to the 16 codons selected by SELO. The notable exception is codon 37, which SELO did not select. Importantly, ShrinkSet selected mutations at codons 10, 32, 46,

47, 50, 54, 76, 84, and 90, and these are known to be associated with Amprenavir resistance (Johnson, Brun-Vézinet, Clotet, Günthard, Kuritzkes, Pillay, Schapiro, and Richman 2010).

### Analysis of salary data set

Fan and Peng (2004) describe a salary data set that was used in a gender discrimination lawsuit. The accusation was that female employees at a bank were paid substantially less than their male counterparts. Because there were potential confounding variables, a simple comparison of mean salaries by gender was not appropriate. We report the result of the ShrinkSet estimator in comparison to the SCAD-penalized linear and semiparametric models. We use the same metrics as in the previous example, namely, model $R^2$ as a fraction of OLS $R^2$ and number of nonzero components. By these criteria, there is no uniformly best estimator. We see that ShrinkSet does a good job at finding a parsimonious model at the expense of some explanatory power.

| | Model size | $R^2/R^2_{complete}$ |
|---|---|---|
| ShrinkSet | 7 | 0.966 |
| $\text{SCAD}_{linear}$ | 11 | 0.995 |
| $\text{SCAD}_{semiparametric}$ | 11[1] | 0.988 |

Table 5.3: Results of various variable selection procedures on salary data set

The following table (5.4) contains only those coefficients which were retained by the ShrinkSet estimator. ComputerJob is a binary variable indicating whether the employee's tasks regularly require computer skills. Not surprisingly, the associated coefficient is positive. The JobLevel variables are indicator variables indicating whether the employee's job is entry-level (1) or very advanced (6). Since 6 was absorbed in the intercept, we see that all the JobLevel coefficients are negative, and they increase as expected. Notably, gender was not included in the list of relevant predictors, and we conclude that there is little evidence of gender-based discrimination in salary at this bank. Both SCAD models resulted in the same

[1]This number reflects the number of regular nonzero coefficients plus the number of nonparmetric functions.

conclusion, because gender was retained with a high standard deviation. This analysis shows that the occasional tendency of the ShrinkSet procedure to retain some variables for which it gives large approximate $p$-values (see Table 5.2) is not pervasive. In addition, it shows that the ShrinkSet procedure, at least in some situations, outperforms SCAD in terms of automatically thresholding nonsignificant coefficients to zero.

| Variable | Point estimate | Std. error | Approx. $p$-value |
|---|---|---|---|
| Intercept | 58.9794 | 1.3222 | 0.0000 |
| ComputerJob | 4.1148 | 0.9168 | 0.0000 |
| JobLevel1 | -27.1037 | 1.4119 | 0.0000 |
| JobLevel2 | -25.1869 | 1.4507 | 0.0000 |
| JobLevel3 | -21.1132 | 1.4468 | 0.0000 |
| JobLevel4 | -15.4761 | 1.5025 | 0.0000 |
| JobLevel5 | -8.8612 | 1.5538 | 0.0000 |

Table 5.4: Factors for predicting salary selected by ShrinkSet procedure

CHAPTER 6

# CONCLUSIONS AND FURTHER RESEARCH

We have proposed a novel variable selection method based upon an obscure penalty function that mimics the $L_0$ penalty. Unlike other variable selection penalties, the ShrinkSet penalty is everywhere-differentiable and everywhere-continuous, allowing for an efficient iterative ridge algorithm capable of finding the approximate maximizer of the associated penalized likelihood in very few iterations. Because the penalty is not singular at the origin, the associated maximizer does not produce sparse solutions, despite having shrunk irrelevant coefficients very close to zero. We therefore employ a zero-thresholding step based upon a shrinkage parameter $\delta$ in order to obtain the ShrinkSet estimator. Unlike some existing methods, the ShrinkSet procedure explicitly estimates this thresholding parameter in its normal implementation, and requires no additional computation to do so.

We have given strong evidence for the asymptotic oracle properties of our estimator and have compared its finite-sample performance to that of several leading methods in various simulation and data analysis settings. In all scenarios studied, the ShrinkSet estimator either outperformed or was on par with other variable selection procedures. In addition, we maintain that the smooth nature of the penalty allows for optimization that is both computationally efficient and pedagogically appealing.

While this project has set forth some of the most important properties of the ShrinkSet estimator, there is still much to be learned. As mentioned in the text, further research could focus on establishing the oracle properties of the robust and generalized linear estimators. In addition, there may be other effective implementations of the ShrinkSet penalty in the $p \gg n$ setting. Finally, formal proofs of the conjectures set forth in this project would help to solidify the ShrinkSet procedure's position in the class of effective variable selection methods.

The fact that there are still unanswered questions should not dissuade readers from using the ShrinkSet procedure, however. It has been almost two decades since the LASSO was first introduced, but statisticians today continue to make great discoveries while studying its properties. We suspect that further investigations into the ShrinkSet procedure would likewise benefit and enrich the field.

# BIBLIOGRAPHY


Akaike, H. (1974), "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

Antoniadis, A., and Fan, J. (2001), "Regularization of Wavelet Approximations," *Journal of the American Statistical Association*, 96, 939–955.

Antoniadis, A., Gijbels, I., and Nikolova, M. (2009), "Penalized likelihood regression for generalized linear models with non-quadratic penalties," *Annals of the Institute of Statistical Mathematics*, Online First.

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384.

—— (1996), "Heuristics of instability and stabilization in model selection," *The Annals of Statistics*, 24, 2350–2383.

—— (2001), "Random Forests," *Machine Learning*, 45, 5–32.

Candès, E., and Tao, T. (2007), "The Dantzig selector: Statistical estimation when $p$ is much larger than $n$," *The Annals of Statistics*, 35, 2313–2351.

Corepetti, T., Heitz, D., Arroyo, G., Mémin, E., and Santa-Cruz, A. (2006), "Fluid experimental flow estimation based on an optical-flow scheme," *Experiments in Fluids*, 40, 80–97.

Dicker, L. (2010), "Regularized regression methods for variable selection and estimation (PhD dissertation, Harvard University)," *Dissertations and Theses: Full Text [ProQuest online database]*, Publication number: AAT 3414668.

Doksum, K., Tang, S., and Tsui, K.-W. (2008), "Nonparametric Variable Selection: The EARTH Algorithm," *Journal of the American Statistical Association*, 103, 1609–1620.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499.

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J., and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70, 859–911.

Fan, J., and Peng, H. (2004), "Nonconcave Penalized Likelihood with a Diverging Number of Parameters," *The Annals of Statistics*, 32, 928–961.

Fessler, J. A. (1996), "Mean and Variance of Implicitly Defined Biased Estimators (Such as Penalized Maximum Likelihood) : Applications to Tomography," *IEEE Transactions on Image Processing*, 5, 493–506.

Frank, I. E., and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–135.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22.

Fu, W. J. (1998), "Penalized Regressions: The Bridge versus the Lasso," *Journal of Computational and Graphical Statistics*, 7, 397–416.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning* (2nd ed.), New York: Springer.

Hoerl, A. E., and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67.

Huber, P. (1964), "Robust Estimation of a Location Parameter," *The Annals of Statistics*, 53, 73–101.

Johnson, V. A., Brun-Vézinet, F., Clotet, B., Günthard, H., Kuritzkes, D., Pillay, D., Schapiro, J., and Richman, D. (2010), "Update of the Drug Resistance Mutations in HIV-1: December 2010," *Topics in HIV Medicine*, 18, 156–163.

Lian, H. (2010), "Flexible Shrinkage Estimation in High-Dimensional Varying Coefficient Models," *arXiv.org*, arXiv:1008.2271v1.

Mallows, C. L. (1973), "Some Comments on $C_p$," *Technometrics*, 15, 661–675.

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, 83, 1023–1036.

Nikolova, M. (2005), "Analysis of the Recovery of Edges in Images and Signals by Minimizing Non-convex Regularized Least-squares," *Multiscale Modeling and Simulation*, 4, 960–991.

Rhee, S.-Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L., and Shafer, R. W. (2006), "Genotypic Predictors of Human Immunodeficiency Virus Type 1 Drug Resistance," *Proceedings of the National Academy of Sciences of the United States of America*, 46, 17355–17360.

Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Wang, H., Li, R., and Tsai, C.-L. (2007), "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviaation Method," *Biometrika*, 94, 553–568.

Wikel, J. H., and Dow, E. R. (1993), "The use of neural networks for variable selection in QSAR," *Bioorganic and Medicinal Chemistry Letters*, 3, 645–651.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H., and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B (Methodological)*, 67, 301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2007), "On the "Degrees of Freedom" of the LASSO," *The Annals of Statistics*, 35, 2173–2192.

Zou, H., and Li, R. (2008), "One-step sparse estiamtes in nonconcave penalized likelihood models," *The Annals of Statistics*, 36, 1509–1533.