VARIABLE SELECTION AND ESTIMATION WITH THE SEAMLESS-L_0 PENALTY

Lee Dicker, Baosheng Huang, and Xihong Lin

Rutgers University, Beijing Institute of Technology, and Harvard School of Public Health

Abstract: Penalized least squares procedures that directly penalize the number of variables in a regression model (L_0 penalized least squares procedures) enjoy nice theoretical properties and are intuitively appealing. On the other hand, L_0 penalized least squares methods also have significant drawbacks in that implementation is NP-hard and computationally unfeasible when the number of variables is even moderately large. One of the challenges is the discontinuity of the L_0 penalty. We propose the seamless-L_0 (SELO) penalty, a smooth function on $[0, \infty)$ that very closely resembles the L_0 penalty. The SELO penalized least squares procedure is shown to consistently select the correct model and is asymptotically normal, provided the number of variables grows more slowly than the number of observations. SELO is efficiently implemented using a coordinate descent algorithm. Since tuning parameter selection is crucial to the performance of the SELO procedure, we propose a BIC-like tuning parameter selection method for SELO, and show that it consistently identifies the correct model while allowing the number of variables to diverge. Simulation results show that the SELO procedure with BIC tuning parameter selection performs well in a variety of settings – outperforming other popular penalized least squares procedures by a substantial margin. Using SELO, we analyze a publicly available HIV drug resistance and mutation dataset and obtain interpretable results.

Key words and phrases: Penalized least squares, oracle property, coordinate descent, tuning parameter selection, BIC.

1 Introduction

Penalized least squares procedures provide an attractive approach to the variable selection and estimation problem, simultaneously identifying predictors associated with a given outcome and estimating their effects. An important class of PLS procedures rely on the L_0 penalty function (C_p (Mallows (1973)), AIC (Akaike (1974)), BIC (Schwarz (1978)), and RIC (Foster and George (1994)), for instance). Here, potential estimators are penalized according to the number of non-zero parameter estimates; thus, model complexity is penalized in a straightforward and intuitive way. In addition to their intuitive appeal, these methods arise naturally in prediction error and risk minimization (C_p and RIC, respectively), information theoretic (AIC), and Bayesian (BIC) approaches to the variable selection and estimation problem. One drawback is instability of the resulting estimators (Breiman (1996)) due to the fact that the L_0 penalty is not continuous at 0. Another perhaps more significant drawback is that implementing L_0 PLS procedures is NP-hard and can involve an exhaustive search over all possible models that is computationally infeasible when the number of potential predictors is even moderately large.

Penalized least squares methods based on continuous penalty functions have been proposed as alternatives. These methods are based on the now well-known fact that non-smooth (but continuous) penalty functions can lead to variable selection. The LASSO (Tibshirani (1996)), the L_1 PLS procedure, is perhaps the most popular. However, it may not consistently select the correct model and is not necessarily asymptotically normal (Knight (2000); Zou (2006)). Notable procedures that remedy this are the adaptive LASSO (Zou (2006)), SCAD (Fan and Li (2001)), and the recently proposed MC+ (Zhang (2010)). SCAD and MC+ replace the L_1 penalty of LASSO with a quadratic spline penalty function designed to reduce bias. In this paper, we propose the seamless L_0 (SELO) penalty that, unlike other continuous penalty functions, is explicitly designed to mimic the L_0 penalty. In addition to possessing good theoretical properties, we have found that substantial performance gains may be realized by using the SELO penalty.

The SELO penalty very closely approximates the L_0 penalty function, while addressing the stability of estimates and computational feasibility. Theoretical properties enjoyed by SELO estimators include the oracle property and our asymptotic framework allows the number of predictors, d, to tend to infinity, along with the number of observations n, provided $d/n \to 0$.

The practical performance of PLS procedures depends heavily on the choice of a tuning parameter, λ. Here, we propose a BIC tuning parameter selector that performs well when used in conjunction with SELO. This SELO/BIC procedure consistently identifies the correct model if $d/n \to 0$ and other regularity conditions are met.

We demonstrate that the SELO/BIC and SELO with data validation-based tuning parameter selection procedures perform well in a variety of simulation settings – outperforming the popular LASSO, adaptive LASSO, and SCAD procedures, especially in terms of model selection criteria. We also compare SELO/BIC

to MC+ and find that SELO/BIC performs very favorably. Finally, we show that SELO/BIC gives concise and interpretable results when applied to an HIV drug resistance and mutation dataset described by Rhee et al. (2006); these results are supported by previous work on the subject.

2 Variable selection and estimation with the seamless-L_0 penalty

2.1 Linear models and penalized least squares

Consider the linear model

$$y = X\beta^* + \epsilon, \tag{1}$$

where $y = (y_1, ..., y_n) \in \mathbb{R}^n$ is a vector of n observed outcomes, X is an $n \times d$ matrix of predictors, $\beta^* = (\beta_1^*, ..., \beta_d^*) \in \mathbb{R}^d$ is the unknown parameter of interest, and $\epsilon \in \mathbb{R}^n$ is a vector of iid additive noise with mean 0 and variance σ^2. Denote the columns of X by $X_1, ..., X_d \in \mathbb{R}^n$ and the rows of X by $x_1, ..., x_n \in \mathbb{R}^d$. Let $A = \{j;\ \beta_j^* \neq 0\}$ be the true model and suppose that d_0 is the size of the true model. That is, suppose that $|A| = d_0$, where $|A|$ denotes the cardinality of A.

When discussing variable selection, it is convenient to have concise notation for referring to sub-vectors and sub-matrices. For $S \subseteq \{1, ..., d\}$, let $\beta_S = (\beta_j)_{j \in S} \in \mathbb{R}^{|S|}$ be the $|S|$-dimensional sub-vector of β containing entries indexed by S and let X_S be the $n \times |S|$ matrix obtained from X by extracting columns corresponding to S. Given a $d \times d$ matrix C and subsets $S_1, S_2 \subseteq \{1, ..., d\}$, let C_{S_1,S_2} be the $|S_1| \times |S_2|$ sub-matrix of C with rows determined by S_1 and columns determined by S_2.

All of the variable selection and estimation procedures that we consider in this paper are PLS procedures. A generic PLS estimator minimizes the objective

$$\frac{1}{2n}||y - X\beta||^2 + \sum_{j=1}^{d} p(\beta_j), \tag{PL}$$

where $||\cdot||$ denotes the L_2 norm and $p(\cdot)$ is a penalty function that generally depends on some tuning parameter, often denoted by λ. Note that we may penalize different coefficients in different ways by allowing $p(\cdot)$ to depend on j in (PL).

LASSO is the PLS procedure with the L_1 penalty $p(\beta_j) = p_\lambda(\beta_j) = \lambda|\beta_j|$ and is perhaps the most popular and widely studied non-L_0 penalized procedure. However, LASSO estimates may be biased (Fan and Li (2001)) and inconsistent for model selection (Zou (2006)). This implies that the LASSO does not have the oracle property of Fan and Li (2001) (a variable selection and estimation procedure is said to have the oracle property if it selects the true model, A, with probability tending to one, and if the estimated coefficients are asymptotically normal, with the same asymptotic variance as the least squares estimator based on the true

model).

The adaptive LASSO is a weighted version of LASSO which has the oracle property (Zou (2006)). Slightly abusing notation, the adaptive LASSO penalty is defined by $p(|\beta_j|) = \lambda w_j |\beta_j|$, where w_j is a data-dependent weight. A typical weight, if $d < n$, is $w_j = |\hat{\beta}_j^{(0)}|^{-1}$, where $\hat{\beta}^{(0)}$ is the ordinary least squares (OLS) estimator of β^*.

SCAD is another popular PLS procedure. The SCAD penalty is the continuous function defined by

$$p'_{\mathrm{SCAD}}(\beta_j) = \lambda \mathrm{sgn}(\beta_j) \left[I\{|\beta_j| \leq \lambda\} + \frac{\max\{a\lambda - \beta_j, 0\}}{(a-1)\lambda} I\{|\beta_j| > \lambda\} \right],$$

for $\beta_j \neq 0$, and $p_{\mathrm{SCAD}}(0) = 0$, where $a > 2$ is another tuning parameter; Fan and Li (2001) recommend taking $a = 3.7$ and we follow this recommendation throughout. Notice that the SCAD penalty is a quadratic spline with knots at $\pm\lambda$ and $\pm a\lambda$. Fan and Li (2001) showed that the SCAD procedure has the oracle property when $\lambda = \lambda_n$ satisfies certain conditions as $n \to \infty$.

The last penalty we introduce is the minimax concave penalty (Zhang (2010)). We refer to this recently proposed penalty and the associated PLS procedure as MC+ ("+" refers to the algorithm used for implementing MC+). The MC+ penalty is

$$p_{\mathrm{MC+}}(\beta_j) = \lambda \left[|\beta_j| - \frac{|\beta_j|^2}{2\gamma\lambda} \right] I\{0 \leq |\beta_j| < \gamma\lambda\} + \frac{\lambda^2\gamma}{2} I\{|\beta_j| \geq \gamma\lambda\}.$$

Like the SCAD penalty, the MC+ penalty is a quadratic spline. The parameter $\gamma > 0$ determines the concavity of $p_{\mathrm{MC+}}$. Zhang (2010) proved that the MC+ procedure can select the correct model with probability tending to 1 and that MC+ estimators have good properties in terms of L^p-loss, provided λ and γ satisfy certain conditions. Zhang's results in fact allow for $d \gg n$.

2.2 The seamless-L_0 penalty

L_0 penalties have the form

$$p_\lambda(\beta_j) = \lambda I\{\beta_j \neq 0\} = \begin{cases} \lambda & \text{if } \beta_j \neq 0, \\ 0 & \text{if } \beta_j = 0 \end{cases}$$

and directly penalize non-zero parameter estimates. These penalties determine an important class of PLS procedures that includes AIC, BIC, and RIC. L_0 penalties have a strong intuitive appeal and desirable theoretical properties. However, the associated variable selection and estimation procedures tend to be unstable (Breiman (1996), especially when the data contains only a weak signal, and are computationally infeasible for even moderately large d, as implementations generally require a combinatorial search. This is largely due to the fact that the L_0 penalty is discontinuous. We introduce a continuous approximation to

the L_0 penalty,

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

There is a tuning parameter $\tau > 0$, in addition to λ, and when τ is small, $p_{\text{SELO}}(\beta_j) \approx \lambda I\{\beta_j \neq 0\}$. In practice, we have found that when the data are standardized so that $X_j^T X_j = n$, taking $\tau = 0.01$ gives good results. Since the SELO penalty is continuous, the associated PLS procedure is more stable than L_0 procedures. Furthermore, we will see that approximate minima of the SELO penalized objective,

$$\frac{1}{2n}||y - X\beta||^2 + \sum_{j=1}^{d} p_{\text{SELO}}(\beta_j), \tag{SELO}$$

can be rapidly computed using a coordinate descent algorithm.

The SELO penalty function is plotted in Figure 1, along with the SCAD, L_1 (LASSO), and L_0 penalties (left panel), and the MC+ penalty for various values of γ (right panel). Notice that the seamless-L_0 penalty mimics the L_0 penalty much more closely than the L_1 or SCAD penalties. We point out that the L_1 penalty is unbounded and this may lead to estimation bias (Fan and Li (2001)). The SCAD penalty is bounded and, like the SELO method, enjoys the oracle property; however, in Section 5, we will see that the SELO method offers substantial performance gains over SCAD in a variety of simulated settings. The MC+ penalty is plotted for values of γ ranging from $\gamma = 1.01$ (which has the greatest concavity and most closely resembles the SELO penalty) to $\gamma = 5$. The adaptive LASSO penalty is not plotted in Figure 1; the penalty varies with the weights w_j and, for each w_j, the adaptive LASSO penalty resembles the LASSO penalty with differing slope. As shown in Section 5, the SELO estimator outperforms the adaptive LASSO estimator in various settings.

2.3 Theoretical properties of the SELO estimator

The main results in this section show that the SELO estimator has the oracle property when λ and τ are chosen appropriately. That is, SELO consistently selects the correct model A and the SELO estimator is asymptotically normal with the same asymptotic variance as the OLS estimator based on A. We consider an asymptotic regime in which d may tend to infinity with n. The following conditions play a role in our analysis.

(A) $n \to \infty$ and $d\sigma^2/n \to 0$.

(B) $\rho\sqrt{n/(d\sigma^2)} \to \infty$, where $\rho = \min_{j\in A} |\beta_j^*|$.

(C) There exist positive constants $r_0, R_0 > 0$ such that $r_0 \leq \lambda_{min}(n^{-1}X^TX) < \lambda_{\max}(n^{-1}X^TX) \leq R_0$, where $\lambda_{min}(n^{-1}X^TX)$ and $\lambda_{max}(n^{-1}X^TX)$ are the smallest and largest eigenvalues of $n^{-1}X^TX$, respectively.

(D) $\lambda = O(1)$, $\lambda\sqrt{n/(d\sigma^2)} \to \infty$, and $\tau = O[d^{-1}(d\sigma^2/n)^{3/2}]$.

(E) $\lim_{n\to\infty} n^{-1} \max_{1\le i\le n} \sum_{j=1}^{d} x_{ij}^2 = 0$.

(F) $E\left(|\epsilon_i/\sigma|^{2+\delta}\right) < M$ for some $\delta > 0$ and $M < \infty$.

Since d may vary with n, it is implicit that β^* may vary with n. Additionally, we allow the model A and the distribution of ϵ (in particular, σ^2) to change with n. Condition (A) limits how d and σ^2 may grow with n; it is substantially weaker than required by Fan and Peng (2004) who require $d^5/n \to 0$, and slightly weaker than required by Zou and Zhang (2009) who require $\log(d)/\log(n) \to \nu \in [0, 1)$. Other authors have studied PLS methods in settings where $d > n$ with growth conditions on d weaker than Condition (A), but when it is relaxed, additional stronger conditions are needed to obtain desirable theoretical properties. Condition (B) gives a lower bound on the size of the smallest nonzero entry of β^*; it is allowed to vanish asymptotically, provided it does not do so faster than $\sqrt{d\sigma^2/n}$. Similar conditions are found in (Fan and Peng (2004)) and (Zou and Zhang (2009)). Condition (C) is an identifiability condition. Condition (D) restricts the rates of the tuning parameters λ and τ, but does not constrain the minimum size of τ. In practice, we have found that one should not take τ too small in order to preserve stability of the SELO estimator. Conditions (E) and (F) are used to prove asymptotic normality of SELO estimators and are related to the Lindeberg condition of the Lindeberg-Feller CLT (Durrett (2005)). A proof of Theorem 1 is in the Appendix.

Theorem 1. *Suppose that (A)-(F) hold. There exists a sequence of $\sqrt{n/(d\sigma^2)}$-consistent local minima of (SELO), $\hat{\beta}$, such that:*

(i) [Model selection consistency] $\lim_{n\to\infty} P(\{j;\ \hat{\beta}_j \neq 0\} = A) = 1$.

(ii) [Asymptotic normality and efficiency]

$$\sqrt{n}B_n(n^{-1}X_A^T X_A/\sigma^2)^{1/2}(\hat{\beta}_A - \beta_A^*) \to N(0, G)$$

in distribution, where B_n is an arbitrary $q \times |A|$ matrix such that $B_n B_n^T \to G$.

Remark 1. If d, β^*, and σ^2 are fixed, then Theorem 1 (ii) implies that $\hat{\beta}_A$ has the same asymptotic distribution as the ordinary least squares estimator of β_A^*, given knowledge of the model A in advance.

Remark 2. Though SELO and other PLS methods for variable selection and estimation are primarily useful when β^* is sparse, we make no assumptions about the sparsity level of β^* in Theorem 1.

Remark 3. In any implementation of SELO, concrete values of the tuning parameters λ and τ must be selected. In Section 3 we propose a BIC tuning parameter selection procedure and prove that when SELO is implemented with BIC tuning parameter selection, the resulting estimator consistently selects the correct model.

2.4 A standard error formula

Let $\hat{\beta} = \hat{\beta}(\tau, \lambda)$ be a local minimizer of (SELO). Following Fan and Li (2001) and Fan and Peng (2004), standard errors of $\hat{\beta}$ may be estimated by using quadratic approximations to (SELO). Indeed, the approximation

$$p_{\mathrm{SELO}}(\beta_j) \approx p_{\mathrm{SELO}}(\beta_{j0}) + \frac{1}{2|\beta_{j0}|} p'_{\mathrm{SELO}}(\beta_{j0})(\beta_j^2 - \beta_{j0}^2), \text{ for } \beta_j \approx \beta_{j0},$$

suggests that (SELO) may be replaced by

$$\text{minimize } \frac{1}{n}||y - X\beta||^2 + \sum_{j=1}^{d} \frac{p'_{\mathrm{SELO}}(\beta_{j0})}{|\beta_{j0}|}\beta_j^2,$$

at least for the purposes of obtaining standard errors. Using this expression, we obtain a sandwich formula for the estimated standard error of $\hat{\beta}_{\hat{A}}$, where $\hat{A} = \{j; \hat{\beta}_j \neq 0\}$:

$$\widehat{\mathrm{cov}}(\hat{\beta}_{\hat{A}}) = \hat{\sigma}^2 \left\{ X_{\hat{A}}^T X_{\hat{A}} + n\Delta_{\hat{A},\hat{A}}(\hat{\beta}) \right\}^{-1} X_{\hat{A}}^T X_{\hat{A}} \left\{ X_{\hat{A}}^T X_{\hat{A}} + n\Delta_{\hat{A},\hat{A}}(\hat{\beta}) \right\}^{-1} \tag{2}$$

where $\Delta(\beta) = \mathrm{diag}\{p'_{\mathrm{SELO}}(|\beta_1|)/|\beta_1|, ..., p'_{\mathrm{SELO}}(|\beta_d|)/|\beta_d|\}$, $\hat{\sigma}^2 = (n - \hat{d}_0)^{-1}||y - X\hat{\beta}||^2$, and $\hat{d}_0 = |\hat{A}|$ is the number of elements in $\hat{A}$. Under the conditions of Theorem 1,

$$B_n X_A^T X_A \widehat{\mathrm{cov}} \left(\hat{\beta}_{\hat{A}} \right) B_n^T / \sigma^2 \to G.$$

3 Tuning parameter selection

Tuning parameter selection is an important issue in most PLS procedures. One often proceeds by finding estimators that correspond to a range of tuning parameter values (referred to as a solution path). The preferred estimator is then identified along the solution path as one corresponding to a tuning parameter value that optimizes some criteria, such as GCV (Breiman (1995); Tibshirani (1996); Fan and Li (2001)), AIC (Zou, Hastie, and Tibshirani (2007)), or BIC (Wang, Li, and Tsai (2007); Wang and Leng (2007); Zou, Hastie, and Tibshirani (2007); Wang, Li, and Leng (2009)). It is well known that GCV and AIC-based methods are not consistent for model selection in the sense that, as $n \to \infty$, they may select irrelevant predictors with non-vanishing probability (Shao (1993); Wang, Li, and Tsai (2007)). BIC-based tuning parameter selection (Schwarz, 1978) has been shown to be consistent for model selection in several settings (Wang, Li, and Tsai (2007); Wang and Leng (2007); Zou, Hastie, and Tibshirani (2007)). Thus, if variable selection and identification of the true model, A, is the primary goal, then BIC tuning parameter selection may be preferred over GCV and AIC. We propose a BIC tuning parameter selector for the SELO procedure and show that the SELO/BIC procedure is consistent for model selection, even if the number of predictors diverges (i.e.

$d \to \infty$).

BIC procedures are often implemented by minimizing

$$\mathrm{BIC}_0 = \mathrm{BIC}_0(\hat{\beta}) = \log(\hat{\sigma}^2) + \frac{\log(n)}{n}\widehat{\mathrm{DF}} \tag{3}$$

over a collection of estimators, $\hat{\beta}$, where $\hat{\sigma}^2$ is an estimator of the residual variance and $\widehat{\mathrm{DF}}$ is an estimator of the degrees of freedom corresponding to $\hat{\beta}$. We propose estimating the degrees of freedom for the SELO estimator by the number of selected coefficients: $\widehat{\mathrm{DF}} = \hat{d}_0$, where $\hat{d}_0 = |\{j; \hat{\beta}_j \neq 0\}|$. This estimator of DF is partially motivated by (Zou, Hastie, and Tibshirani (2007)), where connections between the degree of freedom for LASSO and Stein's unbiased risk estimation theory are discussed. To estimate the residual variance, we use $\hat{\sigma}^2 = (n - \hat{d}_0)^{-1}||y - X\hat{\beta}||^2$; here, $n - \hat{d}_0$ is used to account for degrees of freedom lost to estimation. Additionally, in order to account for a diverging number of parameters, we allow for additional flexibility in our BIC criterion by replacing $\log(n)$ in BIC_0 with a positive number k_n that depends on the sample size n. This follows Wang, Li, and Leng (2009), who showed that if $d \to \infty$, then it may be necessary to choose $k_n > \log(n)$ to obtain model selection consistency with BIC. Thus, our BIC criterion is

$$\mathrm{BIC}_{k_n} = \mathrm{BIC}_{k_n}(\hat{\beta}) = \log\left(\frac{||y - X\hat{\beta}||^2}{n - \hat{d}_0}\right) + \frac{k_n}{n}\hat{d}_0.$$

With $d \to \infty$, Wang, Li and Leng (2009) proved that if $k_n/\log(n) \to \infty$ and ϵ_i is Gaussian, then, under certain additional conditions, BIC is consistent for model selection. Our Theorem 2 below is a model selection consistency result for BIC (when used in conjunction with SELO) and it extends Wang, Li, and Leng's (2009) results in two directions:

(i) It allows for a broader class of errors distributions. Theorem 2 (a) applies to ϵ_i with a $(2+\delta)$-th moments; Theorem 2 (b)-(c) applies to sub-Gaussian ϵ_i.

(ii) Theorem 2 (c) delineates circumstances under which the classical BIC (with $k_n = \log(n)$) is consistent for model selection, even when $d \to \infty$.

Modified version of conditions (A), (B), and (F) are required for the different parts of Theorem 2.

(A2) $n \to \infty$, $dk_n/n \to 0$ and $\sigma^2 = O(1)$.

(B2) $\rho\sqrt{n/(dk_n)} \to \infty$, where $\rho = \min_{j \in A}|\beta_j^*|$.

(F2) The errors $\epsilon_1, ..., \epsilon_n$ are subgaussian in the sense that there exists $\sigma_0 > 0$ such that $E(e^{t\epsilon_i}) \leq e^{\sigma_0^2 t^2/2}$ for all $t \in \mathbb{R}$.

Note that conditions (A2), (B2), and (F2) are stronger than conditions (A), (B), and (F), respectively. However, (A2)-(B2) are required for Wang, Li, and Leng's (2009) results on model selection consistency and BIC

when the number of predictors diverges (see condition 4 in their paper). Furthermore, condition (F2) is only required in parts (b)-(c) of Theorem 2. A proof of Theorem 2 is found in the Appendix.

Theorem 2. *Suppose that conditions (A2)-(B2), (C), and (E) hold. Suppose further that $\Omega \subseteq \mathbb{R}^2$ is a subset which contains a sequence $(\lambda, \tau) = (\lambda_n^*, \tau_n^*)$ such that condition (D) holds. Let $\hat{\beta}^* = \hat{\beta}(\lambda_n^*, \tau_n^*)$ be the local minima of (SELO) described in Theorem 1 and let $\mathrm{BIC}_{k_n}^- = \inf\{\mathrm{BIC}_{k_n}\{\hat{\beta}(\lambda, \tau)\};\ (\lambda, \tau) \in \Omega,\ \hat{A} \neq A\}$.*

(a) [$(2+\delta)$-th moments] *If condition (F) holds and $\liminf_{n\to\infty} k_n/n^{2/(2+\delta)} > 0$, then*

$$\lim_{n\to\infty} P\left[\mathrm{BIC}_{k_n}^- > \mathrm{BIC}_{k_n}\left\{\hat{\beta}(\lambda_n^*, \tau_n^*)\right\}\right] = 1. \tag{4}$$

(b) [Subgaussian] *If condition (F2) holds and $k_n/\log(n) \to \infty$, then (4) holds.*

(c) [Subgaussian; $k_n = \log(n)$] *Suppose that condition (F2) holds. Let $R_1 \in \mathbb{R}$ be a constant such that $\max_{1\leq j\leq d} n^{-1}||X_j||^2 \leq R_1$ and recall from condition (C) that r_0 is a constant satisfying $0 < r_0 \leq \lambda_{min}(n^{-1}X^TX)$. If there is a constant $\zeta > 0$ such that*

$$d = o\left\{n^{\sigma^2 r_0/(2\sigma_0^2 R_1)-\zeta}\right\} \tag{5}$$

then (4) holds with $k_n = \log(n)$.

Remark 1. Theorem 2 implies that if $\hat{\beta}(\hat{\lambda}, \hat{\tau})$ is chosen to minimize BIC with an appropriately chosen k_n, then $\hat{\beta}(\hat{\lambda}, \hat{\tau})$ is consistent for model selection. In other words, if $\mathrm{BIC}_{k_n}\{\hat{\beta}(\hat{\lambda}, \hat{\tau})\} = \mathrm{BIC}_{k_n}^-$, then $\lim_{n\to\infty} P[\{j; \hat{\beta}_j(\hat{\lambda}, \hat{\tau}) \neq 0\} = A] = 1$.

Remark 2. Part (a) of Theorem 2 describe how k_n should be chosen if ϵ_i has a $(2+\delta)$-th moment. In particular, if $k_n = n^{2/(2+\delta)}$, then the SELO/BIC procedure is consistent for model selection under the given conditions. Note that if ϵ_i has higher order moments, then k_n may be chosen to be smaller. Under (A2)-(B2), smaller k_n is generally desirable because this entails weaker conditions on d and ρ, i.e. smaller k_n allows for larger d and smaller ρ.

Remark 3. Parts (b)-(c) of Theorem 2 describe how k_n may be chosen effectively if ϵ_i is subgaussian. Part (b) implies that consistent model selection is ensured if k_n satisfies $k_n/\log(n) \to \infty$. Part (c) implies that if (5) holds, then one can take $k_n = \log(n)$ to achieve consistent model selection with SELO/BIC; in this case, the BIC penalty is the same as in the classical (fixed d) BIC criterion (3). It is important to note that the additional condition (5) for Theorem 2 (c) further limits the size of d (though $d \to \infty$ is still allowed). However, if d, n are given and they satisfy (5), then choosing a smaller k_n for the BIC criterion is still desirable because condition (B2) is less restrictive for smaller k_n.

Remark 4. The constant R_1 in Theorem 2 (c) exists by condition (C).

Remark 5. In Theorem 2 (c), notice that if $k_n = \log(n)$, then (5) implies (A2). Hence, requiring condition (A2) is redundant in Theorem 2 (c). Roughly speaking, the condition (5) allows for larger d if the predictor matrix X is "well-behaved" and if σ_0^2 is close to σ^2 (recall that σ_0^2 is defined in (F2); it follows that $\sigma_0^2 \geq \sigma^2$ and if ϵ_i is gaussian, then $\sigma_0^2 = \sigma^2$). In any event, condition (5) implies $d/n^{1/2} \to 0$.

In Section 5, we describe the results of several simulation studies and a data analysis where SELO with BIC tuning parameter selection was implemented. In all of our implementations we took $k_n = \log(n)$; that is, we used the BIC criterion

$$\mathrm{BIC}(\hat{\beta}) = \mathrm{BIC}_{\log(n)}(\hat{\beta}) = \log\left(\frac{||y - X\hat{\beta}||^2}{n - \hat{d}_0}\right) + \frac{\log(n)}{n}\hat{d}_0. \tag{6}$$

Broadly speaking, Theorem 2 (c) provides justification for using this criterion in settings where X is well-behaved, the errors ϵ_i are not too extreme, and d is large, but significantly smaller than n. In addition to performing effectively in all of the settings considered in this paper, the BIC criterion with $k_n = \log(n)$ is appealing because of its similarity to the classical BIC criterion (3). Furthermore, we have found that using a larger k_n in BIC_{k_n}, which Theorem 2 suggests may be necessary to ensure consistent model selection in certain situations, can lead to underfitting (i.e. omitting important predictors from the selected model); this, in turn, may yield estimators with relatively poor prediction and estimation error.

4 Implementation: coordinate descent

In this section, we describe a simple and efficient algorithm for obtaining SELO estimators for finding a SELO solution path. This is necessary for effectively implementing any of the SELO tuning parameter selection procedures described above. An associate editor pointed out that this algorithm was also recently proposed for a very general class of penalized likelihood methods by Fan and Lv (2011), who refer to it as "iterative coordinate ascent" (ICA). Coordinate descent algorithms for L_1 and L_2 penalized likelihood methods have also been described by Friedman, Hastie, and Tibshirani (2010).

The idea of the coordinate descent algorithm is to find local optima of a multivariate optimization problem by solving a sequence of univariate optimization problems. This can be very effective if the univariate optimization problems are simple. Let

$$Q(\beta) = \frac{1}{2n}||y - X\beta||^2 + \sum_{j=1}^{d} p_{\mathrm{SELO}}(\beta_j)$$

and, for fixed $\beta_{-k} = (\beta_1, ..., \beta_{k-1}, \beta_{k+1}, ..., \beta_d)$, take $Q_k(\beta_k; \beta_{-k}) = Q(\beta)$. Given β_{-k}, it is straightforward

to minimize $Q_k(\,\cdot\,;\beta_{-k})$. Indeed, differentiating, one observes that finding critical points of $Q_k(\,\cdot\,;\beta_{-k})$ amounts to finding the roots of cubic equations, which may be done very rapidly using Cardano's formula or some other procedure. The coordinate descent algorithm is implemented by minimizing $Q_k(\,\cdot\,;\beta_{-k})$ and using the solution to update β; at the next step, $Q_{k+1}(\,\cdot\,;\beta_{-(k+1)})$ is minimized and the minimizer is again used to update β. In this way, we cycle through the indices $k = 1, ..., d$; this can be performed multiple times until some convergence criterion is met. More precisely, for an integer k, define $\bar{k} \in \{1, ..., d\}$ so that $k - \bar{k}$ is divisible by d; the coordinate descent algorithm is described below.

ALGORITHM 1 (SELO-CD).

(1) *Let $\beta^{(1)} \in \mathbb{R}^d$ be some initial value and set $k = 1$.*

(2) *Let $\tilde{\beta} = \operatorname{argmin}_\beta Q_{\bar{k}}(\beta; \beta^{(k)}_{-\bar{k}})$.*

(3) *Define $\beta^{(k+1)}$ by $\beta^{(k+1)}_{-\bar{k}} = \beta^{(k)}_{-\bar{k}}$ and $\beta^{(k+1)}_{\bar{k}} = \tilde{\beta}$.*

(4) *If $|\beta^{(k+1)} - \beta^{(k)}|$ is small or k is very large, exit and return $\beta^{(k+1)}$; otherwise, increment k by 1 and go to step (2).*

In general, determining theoretical convergence properties of algorithms for non-convex minimization problems is difficult. However, it is clear that $Q(\beta^{(k+1)}) \leq Q(\beta^{(k)})$. Furthermore, if $\hat{\beta}$ minimizes $Q(\cdot)$ and $\beta^{(1)}$ lies in some ball centered about $\hat{\beta}$ on which $Q(\cdot)$ is convex, then it is easy to see that $\hat{\beta}^{(k)} \to \hat{\beta}$. In practice, we have found that numerical convergence of $\beta^{(k)}$ generally occurs rapidly, when a reasonable initial value is chosen.

Assuming convergence, SELO-CD returns the minimum of (SELO), for a fixed pair of tuning parameters, (λ, τ). To obtain a SELO solution path, we repeatedly implement SELO-CD for a range of values of (λ, τ). To speed up the implementation, we utilize warm starts. This means that given the output of SELO-CD, $\hat{\beta} = \hat{\beta}(\lambda, \tau)$, for some fixed pair of tuning parameters, (λ, τ), we use $\hat{\beta}$ as the initial estimate for implementing SELO-CD at nearby tuning parameter values. We have found that using warm starts dramatically decreases computing time.

The first step in finding a SELO solution path involves determining values of (λ, τ) for which SELO estimators will be obtained. For each τ, we consider a wide range of values for λ: $\lambda_{\max} = \lambda_0 > \cdots > \lambda_M = 0$, for some large number M ($M = 100$ in our implementations), where

$$\lambda_{\max} := \frac{||y||^2}{2n} \log(2) \left\{ \log\left(\frac{||y||^2}{||y||^2 + 2n\tau ||X^T y||_\infty} + 1 \right) \right\}^{-1}.$$

Our choice of $\lambda_{\max}$ is motivated by the fact that

$$\underset{\beta}{\operatorname{argmin}} \frac{1}{2n}||y - X\beta||^2 + \sum_{j=1}^{d} p_{\mathrm{SELO}}(\beta_j; \lambda, \tau) = 0$$

whenever $\lambda \geq \lambda_{\max}$. Thus, $\hat{\beta}(\lambda_0, \tau) = 0$ and $\hat{\beta}(\lambda_k, \tau)$ grows with k; in other words, given τ, the value $\lambda_0 = \lambda_{\max}$ tells us where the solution path begins. For each τ and $\lambda_{\max} = \lambda_0 > \cdots > \lambda_M = 0$, SELO estimators may be rapidly obtained with the following algorithm.

ALGORITHM 2 (`SELO-CD-PATH`).

(1) *Let $\hat{\beta}^{(m)} = 0 \in \mathbb{R}^d$ and set $m = 0$.*

(2) *Run Algorithm 1, with initial value $\beta^{(1)} = \hat{\beta}^{(m)}$ and $\lambda = \lambda_{m+1}$; let $\hat{\beta}^{(m+1)}$ equal the output of Algorithm 1.*

(3) *If $m \geq M - 1$, exit and return $\{\hat{\beta}^{(m)}\}_{m=0}^{M}$; otherwise, increment m by 1 and go to step (2).*

Remark 1. Implementing `SELO-CD-PATH` for each value of τ and each sequence $\lambda_{\max} = \lambda_0 > \cdots > \lambda_M = 0$ to be considered gives the entire SELO solution path.

Remark 2. In practice, we have found that if the columns of X are standardized so that that $||X_j||^2 = n$, for $j = 1, ..., p$, then taking $\tau = 0.01$ or selecting τ from a relatively small range of possible values works well.

5 Simulation studies and a data example

5.1 Simulation methodology

All of our simulations were based on datasets of independent observations, (y_i, x_i^T), $i = 1, ..., n$, where $y_i = x_i^T\beta^* + \epsilon_i$. Throughout, $x_i \sim N(0, \Sigma)$ and $\epsilon_i \sim N(0, \sigma^2)$ were drawn independently. We took $\Sigma = (\sigma_{ij})$, where $\sigma_{ij} = 0.5^{|i-j|}$. In addition to SELO, we considered the LASSO, adaptive LASSO (with weights $w_j = |\hat{\beta}_j^{(0)}|^{-1}$, where $\hat{\beta}^{(0)}$ is the OLS estimator), SCAD, and MC+ PLS procedures. Covariates were standardized to have $||X_j|| = n$, $j = 1, ..., n$, prior to obtaining estimates; however, all summary statistics discussed below pertain to estimators transformed to the original scale. In our simulations, SELO solution paths were found using `SELO-CD-PATH`. LASSO, adaptive LASSO, and SCAD solution paths were also computed using coordinate descent algorithms. We used the PLUS algorithm (implemented in the `plus` R library) to find MC+ estimators. For SELO, LASSO, adaptive LASSO, and SCAD, tuning parameter selection was performed with BIC and, alternatively, a prediction error minimization-based procedure referred to as data-validation (DV) tuning parameter selection. For MC+, we used data-validation based tuning parameter selection and a method proposed by Zhang (2010) that involves estimating a specific value of λ with good

theoretical properties, in addition to BIC. For SELO tuning parameter selection, we considered two types of solution paths: one where $\tau = 0.01$ was fixed and we tuned over $\lambda \in \{\lambda_0, \lambda_1, ..., \lambda_M\}$, where $\lambda_0 = \lambda_{\max}$, and a second where we tuned over $\tau \in \{0.001, 0.01, 0.1, 0.5\}$ and $\lambda \in \{\lambda_0, \lambda_1, ..., \lambda_M\}$. For selecting the parameter γ when implementing MC+, we followed Zhang (2010) and, for each simulated dataset, took $\gamma = 2/(1 - \max_{i \neq j} |X_i^T X_j|/n)$.

BIC tuning parameter selection for SELO was discussed in Section 4. In all of our SELO/BIC simulations, we used the BIC criterion (6). In addition to using BIC tuning parameter selection with SELO, we used BIC tuning parameter selection with LASSO, adaptive LASSO, SCAD, and MC+. For these PLS methods, we used previously proposed versions of BIC that are special cases of (3) with different values of $\hat{\sigma}^2$ and $\widehat{\mathrm{DF}}$ (where available). In particular, for SCAD, we followed Wang, Li, and Tsai (2007) in using a BIC tuning parameter selector with $\hat{\sigma}^2 = n^{-1}||y - X\hat{\beta}||^2$ and

$$\widehat{\mathrm{DF}} = \mathrm{tr}\left\{X_{\hat{A}}\left(X_{\hat{A}}^T X_{\hat{A}} + n\Delta^{\mathrm{SCAD}}_{\hat{A},\hat{A}}\right)^{-1} X_{\hat{A}}^T\right\}, \tag{7}$$

where $\Delta^{\mathrm{SCAD}} = \mathrm{diag}\left\{p'_{\mathrm{SCAD}}(\hat{\beta}_1)/|\hat{\beta}_1|, ..., p'_{\mathrm{SCAD}}(\hat{\beta}_d)/|\hat{\beta}_d|\right\}$. That is, for the SCAD/BIC procedure, the tuning parameter λ was selected to minimize BIC_0, using the above values of $\hat{\sigma}^2$ and $\widehat{\mathrm{DF}}$. For LASSO, there is less consensus in the literature on how to implement BIC tuning parameter selection. We took $\hat{\sigma}^2 = n^{-1}||y - X\hat{\beta}||^2$ and, following Zou, Hastie, and Tibshirani (2007), we took $\widehat{\mathrm{DF}} = \hat{d}_0$ (the BIC objective in (Zou, Hastie, and Tibshirani (2007) is derived under the assumption that σ^2 is known and takes a slightly different form than BIC_0); we used the same $\hat{\sigma}^2$ and $\widehat{\mathrm{DF}}$ for the adaptive LASSO/BIC procedure. To our knowledge, BIC tuning parameter selection for MC+ has not been discussed in the literature. For MC+/BIC we mimicked the BIC procedure for SELO and took $\hat{\sigma}^2 = (n-d)^{-1}||y - X\hat{\beta}||^2$ and $\widehat{\mathrm{DF}} = \hat{d}_0$. In Sections 5.4-5.5, we discuss the results of a simulation study where we implement BIC selection procedures for LASSO, adaptive LASSO, and SCAD that mimic the BIC procedure proposed for SELO more closely.

Data validation procedures may provide the simplest example of prediction error minimization-based tuning parameter selection. We implemented DV tuning parameter selection as follows. After obtaining a solution path of estimators, we generated an independent validation dataset, $(\tilde{y}_i, \tilde{x}_i^T)$, $i = 1, ..., n$, under the same conditions as the original data, (y_i, x_i), $i = 1, ..., n$. The estimator in the solution path that minimized

$$\sum_{i=1}^{n} (\tilde{y}_i - \tilde{x}_i^T \hat{\beta})^2 \tag{8}$$

was selected. The quantity (8) is a surrogate for the prediction error associated with $\hat{\beta}$; thus, we expect the prediction error of the selected estimator to be relatively low. Notice that DV tuning parameter selection requires a validation dataset and utilizes more data than BIC tuning parameter selection. Thus, we should

not be surprised if the DV selector outperforms BIC by some metrics, especially metrics related to prediction error. In fact, the simulation results below indicate that BIC generally outperforms DV in terms of model selection metrics and that, in a wide range of setting, the tuning parameter selection procedures perform very similarly in terms of prediction error metrics.

In addition to DV and BIC, we implemented a third method for MC+ tuning parameter selection that was proposed by Zhang (2010). This method, referred to as universal tuning, or just λ_{univ}, consists of taking $\lambda = \hat{\lambda}_{univ} = \hat{\sigma}\sqrt{2\log(d)/n}$, where $\hat{\sigma}^2 = (n-d)^{-1}||y - X\hat{\beta}||^2$. Zhang (2010) showed that λ_{univ} possesses good theoretical properties – at least asymptotically - for both variable selection and estimation.

5.2 Simulation study I: $d = 8$

Our first simulation study involved a modest number of predictors. We set $\beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0)^T \in \mathbb{R}^8$. Thus, $d = 8$ and $d_0 = 3$. We fixed $\sigma^2 = 9$ and simulated datasets with $n = 50$ and $n = 100$ observations. For each value of n, we simulated 1000 independent datasets $\{(y_1, x_1^T), ..., (y_n, x_n^T)\}$ and, for each dataset, we computed estimates of β^*. For each estimate, $\hat{\beta}$, we recorded:

- the model size, $\hat{A} = \{j;\ \hat{\beta}_j \neq 0\}$;
- an indicator of whether or not the true model was selected, $I\{\hat{A} = A\}$;
- the false positive rate, $|\hat{A} \setminus A|/|\hat{A}|$, where $A \setminus \hat{A}$ denotes the set difference of A and $\hat{A}$;
- the false negative rate, $|A \setminus \hat{A}|/(d - |\hat{A}|)$;
- and the model error, $(\hat{\beta} - \beta^*)^T\Sigma(\hat{\beta} - \beta^*)$.

Results for SELO, LASSO, adaptive LASSO, SCAD, and MC+ are summarized in Table 1.

Table 1 indicates that SELO consistently selects a smaller model and the correct model more frequently than any of the other procedures, regardless of tuning method. These competing methods overfit the model considerably, when compared with SELO. For instance, when $n = 100$ and DV tuning was used (results for DV tuning are contained in the top half of Table 1), SELO with fixed $\tau = 0.01$ selected models with average size 3.669 and the correct model was selected in 70.9% of the simulated datasets; on the other hand, LASSO, adaptive LASSO, SCAD, and MC+ selected models with average size 5.791, 4.472, 4.324, and 4.530 and selected the correct model in 5.8%, 30.7%, 35.3%, and 30.4% of simulated datasets, respectively. In terms of model error, it appears that all the methods have similar performance when $n = 50$; when $n = 100$, SELO performs comparably with adaptive LASSO, SCAD, and MC+ and all of these methods outperform LASSO. For example, for $n = 100$ and DV tuning, the mean model error for SELO with $\tau \in \{0.001, 0.01, 0.1, 0.5\}$ was 0.352; the mean model error for LASSO, adaptive LASSO, SCAD, and MC+ was 0.521, 0.430, 0.384, and 0.401, respectively. Notice that when data validation tuning parameter selection is used, SELO with tuning

over $\tau \in \{0.001, 0.01, 0.1, 0.5\}$ seems to give improved model error when compared to SELO with $\tau = 0.01$ fixed. This is not surprising; however, the differences between results for fixed τ and variable τ appear to be fairly small.

The bottom half of Table 1 contains results for BIC tuning parameter selection and, for MC+, universal tuning parameter selection (λ_{univ}). It is reasonable to compare the MC+/λ_{univ} procedure to BIC procedures because they all utilize the same amount of data (i.e. no validation data is used). However, before taking a closer look at the BIC and λ_{univ} results, we make some general observations about how the BIC and λ_{univ} results compare to the DV results. Keeping in mind that the DV procedures utilize more data than BIC and λ_{univ}, note that for LASSO, adaptive LASSO, MC+, and SELO, the correct model was selected considerably more often with BIC and λ_{univ}, while model error tended to be slightly lower when DV tuning parameter selection was used (one notable exception was that the model error for MC+/λ_{univ} was substantially higher than for the other methods). For example, for SELO with $\tau = 0.01$ and $n = 100$, the correct model was selected 17% more often when BIC tuning parameter selection was used, as opposed to DV tuning, while average model error was decreased by 0.014 when DV tuning is used, as opposed to BIC tuning. This is not surprising because BIC is consistent for model selection, while data validation tuning is designed to minimize prediction error. Focusing now on the BIC and λ_{univ} results, it is apparent that SELO/BIC outperformed all corresponding BIC and λ_{univ} methods in terms of selecting the correct model with the highest frequency. Furthermore, SELO/BIC performed comparably to the other BIC procedures in terms of model error. As mentioned above, MC+/λ_{univ} appeared to perform rather poorly in terms of model error.

Overall, the results in Table 1 indicate that SELO performs very favorably when compared with the other estimators, in terms of (i) selecting the correct model and (ii) model error. We also point out that the SELO simulation results are similar for fixed $\tau = 0.01$ and for tuning over $\tau \in \{0.001, 0.01, 0.1, 0.5\}$. This suggests that tuning over different values of τ may, in practice, offer only modest gains over a fixed τ approach.

For each of the 1000 simulated dataset, we also estimated the variance of the SELO estimators using (2). Results are reported in Table 2. When $n = 100$, the variance estimates corresponding to non-zero entries in β^* agreed well with the empirical variances. However, the performance of variance estimators corresponding to entries of β^* equal to zero and the performance of variance estimators when $n = 50$ was less reliable; in these situations, the variance estimates appeared to underestimate the true variability of SELO estimators. Similar phenomena were observed for the quadratic approximation based standard error estimators for the adaptive LASSO and SCAD estimators.

5.3 Simulation study II: $d = 20$

In our second study, we took $\beta^* \in \mathbb{R}^{20}$, with $\beta_1^* = 3$, $\beta_2^* = 1.5$, $\beta_5^* = 2$, and $\beta_j^* = 0$, if $j \neq 1$, 2, or 5. Thus $d = 20$ and, as in the previous example, $d_0 = 3$. We fixed $\sigma^2 = 9$ and for each value of n, with $n \in \{50, 100\}$, we simulated 1000 independent datasets $\{(y_1, x_1^T), ..., (y_n, x_n^T)\}$ and, for each dataset, we computed estimates of β^*. Results from this simulation study are found in Table 3.

The results in Table 3 are similar to the results in Table 1, indicating that when data validation was used for tuning parameter selection, SELO outperformed LASSO, adaptive LASSO, SCAD, and MC+ in terms of model selection criteria ("Correct model"); in terms of model error, when $n = 50$, SELO performed comparably to its competitors, while when $n = 100$, it performed comparably to adaptive LASSO, SCAD, and MC+, and all the four methods outperformed LASSO. In fact, in many settings, SELO selected the correct model far more frequently than the other procedures. SELO and SCAD seemed to perform the best in terms of model error – SELO with $\tau \in \{0.001, 0.01, 0.1, 0.5\}$ performed especially well.

When BIC and λ_{univ} tuning parameter selection was used, Table 3 indicates that SELO performed the best in terms of model selection criteria (except for the $n = 50$ setting, where λ_{univ} selected the correct model more frequently than the other methods; however, in this setting, MC+/λ_{univ} had the highest model error) and performed well in terms of model error.

The comparison of BIC and DV tuning parameter selection in Table 3 is more nuanced for the $d = 20$ simulations than for the $d = 8$ simulations in Table 1. Model error results for DV tuning are better than those for BIC tuning in all settings, which is to be expected. For LASSO, adaptive LASSO, and MC+, BIC consistently selected smaller models and selected the correct model more frequently than DV tuning. On the other hand, for SCAD and SELO, data validation selected the correct model more frequently than BIC in some settings. For instance, when $n = 100$ and SELO with $\tau = 0.01$ is implemented, BIC selected the correct model in 70.2% of all simulated datasets, while DV selected the correct model in 76.5% of all simulated datasets. We point out that the comparative performance gains of DV tuning in these settings may be related to the fact that DV utilizes more data than the BIC procedures.

Standard error estimates were also computed for SELO estimates in this simulation study. In the interests of brevity, these results are not fully reported here. However, we mention that standard error estimates corresponding to non-zero entries in β^* – β_1^* , β_2^*, and β_5^* – agreed well with empirical standard errors for $n = 100$, while the variance estimates for $n = 50$ were less reliable.

5.4 Simulation study III: Common BIC criterion

The simulation results described in Section 5.2 and 5.3 give an indication of the performance of the proposed SELO methods, in comparison with current recommended implementations (or reasonable variations thereof) of other PLS methods. Table 4 summarizes the results of a simulation study in which (6), the

BIC criterion used with for SELO, was used for each PLS method. Thus, results for SELO and MC+ contained in Table 4 were also reported in Tables 1 and 3. Results indicate that when this common BIC criterion is utilized, SELO performs well when compared to the alternative methods. Indeed, SELO outperforms all other methods in terms of model selection criteria and performs comparably in terms of model error.

Comparing the results for the different BIC procedures implemented with LASSO, adaptive LASSO, and SCAD, it appears that the common BIC criterion (6) typically leads to improved model selection performance, at the expense of a slight increase in model error for the simulation settings considered here. The most substantial differences are for SCAD. For instance, in the $d = 20$, $n = 50$ results, the frequency with which that correct model was selected by SCAD/BIC jumps from 0.043 in Table 2 to 0.251 in Table 4.

5.5 Simulation study IV: Large d analysis

We examined the performance of the various PLS methods for d substantially larger than in the previous studies. In particular, we took $d = 339$, $n = 800$, $\sigma^2 = 36$, and $\beta^* = (3J_{37}^T, -2J_{37}^T, 1J_{37}^T, 0J_{228}^T)^T \in \mathbb{R}^{339}$, where $J_k \in \mathbb{R}^k$ is the vector with all entries equal to 1. Thus, $d_0 = 111$. We used BIC tuning parameter selection in this study. For SELO and MC+, we used the BIC criterion (6). For each of LASSO, adaptive LASSO, and SCAD, we implemented two BIC selection procedures: The corresponding BIC procedure from simulations I-II and the common BIC procedure (6) that was used for all estimation methods in simulation III. For SELO, the tuning parameter $\tau = 0.01$ was taken to be fixed. We simulated 100 independent datasets $\{(y_1, x_1^T), ..., (y_n, x_n^T)\}$ and, for each dataset, we computed estimates of β^*. Results from this simulation study are found in Table 5.

Perhaps the most striking aspect of the results presented in Table 5 is that no method ever selected the correct model in this simulation study, but with d, d_0, and σ^2 substantially larger than in the previous simulation studies this is not too surprising. On average, SELO selected the most parsimonious models of all methods (average model size, 104.04) and had the smallest model error (11.038). Since $d_0 = 111$, it is clear that SELO underfit in some instances (indeed, this is reported in Table 5: $F- = 8.03$). In fact, all of the methods in this study underfit to some extent (i.e. $F- > 0$), perhaps due to the fact that many of the non-zero entries in β^* were small relative to the noise level $\sigma^2 = 36$.

Comparing the different BIC criteria for LASSO, adaptive LASSO, and SCAD, it is apparent that the common BIC criterion (6) (used in simulation III) selected a more parsimonious model than the BIC procedures that were described in Section 5.1 and used in simulations I-II at the cost of a somewhat inflated model error. This is not unexpected, because the degrees of freedom adjustment in the denominator of $\hat{\sigma}^2 = (n - \hat{d}_0)^{-1}||y - X\hat{\beta}||^2$, which is used in the common BIC criterion (6), inflates the BIC of larger models. In the present study, the diferences between the two BIC implementations for SCAD were especially notable. This may be related to the fact that for SCAD, the two BIC criteria use different estimates of σ^2 and

different $\widehat{\mathrm{DF}}$, while in the BIC criteria for LASSO and adaptive LASSO, only the estimates of σ^2 differ.

5.6 HIV-1 drug resistance and codon mutation

Rhee et al. (2006) describe a publicly available dataset that contains HIV-1 drug resistance and mutation information on 848 HIV-1 isolates. We investigated the association between resistance to the drug Amprenavir, a protease inhibitor, and the presence of mutations at various protease codons using SELO/BIC and other methods. Identifying codon mutations that effect drug resistance and estimating the magnitudes of these effects is clinically important. Our outcome of interest was IC_{50}, a continuous measure of HIV-1 drug resistance; higher IC_{50} corresponds to greater drug resistance. The dataset contains mutation information for 99 protease codons. We used a binary predictor to indicate the presence of a mutation at each codon location. After removing samples with missing data, and codons with fewer than three observed mutations, the resulting dataset contained $n = 768$ samples and $d = 76$ predictor codons. We analyzed this dataset.

After taking the logarithm of IC_{50}, centering the data, and standardizing the predictors to have length n, we found SELO, LASSO, adaptive LASSO, SCAD, and MC+ (with $\gamma = 2/(1-\max_{i\neq j}|X_i^T X_j|/n)$) estimates. The BIC criterion (6) was used to select tuning parameters. We also computed the OLS estimate of the codon mutations' effect on IC_{50}. Results are summarized in Table 4: the OLS estimator does not perform variable selection; estimated effects are non-zero for all 76 codons. LASSO, adaptive LASSO, SCAD, and MC+ select 23, 14, 24, and 24 codon mutations associated with Amprenavir resistance, respectively, out of 76 total codon mutations. SELO, selects 16 mutations associated with Amprenavir resistance, a substantially simpler model than LASSO, SCAD, and MC+, and only slightly more complex than the adaptive LASSO. Furthermore, as indicated by the columns labeled R^2 and R^2/R^2_{OLS} in Table 6, the SELO estimator describes more variability in the data than any of the PLS alternatives, and nearly as much as the OLS estimator.

The codons selected by the SELO procedure are displayed in Table 7, along with the corresponding point estimates, standard errors, and p-values based on a normal approximation (which is justified by Theorem 1). All of the codons selected by SELO are selected by adaptive LASSO. SELO selects two codons which are not selected by LASSO (codons 64 and 71) and one codon not selected by SCAD or MC+ (codon 71). All of the codons selected by SELO are highly significant. Furthermore, mutations at codons 10, 32, 46, 47, 50, 54, 76, 84, and 90 are known to be associated with Amprenavir resistance (Johnson et al. (2008)). Overall, in this analysis, SELO leads to a relatively simple model of the association between drug resistance and codon mutations that explains the data very nearly as well as substantially more complex models.

6 Discussion

The concavity parameter. Both SELO and MC+ depend on λ and a concavity parameter (τ for SELO and γ

for MC+). In principle, when implementing these PLS methods, one may tune over a fine two-dimensional grid comprised of different values of λ and the concavity parameter. However, the computational cost of two-dimensional tuning parameter selection may become burdensome. Our numerical results suggest that the performance of SELO is fairly robust to the choice of τ and we have found that $\tau = 0.01$ seems to give reasonable results in all of simulation settings considered here, along with the data analysis.

One-step methods. Following Zou and Li (2008), a one-step version of the SELO procedure may be implemented. Like the `SELO-CD-PATH` procedure, the one-step version has the oracle property of Fan and Li (2001). In simulations we have found that `SELO-CD-PATH` outperforms the one-step procedure regardless of the tuning parameter selection method, perhaps because of the one-step procedure's dependence on an initial estimator. Thus, we recommend using `SELO-CD-PATH` over one-step procedures, especially for relatively noisy data where the initial estimator utilized by the one-step procedure may be unreliable.

L_2-regularization. Motivated by the elastic net (Zou and Hastie (2005)) and the adaptive elastic net (Zou and Zhang (2009)), one could consider a mixed penalty involving p_{SELO} and an L_2-norm penalty:

$$\lambda_1 ||\beta||^2 + \sum_{j=1}^{d} p_{\text{SELO},\lambda_2,\tau}(\beta_j).$$

The elastic net and the adaptive elastic net have been observed to outperform PLS methods that do not involve an L_2 penalty in a variety of settings. Studying the performance of PLS methods that utilize SELO and an L_2 norm penalty is a potentially interesting area for future research.

More predictors than observations. In this paper, we do not address the situation where $d > n$. Existing theoretical results for PLS methods that utilize other penalty functions and are valid for $d > n$ typically require more stringent conditions on the matrix $n^{-1}X^TX$, the sparsity level (d_0 must not be too large), and the distribution of the additive error ϵ_i. If one adopts additional assumptions in this direction, it is plausible that results similar to those found in (Kim, Choi, and Oh (2008)), (Zhang (2010)), and (Fan and Lv (2011)), which apply to other PLS methods, may hold for SELO. Further research in this direction is needed. It is worth pointing out that existing $d > n$ results for PLS methods with a general penalty function do not appear to apply directly to SELO.

GLMs. The SELO procedure may be extended to generalized linear models (McCullagh and Nelder (1989)). As with linear models, one can prove that SELO for GLMs is consistent for model selection and asymptotically efficient, provided tuning parameters follow the appropriate rate. For GLMs, SELO estimators may be found using a coordinate descent algorithm and tuning parameters may be selected to minimize a BIC criteria based on the model deviance or negative log-likelihood. Small scale simulation studies indicate that SELO for GLMs performs well.

Acknowledgments

The authors would like to thank the editors and three referees for their many helpful comments. The authors also thank Zilin Li for pointing out an error in an earlier version of the paper. This research was supported by grants from the US National Cancer Institute and the National Institute of Environmental Health.

Appendix

To prove Theorem 1, we argue as in (Fan and Peng, 2004). Lemmas 1 and 2 imply that there is a $\sqrt{n/(d\sigma^2)}$-consistent local minimizer of the SELO objective that consistently selects A; then Theorem 1 follows readily.

Lemma 1. *Assume that conditions (A)-(D) hold and let*

$$Q_n(\beta) = \frac{1}{2n}||y - X\beta||^2 + \sum_{j=1}^{d} p_{\mathrm{SELO}}(\beta_j). \tag{9}$$

Then for every $r \in (0,1)$, there exists a constant $C_0 > 0$ such that

$$\liminf_{n\to\infty} P\left[\operatorname*{argmin}_{||\beta-\beta^*||\le C\sqrt{d\sigma^2/n}} Q_n(\beta) \subseteq \left\{\beta \in \mathbb{R}^d;\ ||\beta-\beta^*|| < C\sqrt{d\sigma^2/n}\right\}\right] > 1-r,$$

whenever $C \ge C_0$.

Proof. Let $\alpha_n = \sqrt{d\sigma^2/n}$ and fix $r \in (0,1)$. To prove the lemma, it suffices to show that if $C > 0$ is large enough, then

$$Q_n(\beta^*) < \inf_{||u||=1} Q_n(\beta^* + C\alpha_n u)$$

holds for all n sufficiently large, with probability at least $1-r$. Define $D_n(u) = Q_n(\beta^* + C\alpha_n u) - Q_n(\beta^*)$ and note that

$$\begin{aligned} D_n(u) &= \frac{1}{2n}\left(C^2\alpha_n^2||Xu||^2 - 2C\alpha_n\epsilon^T Xu\right) \\ &\qquad + \sum_{j=1}^{d}\left[p_{\mathrm{SELO}}(\beta_j^* + C\alpha_n u_j) - p_{\mathrm{SELO}}(\beta_j^*)\right] \\ &\ge \frac{1}{2n}\left(C^2\alpha_n^2||Xu||^2 - 2C\alpha_n\epsilon^T Xu\right) \\ &\qquad + \sum_{j\in K(u)}\left[p_{\mathrm{SELO}}(\beta_j^* + C\alpha_n u_j) - p_{\mathrm{SELO}}(\beta_j^*)\right], \end{aligned}$$

where $K(u) = \{j; p_{\mathrm{SELO}}(\beta_j^* + C\alpha_n u_j) - p_{\mathrm{SELO}}(\beta_j^*) < 0\}$. Condition (B) and the fact that p_{SELO} is concave

on $[0,\infty)$ imply that, for each C, $p_{\mathrm{SELO}}(\beta_j^* + C\alpha_n u_j) - p_{\mathrm{SELO}}(\beta_j^*) \geq -C\alpha_n|u_j|p'_{\mathrm{SELO}}(\beta_j^* + C\alpha_n u_j)$ for all $||u|| = 1$ and $j \in K(u)$, when n is sufficiently large. Thus, for n large enough,

$$\begin{aligned} D_n(u) &\geq \frac{1}{2n}\left(C^2\alpha_n^2||Xu||^2 - 2C\alpha_n\epsilon^T Xu\right) - \sum_{j\in K(u)} C\alpha_n|u_j|p'_{\mathrm{SELO}}(\beta_j + C\alpha_n u_j) \\ &\geq \frac{1}{2n}\left(C^2\alpha_n^2||Xu||^2 - 2C\alpha_n\epsilon^T Xu\right) - \frac{Cd\lambda\tau\alpha_n}{\rho^2\log(2)}. \end{aligned} \tag{10}$$

By (C),

$$\frac{1}{2n}C^2\alpha_n^2||Xu||^2 \geq \frac{\lambda_{min}(n^{-1}X^TX)}{2}C^2\alpha_n^2, \tag{11}$$

$$\frac{1}{n}C\alpha_n|\epsilon^T Xu| \leq \frac{C\alpha_n}{\sqrt{n}}\left|\left|\frac{1}{\sqrt{n}}X^T\epsilon\right|\right| = O_P(C\alpha_n^2). \tag{12}$$

Furthermore, (D) implies

$$\frac{Cd\lambda\tau\alpha_n}{\rho^2\log(2)} = o(C\alpha_n^2). \tag{13}$$

From (10)-(13), we conclude that if $C > 0$ is large enough, then $\inf_{||u||=1} D_n(u) > 0$ holds for all n sufficiently large, with probability at least $1 - r$. This proves the lemma. □

Lemma 2. *Assume that (A)-(D) hold, let $Q_n(\beta)$ be as at (9), and fix $C > 0$. Then*

$$\lim_{n\to\infty} P\left[\operatorname*{argmin}_{||\beta-\beta^*||\leq C\sqrt{d\sigma^2/n}} Q_n(\beta) \subseteq \{\beta\in\mathbb{R}^d;\ \beta_{A^c} = 0\}\right] = 1,$$

where $A^c = \{1,...,d\}\setminus A$ is the complement of A in $\{1,...,d\}$.

Proof. Suppose that $\beta\in\mathbb{R}^d$ and that $||\beta-\beta^*|| < C\sqrt{d\sigma^2/n}$. Define $\tilde\beta\in\mathbb{R}^d$ by $\tilde\beta_{A^c} = 0$ and $\tilde\beta_A = \beta_A$. Similar to the proof of Lemma 1, if $D_n(\beta,\tilde\beta) = Q_n(\beta) - Q_n(\tilde\beta)$, then

$$\begin{aligned} D_n(\beta,\tilde\beta) &= \frac{1}{2n}||y - X\beta||^2 - \frac{1}{2n}||y - X\tilde\beta||^2 + \sum_{j\in A^c} p_{\mathrm{SELO}}(\beta_j) \\ &= \frac{1}{2n}||y - X\tilde\beta - X(\beta-\tilde\beta)||^2 - \frac{1}{2n}||y - X\tilde\beta||^2 + \sum_{j\in A^c} p_{\mathrm{SELO}}(\beta_j) \\ &= \frac{1}{2n}(\beta-\tilde\beta)^T X^TX(\beta-\tilde\beta) - \frac{1}{n}(\beta-\tilde\beta)^T X^T(y - X\tilde\beta) + \sum_{j\in A^c} p_{\mathrm{SELO}}(\beta_j) \\ &= O_P\left(||\beta-\tilde\beta||\sqrt{d\sigma^2/n}\right) + \sum_{j\in A^c} p_{\mathrm{SELO}}(\beta_j). \end{aligned} \tag{14}$$

On the other hand, since the SELO penalty is concave,

$$p_{\mathrm{SELO}}(\beta_j) \geq \frac{\lambda}{\log(2)}\log\left[\frac{C}{C + \tau\sqrt{n/(d\sigma^2)}} + 1\right]|\beta_j|$$

for $j \in A^c$. Thus,

$$\sum_{j \in A^c} p_{\mathrm{SELO}}(\beta_j) \geq \frac{\lambda}{\log(2)} \log \left[\frac{C}{C + \tau\sqrt{n/(d\sigma^2)}} + 1 \right] ||\beta - \tilde{\beta}||. \tag{15}$$

By (D),

$$\liminf_{n\to\infty} \log \left[\frac{C}{C + \tau\sqrt{n/(d\sigma^2)}} + 1 \right] > 0.$$

It follows from (14)-(15) that there is a constant $\tilde{C} > 0$ such that

$$\frac{D_n(\beta, \tilde{\beta})}{||\beta - \tilde{\beta}||} \geq \tilde{C}\lambda + O_P(\sqrt{d\sigma^2/n}).$$

Since $\lambda\sqrt{n/(d\sigma^2)} \to \infty$ by condition (D), the result follows. □

Proof of Theorem 1. Taken together, Lemmas 1 and 2 imply that there exists a sequence of local minima $\hat{\beta}$ of (SELO) such that $||\hat{\beta} - \beta^*|| = O_P(\sqrt{d\sigma^2/n})$ and $\hat{\beta}_{A^c} = 0$. Indeed one may take $\hat{\beta}$ to be the element of $M_n = \{\text{local minima of (SELO)}\} \cap \{\beta \in \mathbb{R}^d;\ \beta_{A^c} = 0\}$ that is closest to β^* in ℓ^2-norm ($\hat{\beta}$ may be defined to be any local minima of (SELO) on the event the $M_n = \emptyset$). Part (i) of the theorem follows immediately. To prove part (ii), observe that on the event $\{j;\ \hat{\beta}_j \neq 0\} = A$, basic calculus implies that we must have

$$\hat{\beta}_A = \beta^*_A + (X_A^T X_A)^{-1} X_A^T \epsilon - (n^{-1} X_A^T X_A)^{-1} p'_A(\hat{\beta}),$$

where $p'_A(\hat{\beta}) = (p'_{\mathrm{SELO}}(\hat{\beta}_j))_{j\in A}$. It follows that

$$\begin{aligned} \sqrt{n} B_n (n^{-1} X_A^T X_A/\sigma^2)^{1/2} (\hat{\beta}_A - \beta^*_A) \quad &= \quad B_n (\sigma^2 X_A^T X_A)^{-1/2} X_A^T \epsilon \\ &\quad -nB_n (\sigma^2 X_A^T X_A)^{-1/2} p'_A(\hat{\beta}) \end{aligned}$$

whenever $\{j;\ \hat{\beta}_j \neq 0\} = A$. Now note that conditions (B)-(D) imply that

$$||n(\sigma^2 X_A^T X_A)^{-1/2} p'_A(\hat{\beta})|| = O_P\left(\sqrt{nd/\sigma^2}\frac{\lambda\tau}{\rho^2}\right) = o_P(1). \tag{16}$$

Thus,

$$\sqrt{n} B_n (n^{-1} X_A^T X_A/\sigma^2)^{1/2} (\hat{\beta}_A - \beta^*_A) = B_n (\sigma^2 X_A^T X_A)^{-1/2} X_A^T \epsilon + o_P(1).$$

To complete the proof of (ii), we use the Lindeberg-Feller CLT (Durrett, 2005) to show that

$$B_n (\sigma^2 X_A^T X_A)^{-1/2} X_A^T \epsilon \to N(0, G) \tag{17}$$

in distribution. Observe that

$$B_n(\sigma^2 X_A^T X_A)^{-1/2} X_A^T \epsilon = \sum_{i=1}^n w_{i,n},$$

where $w_{i,n} = B_n(\sigma^2 X_A^T X_A)^{-1/2} x_{i,A}\epsilon_i$. Fix $\delta_0 > 0$ and let $\eta_{i,n} = x_{i,A}^T (X_A^T X_A)^{-1/2} B_n^T B_n (X_A^T X_A)^{-1/2} x_{i,A}$. Then

$$\begin{aligned} E[||w_{i,n}||^2;\ ||w_{i,n}||^2 > \delta_0] &= \eta_{i,n} E[\epsilon_i^2/\sigma^2;\ \eta_{i,n}\epsilon_i^2/\sigma^2 > \delta_0] \\ &\leq \eta_{i,n} E\left(|\epsilon_i/\sigma|^{2+\delta}\right)^{2/(2+\delta)} P\left\{\eta_{i,n}\epsilon_i^2/\sigma^2 > \delta_0\right\}^{\delta/(2+\delta)} \\ &\leq \eta_{i,n}^{1+\delta/(2+\delta)} \delta_0^{-1} E\left(|\epsilon_i/\sigma|^{2+\delta}\right)^{2/(2+\delta)}. \end{aligned}$$

Since $\sum_{i=1}^n \eta_{i,n} = \mathrm{tr}(B_n^T B_n) \to \mathrm{tr}(G) < \infty$, and since (E) implies

$$\max_{1\leq i\leq n} \eta_{i,n} \leq \lambda_{min}(n^{-1}X^T X)\lambda_{max}(B_n^T B_n) \max_{1\leq i\leq n} \frac{1}{n}\sum_{j=1}^d x_{ij}^2 \to 0,$$

we must have

$$\begin{aligned} \sum_{i=1}^n E[||w_{i,n}||^2;\ ||w_{i,n}||^2 > \delta_0] &\leq \delta_0^{-1} E\left(|\epsilon_i/\sigma|^{2+\delta}\right)^{2/(2+\delta)} \sum_{i=1}^n \eta_{i,n}^{1+\delta/(2+\delta)} \\ &\leq \delta_0^{-1} E\left(|\epsilon_i/\sigma|^{2+\delta}\right)^{2/(2+\delta)} \mathrm{tr}(B_n^T B_n) \max_{1\leq i\leq n} \eta_{i,n}^{\delta/(2+\delta)} \\ &\to 0. \end{aligned}$$

Thus, the Lindeberg condition is satisfied and (17) holds. □

Proof of Theorem 2. Our proof is similar to (Wang, Li, and Leng, 2009): First, we study the BIC corresponding to estimators which fail to select all of the significant variables (underfitting); second, we consider estimators that select too many variables (overfitting). More specifically, we show that estimators that underfit have BIC larger than the OLS estimator fit to the full model. The overfit case is more delicate and requires bounds on the maximum of a collection of random variables (see Lemmas 3-4). Wang, Li, and Leng's (2009) BIC consistency results only apply to gaussian ϵ_i and $k_n/\log(n) \to \infty$. We provide a slightly more general analysis of the overfit case, which enables us to extend their results to non-gaussian ϵ_i and $k_n = \log(n)$.

Without loss of generality, suppose $k_n \geq 1$ and suppose that there is a point $(\lambda_0, \tau_0) \in \Omega$ with $\lambda_0 = 0$. Let $\hat{\beta}_0 = \hat{\beta}(\lambda_0, \tau_0) = (X^T X)^{-1} X^T y$ be the OLS estimator and let $\hat{\beta} = \hat{\beta}(\lambda, \tau)$ be a local minimizer of (SELO) with $(\lambda, \tau) \in \Omega$. Define $\hat{A} = \hat{A}(\lambda, \tau) = \{j;\ \hat{\beta}_j \neq 0\}$ to be the model selected by $\hat{\beta}(\lambda, \tau)$. The first order optimality conditions for SELO imply that

$$\hat{\beta}_{\hat{A}} = (X_{\hat{A}}^T X_{\hat{A}})^{-1} X_{\hat{A}}^T y - n(X_{\hat{A}}^T X_{\hat{A}})^{-1} p'_{\hat{A}}(\hat{\beta}_{\hat{A}}) = \tilde{\beta}_{\hat{A}} - n(X_{\hat{A}}^T X_{\hat{A}})^{-1} p'_{\hat{A}}(\hat{\beta}_{\hat{A}}),$$

where $p'_{\hat{A}}(\beta) = (p'_{\mathrm{SELO}}(\beta_j))_{j\in\hat{A}}$ and $\tilde{\beta}$ is the OLS estimator corresponding to the selected model $\hat{A}$; that is, $\tilde{\beta}_{\hat{A}^c} = 0$ and $\tilde{\beta}_{\hat{A}} = (X_{\hat{A}}^T X_{\hat{A}})^{-1} X_{\hat{A}}^T y$. Then

$$\begin{aligned} ||y - X\hat{\beta}||^2 &= ||y - X\tilde{\beta}||^2 + n^2 p'_{\hat{A}}(\hat{\beta})^T (X_{\hat{A}}^T X_{\hat{A}})^{-1} p'_{\hat{A}}(\hat{\beta}) \\ &= ||y - X\hat{\beta}_0||^2 + ||X(\hat{\beta}_0 - \tilde{\beta})||^2 + n^2 p'_{\hat{A}}(\hat{\beta})^T (X_{\hat{A}}^T X_{\hat{A}})^{-1} p'_{\hat{A}}(\hat{\beta}). \end{aligned} \tag{18}$$

Our first goal is to find a lower bound for $||y - X\hat{\beta}||^2 - ||y - X\hat{\beta}_0||^2$ in cases where $A \setminus \hat{A} \neq \emptyset$. If $A \setminus \hat{A} \neq \emptyset$, then we have

$$\begin{aligned} ||y - X\hat{\beta}||^2 - ||y - X\hat{\beta}_0||^2 &= ||X(\hat{\beta}_0 - \tilde{\beta})||^2 + n^2 p'_{\hat{A}}(\hat{\beta})^T (X_{\hat{A}}^T X_{\hat{A}})^{-1} p'_{\hat{A}}(\hat{\beta}) \\ &\geq nr_0 ||\hat{\beta}_0 - \tilde{\beta}||^2 \\ &= nr_0 \left\{ ||\hat{\beta}_0 - \beta^*||^2 - 2(\hat{\beta}_0 - \beta^*)^T (\tilde{\beta} - \beta^*) + ||\tilde{\beta} - \beta^*||^2 \right\} \\ &\geq nr_0 \left\{ ||\tilde{\beta} - \beta^*||^2 - 2(\hat{\beta}_0 - \beta^*)^T (\tilde{\beta} - \beta^*) \right\} \\ &\geq nr_0 ||\tilde{\beta} - \beta^*|| \left(||\tilde{\beta} - \beta^*|| - 2||\hat{\beta}_0 - \beta^*|| \right) \\ &\geq nr_0 \rho^2 \left(1 - 2\frac{||\hat{\beta}_0 - \beta^*||}{\rho} \right), \end{aligned}$$

where $0 < r_0 < \lambda_{min}(n^{-1}X^T X)$ is defined in condition (C). We use this bound to obtain a lower bound on the difference $\mathrm{BIC}_{k_n}(\hat{\beta}) - \mathrm{BIC}_{k_n}(\hat{\beta}_0)$. Using the fact that $\log(x) \geq 1 - x^{-1}$ for any $x > 0$, we have

$$\begin{aligned} \mathrm{BIC}_{k_n}(\hat{\beta}) - \mathrm{BIC}_{k_n}(\hat{\beta}_0) &= \frac{k_n}{n}(|\hat{A}| - d) + \log\left\{ \frac{(n-d)||y - X\hat{\beta}||^2}{(n-\hat{A})||y - X\hat{\beta}_0||^2} \right\} \\ &\geq 1 - \frac{k_n d}{n} - \frac{n||y - X\hat{\beta}_0||^2}{(n-d)||y - X\hat{\beta}||^2} \\ &= \frac{1}{||y - X\hat{\beta}||^2} \left\{ \left(1 - \frac{k_n d}{n} \right) ||y - X\hat{\beta}||^2 - \frac{n}{n-d} ||y - X\hat{\beta}_0||^2 \right\} \\ &\geq \frac{1}{||y - X\hat{\beta}||^2} \left\{ nr_0\rho^2 \left(1 - \frac{k_n d}{n} \right) \left(1 - 2\frac{||\hat{\beta}_0 - \beta^*||}{\rho} \right) \right. \\ &\quad \left. - \left(\frac{d}{n-d} + \frac{k_n d}{n} \right) ||y - X\hat{\beta}_0||^2 \right\}, \end{aligned}$$

whenever $A \setminus \hat{A} \neq \emptyset$ and $k_n d/n < 1$. Thus, when $k_n d/n < 1$,

$$nr_0\rho^2 \left(1 - \frac{k_n d}{n} \right) \left(1 - 2\frac{||\hat{\beta}_0 - \beta^*||}{\rho} \right) - \left(\frac{d}{n-d} + \frac{k_n d}{n} \right) ||y - X\hat{\beta}_0||^2 > 0 \tag{19}$$

implies

$$\inf\left\{\mathrm{BIC}_{k_n}\{\hat{\beta}(\lambda,\tau)\};\ (\lambda,\tau)\in\Omega,\ A\setminus\hat{A}(\lambda,\tau)\right\} > \mathrm{BIC}_{k_n}(\hat{\beta}_0).$$

Since $||\hat{\beta}_0-\beta^*|| = O_P(\sqrt{d\sigma^2/n})$ and $||y-X\hat{\beta}_0||^2 = O_P(n\sigma^2)$, it follows from conditions (A2)-(B2) that (19) holds with probability tending to 1. We conclude that

$$P\left[\inf\left\{\mathrm{BIC}_{k_n}\{\hat{\beta}(\lambda,\tau)\};\ (\lambda,\tau)\in\Omega,\ A\setminus\hat{A}(\lambda,\tau)\right\} > \mathrm{BIC}_{k_n}(\hat{\beta}_0)\right]\to 1.$$

Therefore, to prove the theorem it suffices to consider overfit models and show that

$$P\left[\inf\left\{\mathrm{BIC}_{k_n}\{\hat{\beta}(\lambda,\tau)\};\ (\lambda,\tau)\in\Omega,\ A\subsetneq\hat{A}(\lambda,\tau)\right\} > \mathrm{BIC}_{k_n}(\hat{\beta}^*)\right]\to 1,$$

where $\hat{\beta}^*$ is a local minimizer of (SELO) with the properties described in Theorem 1.

Recall from Theorem 1 that $P(\{j;\ \hat{\beta}^*_j\neq 0\}=A)\to 1$. In the overfit case, we compare the BIC of the estimator $\hat{\beta}$, when $A\subsetneq\hat{A}$, to the BIC of $\hat{\beta}^*$, when $\{j;\ \hat{\beta}^*_j\neq 0\}=A$. Thus, assume that we are on the event $\{j;\ \hat{\beta}^*_j\neq 0\}=A$ and $A\subsetneq\hat{A}$. Since

$$\begin{aligned}
\log\left\{\frac{||y-X\hat{\beta}||^2}{||y-X\hat{\beta}^*||^2}\right\} &\geq \log\left\{\frac{||y-X\tilde{\beta}||^2}{||y-X\hat{\beta}^*||^2}\right\}\\
&\geq \frac{||y-X\tilde{\beta}||^2-||y-X\hat{\beta}^*||^2}{||y-X\tilde{\beta}||^2}\\
&\geq -\frac{\left|||y-X\tilde{\beta}||^2-||y-X\hat{\beta}^*||^2\right|}{||y-X\hat{\beta}_0||^2},
\end{aligned}$$

it follows that

$$\begin{aligned}
\mathrm{BIC}_{k_n}(\hat{\beta})-\mathrm{BIC}_{k_n}(\hat{\beta}^*) &= \frac{k_n}{n}(|\hat{A}|-|A|)+\log\left(\frac{n-|A|}{n-|\hat{A}|}\right)+\log\left\{\frac{||y-X\hat{\beta}||^2}{||y-X\hat{\beta}^*||^2}\right\}\\
&\geq \frac{k_n}{n}(|\hat{A}|-|A|)-\frac{\left|||y-X\tilde{\beta}||^2-||y-X\hat{\beta}^*||^2\right|}{||y-X\hat{\beta}_0||^2}.
\end{aligned}$$

We study the last term on the right-hand side above in more detail. By condition (C), there exists a constant $R_1\in\mathbb{R}$ such that $\max_{1\leq j\leq d} n^{-1}||X_j||^2\leq R_1$. Notice that by (16),

$$\begin{aligned}
||y-X\hat{\beta}^*||^2 &= \epsilon^T\left\{I-X_A(X_A^TX_A)^{-1}X_A^T\right\}\epsilon+n^2p'_A(\hat{\beta}^*)^T(X_A^TX_A)^{-1}p'_A(\hat{\beta}^*)\\
&= \epsilon^T\left\{I-X_A(X_A^TX_A)^{-1}X_A^T\right\}\epsilon+o_P(\sigma^2)
\end{aligned}$$

and

$$\begin{aligned}\left|||y-X\tilde{\beta}||^2-||y-X\hat{\beta}^*||^2\right| &= \epsilon^T\left\{X_{\hat{A}}(X_{\hat{A}}^TX_{\hat{A}})^{-1}X_{\hat{A}}^T-X_A(X_A^TX_A)^{-1}X_A^T\right\}\epsilon+o_P(\sigma^2)\\ &\leq \frac{1}{nr_0}\left|\left|X_{\hat{A}\setminus A}^T\left\{I-X_A(X_A^TX_A)^{-1}X_A^T\right\}\epsilon\right|\right|^2+o_P(\sigma^2)\\ &\leq \frac{R_1}{r_0}(|\hat{A}|-|A|)\max_{j\notin A}(\epsilon^Tu^{(j)})^2+o_P(\sigma^2),\end{aligned}$$

where the unit vector $u^{(j)}\in\mathbb{R}^n$ is defined by

$$\left|\left|\left\{I-X_A(X_A^TX_A)^{-1}X_A^T\right\}X_j\right|\right|u^{(j)}=\left\{I-X_A(X_A^TX_A)^{-1}X_A^T\right\}X_j,\ \ j\notin A.$$

Combining this with the fact that $||y-X\hat{\beta}_0||^2=\sigma^2(n-d)\{1+o_P(1)\}$, we obtain

$$\mathrm{BIC}_{k_n}(\hat{\beta})-\mathrm{BIC}_{k_n}(\hat{\beta}^*)\geq\frac{|\hat{A}|-|A|}{\sigma^2(n-d)\{1+o_P(1)\}}\left[\sigma^2k_n\left(1-\frac{d}{n}\right)\{1+o_P(1)\}-\frac{R_1}{r_0}\max_{j\notin A}(\epsilon^Tu^{(j)})^2\right].$$

Thus, (4) follows if we can prove

$$\sigma^2k_n\left(1-\frac{d}{n}\right)\{1+o_P(1)\}-\frac{R_1}{r_0}\max_{j\notin A}(\epsilon^Tu^{(j)})^2>0,$$

with probability tending to 1. This will follow if we can show that there is some constant $0<c<1$ such that

$$\lim_{n\to\infty}P\left\{\frac{R_1}{k_nr_0}\max_{j\notin A}\left(\frac{\epsilon^Tu^{(j)}}{\sigma}\right)^2\geq 1-c\right\}=0.$$

Now, Lemma 3 below implies that if $E|\epsilon_i/\sigma|^{2+\delta}<C$, then there is a constant K such that

$$\begin{aligned}P\left\{\frac{R_1}{k_nr_0}\max_{j\notin A}\left(\frac{\epsilon^Tu^{(j)}}{\sigma}\right)^2\geq 1-c\right\} &\leq d\left\{\frac{R_1}{(1-c)k_nr_0}\right\}^{1+\delta/2}\sup_{||u||=1}E\left|\frac{u^T\epsilon}{\sigma}\right|^{2+\delta}\\ &\leq d\left\{\frac{R_1}{(1-c)k_nr_0}\right\}^{1+\delta/2}K.\end{aligned}$$

Since the right-hand side above converges to 0 for any fixed $0<c<1$, part (a) of the theorem follows. If ϵ_i is subgaussian, then Lemma 4 below implies that

$$\begin{aligned}P\left\{\frac{R_1}{k_nr_0}\max_{j\notin A}\left(\frac{\epsilon^Tu^{(j)}}{\sigma}\right)^2\geq 1-c\right\} &\leq d\sup_{||u||=1}P\left\{\frac{R_1}{k_nr_0}\left(\frac{u^T\epsilon}{\sigma}\right)^2\geq 1-c\right\}\\ &\leq 2d\exp\left\{-\frac{(1-c)k_n\sigma^2r_0}{2\sigma_0^2R_1}\right\}.\end{aligned}$$

Parts (b) and (c) of the theorem now follow readily. □

Lemma 3. *Suppose that $\epsilon_1, ..., \epsilon_n$ are iid, $E(\epsilon_i) = 0$, and $E|\epsilon_i|^{2+\delta} \leq C < \infty$ for some constants $C, \delta > 0$. Let $\epsilon = (\epsilon_1, ..., \epsilon_n)^T \in \mathbb{R}^n$. Then there is a constant K such that*

$$\sup_{||u||=1} E|u^T \epsilon|^{2+\delta} < K. \tag{20}$$

Proof. This lemma is a modified version of the main result in (Dharmadhikari, Fabian, and Jogdeo, 1968). Without loss of generality, assume that $C \geq 1$. Now fix

$$K \geq \left\{2^\delta (2+\delta)(1+\delta)C\right\}^{1+\delta/2}.$$

We prove that (20) holds by induction on n. Clearly, (20) is true for $n = 1$. Now assume that (20) is true for some $n - 1 \geq 1$. Fix $u_1, ..., u_n \in \mathbb{R}$ such that $u_1^2 + \cdots + u_n^2 = 1$. Without loss of generality, assume that $|u_n| < 1$ and define $\tilde{u}_i = (1 - u_n^2)^{-1/2} u_i$, $i = 1, ..., n-1$. Let

$$S_n = \sum_{i=1}^{n} u_i \epsilon_i \text{ and } S_{n-1} = \sum_{i=1}^{n-1} \tilde{u}_i \epsilon_i.$$

Then

$$S_n = \sqrt{1 - u_n^2} S_{n-1} + u_n \epsilon_n$$

and, by assumption (the induction hypothesis),

$$E|S_{n-1}|^{2+\delta} \leq K.$$

By Taylor's theorem

$$\begin{aligned} |S_n|^{2+\delta} &= \left|\sqrt{1-u_n^2} S_{n-1}\right|^{2+\delta} + (2+\delta)\text{sign}(S_{n-1}) \left|\sqrt{1-u_n^2} S_{n-1}\right|^{1+\delta} u_n \epsilon_n \\ &\quad + \frac{1}{2}(2+\delta)(1+\delta) \left|\sqrt{1-u_n^2} S_{n-1} + \theta u_n \epsilon_n\right|^{\delta} u_n^2 \epsilon_n^2, \end{aligned}$$

where $0 \leq \theta \leq 1$. Since

$$\left|\sqrt{1-u_n^2} S_{n-1} + \theta u_n \epsilon_n\right|^{\delta} \leq 2^\delta \left\{(1-u_n^2)^{\delta/2} |S_{n-1}|^\delta + |u_n|^\delta |\epsilon_n|^\delta\right\}$$

and

$$E|S_{n-1}|^\delta \epsilon_n^2 \leq \left(E|S_{n-1}|^{2+\delta}\right)^{\delta/(2+\delta)} \left(E|\epsilon_n|^{2+\delta}\right)^{2/(2+\delta)} \leq K^{\delta/(2+\delta)} C^{2/(2+\delta)},$$

it follows that

$$\begin{aligned} E|S_n|^{2+\delta} &\leq (1-u_n^2)^{1+\delta/2}E|S_{n-1}|^{2+\delta} + 2^{\delta-1}(2+\delta)(1+\delta)u_n^2\left\{E|S_{n-1}|^{\delta}\epsilon_n^2 + E|\epsilon_n|^{2+\delta}\right\} \\ &\leq (1-u_n^2)K + 2^{\delta-1}(2+\delta)(1+\delta)u_n^2\left\{K^{\delta/(2+\delta)}C^{2/(2+\delta)} + C\right\} \\ &\leq (1-u_n^2)K + 2^{\delta}(2+\delta)(1+\delta)CK^{\delta/(2+\delta)}u_n^2 \\ &\leq K\left\{1 - u_n^2 + 2^{\delta}(2+\delta)(1+\delta)CK^{-2/(2+\delta)}u_n^2\right\} \\ &\leq K. \end{aligned}$$

This proves the lemma. □

Lemma 4. *Suppose that $\epsilon_1, ..., \epsilon_n$ are iid with $E\epsilon_i = 0$. Suppose further that the distribution of ϵ_i is subgaussian with scale $\sigma_0^2 > 0$, in the sense that*

$$Ee^{t\epsilon_i} \leq e^{\sigma_0^2 t^2/2}, \;\; t \in \mathbb{R}.$$

Let $\epsilon = (\epsilon_1, ..., \epsilon_n)^T \in \mathbb{R}^n$. Then

$$\sup_{||u||=1} P(|u^T\epsilon| \geq c) \leq 2e^{-c^2/(2\sigma_0^2)}.$$

Proof. Fix $u = (u_1, ..., u_n)^T \in \mathbb{R}^n$ with $||u|| = 1$. Then for $t > 0$,

$$\begin{aligned} P(|u^T\epsilon| \geq c) &= P(u^T\epsilon \geq c) + P(u^T\epsilon \leq -c) \\ &\leq e^{-ct}\left(Ee^{tu^T\epsilon} + Ee^{-tu^T\epsilon}\right) \\ &\leq 2e^{-ct}e^{\sigma_0^2 t^2/2}. \end{aligned}$$

The lemma follows by taking $t = c/\sigma_0^2$. □

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE T. Automat. Contr.* **19**, 716–723.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24**, 2350–2383.

Dharmadhakiri, S.W., Fabian, V. and Jogdeo, K. (1968). Bounds on the moments of martingales. *Ann. Math. Stat.* **39**, 1719–1723.

Durrett, R. (2005). *Probability: Theory & Examples*. Thomson, Brooks/Cole, third edition.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**, 1348–1361.

Fan, J. and Lv, J. (2011). Non-concave penalized likelihood with NP-dimensionality. *IEEE T. Inform. Theory* to appear.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928–961.

Foster, D. and George, E. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947–1975.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22.

Johnson, V., Brun-Vezinet, F., Clotet, B., Gunthard, H., Kuritzkes, D., Pillay, D., Schapiro, J., and Richman, D. (2008). Update of the Drug Resistance Mutations in HIV-1: Spring 2008. *Top. HIV Med.* **16**, 62–68.

Kim, Y., Choi, H., and Oh, H. (2008). Smoothly clipped absolute deviation on high dimensions. *J. Am. Stat. Assoc.* **103**, 1665–1673.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356–1378.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498–3528.

Mallows, C. (1973). Some comments on C_p. *Technometrics* **15**, 661–675.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall/CRC, second edition.

Rhee, S., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D., and Shafer, R. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *P. Natl. Acad. Sci. USA* **103**, 17355.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

Shao, J. (1993). Linear model selection by cross-validation. *J. Am. Stat. Assoc.* **88**, 486–494.

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

Wang, H. and Leng, C. (2007). Unified lasso estimation by least squares approximation. *J. Am. Stat. Assoc.* **102**, 1039–1048.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. Roy. Stat. Soc. B.* **71**, 671–683.

Wang, H., Li, R., and Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.

Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. *J. Mach. Learn. Res.* **7**, 2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **101**, 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. B.* **67**, 301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35**, 2173–2192.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36**, 1509–1533.

Zou, H. and Zhang, H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.* **37**, 1733–1751.

Department of Statistics and Biostatistics, Rutgers University, 501 Hill Center, 110 Frelinghuysen Rd., Piscataway, NJ 08854

E-mail: ldicker@stat.rutgers.edu

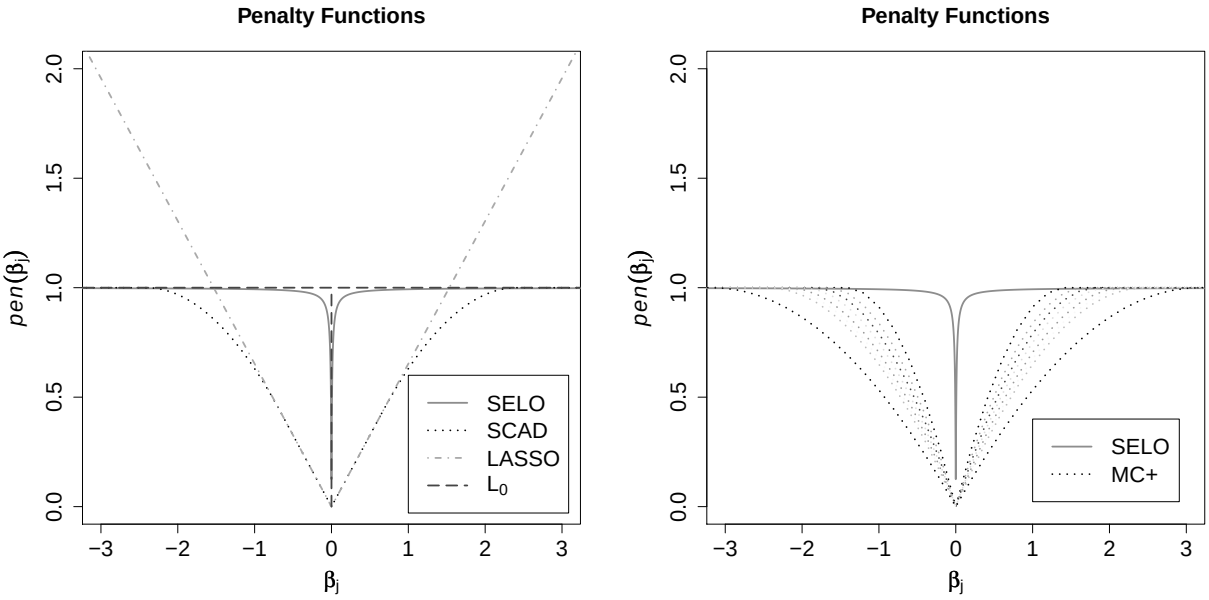
Department of Mathematics, Beijing Institute of Technology, Beijing 100081, China

E-mail: hbaosheng@bit.edu.cn

Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave., Boston, MA 02115

E-mail: xlin@hsph.harvard.edu

Figure 1: Left: SELO, $\lambda = 1, \tau = 0.01$; SCAD, $a = 3.7$, $\lambda = \sqrt{2/(a+1)}$; L_1, $\lambda = \sqrt{2/(a+1)}$; L_0, $\lambda = 1$. Right: SELO, $\lambda = 1$, $\tau = 0.01$; MC+, with $\lambda = \sqrt{2/\gamma}$ and γ taking various values between 1.01 (most concavity – closest to SELO) and 5 (least concavity).



Tuning	n	Method	Model size	Correct model	$F+$	$F-$	Model error
Data	50	LASSO	5.766	0.070	0.438	0.006	1.117
Validation		Adaptive LASSO	4.695	0.209	0.306	0.013	1.083
		SCAD	4.657	0.188	0.310	0.014	1.056
		MC+	4.687	0.201	0.312	0.015	1.090
		SELO					
		$\tau = 0.01$	3.742	0.540	0.147	0.023	1.102
		$\tau \in \{0.001, 0.01, 0.1, 0.5\}$	3.665	0.544	0.141	0.022	1.021
	100	LASSO	5.791	0.058	0.439	0.000	0.521
		Adaptive LASSO	4.472	0.307	0.264	0.001	0.430
		SCAD	4.324	0.353	0.241	0.001	0.384
		MC+	4.530	0.304	0.272	0.000	0.401
		SELO					
		$\tau = 0.01$	3.669	0.709	0.113	0.005	0.394
		$\tau \in \{0.001, 0.01, 0.1, 0.5\}$	3.720	0.670	0.126	0.002	0.352
BIC	50	LASSO	4.191	0.276	0.249	0.009	1.278
		Adaptive LASSO	3.315	0.518	0.120	0.037	1.277
		SCAD	3.798	0.362	0.200	0.023	1.220
		MC+	3.244	0.495	0.113	0.042	1.333
		SELO					
		$\tau = 0.01$	2.913	0.605	0.061	0.056	1.310
		$\tau \in \{0.001, 0.01, 0.1, 0.5\}$	2.904	0.607	0.061	0.055	1.300
λ_{univ}		MC+	3.088	0.557	0.088	0.045	1.991
BIC	100	LASSO	4.065	0.347	0.219	0.000	0.643
		Adaptive LASSO	3.265	0.750	0.066	0.004	0.475
		SCAD	3.645	0.548	0.142	0.002	0.458
		MC+	3.211	0.748	0.063	0.009	0.507
		SELO					
		$\tau = 0.01$	3.061	0.879	0.026	0.008	0.408
		$\tau \in \{0.001, 0.01, 0.1, 0.5\}$	3.052	0.881	0.024	0.008	0.394
λ_{univ}		MC+	3.208	0.763	0.059	0.007	0.774

Table 1: Simulation study I results ($\beta^* \in \mathbb{R}^8$). "Model size," "$F+$," "$F-$," and "Model error" indicate the mean model size, false positive rate, false negative rate, and model error over all 1000 independent datasets. "Correct model" indicates the proportion of times the correct model, A, was selected over the 1000 datasets.

n	Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
50	$\tau = 0.01$	0.237	0.185	0.012	0.009	0.186	0.011	0.010	0.007
		(0.345)	(0.711)	(0.102)	(0.089)	(0.399)	(0.088)	(0.075)	(0.055)
	$\tau \in \{0.001,$	0.237	0.184	0.011	0.009	0.186	0.009	0.008	0.007
	$0.01, 0.1, 0.5\}$	(0.346)	(0.714)	(0.103)	(0.089)	(0.398)	(0.079)	(0.061)	(0.055)
100	$\tau = 0.01$	0.123	0.121	0.003	0.003	0.096	0.002	0.002	0.001
		(0.146)	(0.199)	(0.030)	(0.022)	(0.116)	(0.019)	(0.015)	(0.011)
	$\tau \in \{0.001,$	0.123	0.119	0.003	0.002	0.095	0.002	0.002	0.002
	$0.01, 0.1, 0.5\}$	(0.140)	(0.190)	(0.027)	(0.016)	(0.109)	(0.015)	(0.014)	(0.013)

Table 2: Variability of SELO estimators in simulation study I ($\beta^* \in \mathbb{R}^8$); BIC tuning parameter selection. Mean estimated variance across 1000 simulated datasets and empirical variance of SELO estimates (in parentheses).

Tuning	n	Method	Model size	Correct model	$F+$	$F-$	Model error
Data	50	LASSO	8.270	0.024	0.579	0.002	1.598
Validation		Adaptive LASSO	5.711	0.104	0.413	0.010	1.559
		SCAD	6.063	0.076	0.451	0.009	1.470
		MC+	6.125	0.106	0.458	0.007	1.503
		SELO					
		$\tau = 0.01$	3.289	0.521	0.093	0.018	1.502
		$\tau \in \{0.001, 0.01, 0.1, 0.5\}$	3.452	0.539	0.111	0.014	1.306
	100	LASSO	8.474	0.020	0.583	0.000	0.766
		Adaptive LASSO	5.543	0.174	0.363	0.001	0.583
		SCAD	5.868	0.165	0.401	0.001	0.513
		MC+	6.133	0.126	0.427	0.001	0.589
		SELO					
		$\tau = 0.01$	3.544	0.765	0.079	0.003	0.488
		$\tau \in \{0.001, 0.01, 0.1, 0.5\}$	3.683	0.707	0.107	0.002	0.460
BIC	50	LASSO	5.001	0.208	0.331	0.003	1.928
		Adaptive LASSO	4.198	0.275	0.258	0.015	1.922
		SCAD	6.345	0.043	0.498	0.009	1.896
		MC+	3.675	0.348	0.191	0.015	1.954
		SELO					
		$\tau = 0.01$	3.361	0.411	0.156	0.019	1.939
		$\tau \in \{0.001, 0.01, 0.1, 0.5\}$	3.440	0.420	0.169	0.018	1.984
λ_{univ}		MC+	3.242	0.464	0.126	0.016	2.689
BIC	100	LASSO	4.408	0.328	0.255	0.000	0.963
		Adaptive LASSO	3.583	0.595	0.127	0.002	0.634
		SCAD	5.478	0.154	0.390	0.000	0.567
		MC+	3.470	0.579	0.117	0.004	0.716
		SELO					
		$\tau = 0.01$	3.310	0.702	0.080	0.002	0.571
		$\tau \in \{0.001, 0.01, 0.1, 0.5\}$	3.263	0.696	0.076	0.003	0.578
λ_{univ}		MC+	3.262	0.674	0.079	0.004	1.043

Table 3: Simulation study II results ($\beta^* \in \mathbb{R}^{20}$). "Model size," "$F+$," "$F-$," and "Model error" indicate the mean model size, false positive rate, false negative rate, and model error over all 1000 independent datasets. "Correct model" indicates the proportion of times the correct model, A, was selected over the 1000 datasets.

d	n	Method	Model size	Correct model	$F+$	$F-$	Model error
8	50	LASSO	3.940	0.361	0.992	0.052	1.360
		Adaptive LASSO	3.119	0.567	0.354	0.235	1.289
		SCAD	3.298	0.455	0.539	0.241	1.384
		MC+	3.244	0.495	0.113	0.042	1.333
		SELO					
		$\tau = 0.01$	2.913	0.605	0.061	0.056	1.310
		$\tau \in \{0.001, 0.01, 0.1, 0.5\}$	2.904	0.607	0.061	0.55	1.300
	100	LASSO	3.830	0.411	0.830	0.000	0.677
		Adaptive LASSO	3.168	0.791	0.203	0.035	0.474
		SCAD	3.240	0.702	0.298	0.058	0.488
		MC+	3.211	0.748	0.063	0.009	0.507
		SELO					
		$\tau = 0.01$	3.061	0.879	0.026	0.008	0.408
		$\tau \in \{0.001, 0.01, 0.1, 0.5\}$	3.208	0.763	0.059	0.007	0.394
20	50	LASSO	4.090	0.329	1.181	0.091	2.020
		Adaptive LASSO	3.466	0.370	0.796	0.330	1.793
		SCAD	3.830	0.251	1.161	0.331	2.059
		MC+	3.675	0.348	0.191	0.015	1.954
		SELO					
		$\tau = 0.01$	3.361	0.411	0.156	0.019	1.939
		$\tau \in \{0.001, 0.01, 0.1, 0.5\}$	3.440	0.420	0.169	0.018	1.984
	100	LASSO	4.032	0.403	1.033	0.001	1.030
		Adaptive LASSO	3.331	0.658	0.394	0.063	0.645
		SCAD	3.566	0.511	0.652	0.086	0.719
		MC+	3.470	0.579	0.117	0.004	0.716
		SELO					
		$\tau = 0.01$	3.310	0.702	0.080	0.002	0.571
		$\tau \in \{0.001, 0.01, 0.1, 0.5\}$	3.263	0.696	0.076	0.003	0.578

Table 4: Simulation study III (common BIC criterion). BIC tuning parameter selection used for all methods. "Model size," "$F+$," "$F-$," and "Model error" indicate the mean model size, false positive rate, false negative rate, and model error over all 1000 independent datasets. "Correct model" indicates the proportion of times the correct model, A, was selected over the 1000 datasets.

d	n	Method	Model size	Correct model	$F+$	$F-$	Model error
339	800	LASSO					
		Common BIC	126.14	0	15.56	0.42	15.683
		BIC	130.01	0	19.34	0.33	14.525
		Adaptive LASSO					
		Common BIC	112.95	0	6.17	4.22	13.521
		BIC	115.64	0	8.27	3.63	12.718
		SCAD					
		Common BIC	114.53	0	10.94	7.41	16.970
		BIC	137.17	0	29.26	3.09	11.856
		MC+	112.23	0	7.90	6.67	15.903
		SELO, $\tau = 0.01$	104.04	0	1.07	8.03	11.038

Table 5: Simulation study IV (large d: $\beta^* \in \mathbb{R}^{339}$). Recall that $d_0 = 111$. BIC tuning parameter selection used for all methods. For LASSO, adaptive LASSO, and SCAD, "Common BIC" indicates that the BIC criterion (6) was used for tuning parameter selection; "BIC" indicates that a previously proposed BIC criterion (discussed in Section 5.1) was used for tuning parameter selection. For SELO and MC+, the common BIC criterion (6) was used. "Model size," "$F+$," "$F-$," and "Model error" indicate the mean model size, false positive rate, false negative rate, and model error over all 100 independent datasets. "Correct model" indicates the proportion of times the correct model, A, was selected over the 100 datasets.

Method	Model size	R^2	R^2/R^2_{OLS}
OLS	76	0.772	1
LASSO	23	0.719	0.931
Adaptive LASSO	14	0.724	0.937
SCAD	24	0.739	0.957
MC+	24	0.723	0.936
SELO, $\tau = 0.01$	16	0.740	0.958

Table 6: HIV drug resistance and codon mutation analysis. BIC tuning parameter selection. "Model size" indicates the number of mutations selected by each method; R^2 is equal to one minus the residual sum of squares divided the total sum of squares.

Codon	Point estimate	Standard error	p-value
10	0.77	0.078	$< 2 \times 10^{-16}$
30	0.89	0.14	5.3×10^{-11}
32	0.94	0.17	3.5×10^{-8}
33	0.64	0.095	1.3×10^{-11}
46	0.57	0.072	3.8×10^{-15}
47	1.1	0.22	6.4×10^{-7}
48	0.55	0.15	0.00035
50	0.68	0.15	8.4×10^{-6}
54	0.65	0.082	3.8×10^{-15}
64	-0.29	0.076	0.00014
71	-0.22	0.074	0.0032
76	1.1	0.16	6.7×10^{-13}
84	1	0.086	$< 2 \times 10^{-16}$
88	-1.2	0.12	$< 2 \times 10^{-16}$
90	0.67	0.074	$< 2 \times 10^{-16}$
93	-0.23	0.062	0.00017

Table 7: Codons selected by SELO/BIC.