Spline-Lasso in High-Dimensional Linear Regression

Jianhua GUO, Jianchang HU, Bing-Yi JING, and Zhen ZHANG

We consider a high-dimensional linear regression problem, where the covariates (features) are ordered in some meaningful way, and the number of covariates $p$ can be much larger than the sample size $n$. The fused lasso of Tibshirani et al. is designed especially to tackle this type of problems; it yields sparse coefficients and selects grouped variables, and encourages local constant coefficient profile within each group. However, in some applications, the effects of different features within a group might be different and change smoothly. In this article, we propose a new spline-lasso or more generally, spline-MCP to better capture the different effects within the group. The newly proposed method is very easy to implement since it can be easily turned into a lasso or MCP problem. Simulations show that the method works very effectively both in feature selection and prediction accuracy. A real application is also given to illustrate the benefits of the method. Supplementary materials for this article are available online.

KEY WORDS: Oracle inequality; Spline-lasso; Spline-MCP; Variable selection.

1. INTRODUCTION

Suppose that we have a dataset consisting of $n$ pairs $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i^T = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$ and $y_i$ are either quantitative or $\{0, 1\}$ representing two classes like “healthy” and “sick.” Consider a linear regression model

$$y_i = x_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\beta^T = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ is the unknown parameter, and $\varepsilon_1, \ldots, \varepsilon_n$ are error variables. We are especially interested in problems where the dimension $p$ is high and the sample size $n$ might be small.

When dealing with high-dimensional data, there are usually two important considerations: model sparsity and prediction ability. The lasso (Tibshirani 1996) tries to achieve both objectives, giving an estimate of $\beta$ by

$$\hat{\beta}_{\text{lasso}} = \arg \min_{\beta} \left\{ \sum_i \left( y_i - \sum_j x_{ij} \beta_j \right)^2 + \lambda \sum_j |\beta_j| \right\}.$$

Some convergence and asymptotics results of lasso have been provided by Knight and Fu (2000). There are also many similar methods to lasso with different purposes. In particular, the elastic net of Zou and Hastie (2005) imposes an extra $L_2$-penalty on top of the lasso $L_1$-penalty:

$$\hat{\beta}_{\text{EN}} = \arg \min_{\beta} \left\{ \sum_i \left( y_i - \sum_j x_{ij} \beta_j \right)^2 + \lambda_1 \sum_j |\beta_j| + \lambda_2 \sum_j \beta_j^2 \right\},$$

where the strict convexity of $L_2$ penalty can select correlated influential features to enjoy the grouping effect. Other variants include the adaptive lasso of Zou (2006), SCAD of Fan and Li (2001), minimax concave penalty (MCP) of Zhang (2010), grouped lasso of Yuan and Lin (2006), Dantzig selector of Candès and Tao (2007), fused lasso of Tibshirani et al. (2005), and so on.

In this article, we will focus on the cases where the high-dimensional features $x_i^T = (x_{i1}, \ldots, x_{ip})$ are ordered in some meaningful way, which may lead to high correlation among closely located variables. Such examples include protein mass spectroscopy, the gene expression data in microarray studies, and single nucleotide polymorphisms (SNPs) data in genome-wide association studies (GWAS). Our goal is two-fold: first to select the relevant features, and second to predict $Y_j$ from a future $X_j^T = (X_{j1}, \ldots, X_{jp})$. To tackle this type of problems, Tibshirani et al. (2005) proposed the fused lasso, which gives an estimator of $\beta$ by

$$\hat{\beta}_{\text{fused}} = \arg \min_{\beta} \left\{ \sum_i \left( y_i - \sum_j x_{ij} \beta_j \right)^2 + \lambda_1 \sum_j |\beta_j| + \lambda_2 \sum_{j=2}^p |\beta_j - \beta_{j-1}| \right\}.$$

Or equivalently,

$$\hat{\beta}_{\text{fused}} = \arg \min_{\beta} \sum_i \left( y_i - \sum_j x_{ij} \beta_j \right)^2,$$

subject to

$$\sum_j |\beta_j| \leq s_1 \quad \text{and} \quad \sum_{j=2}^p |\beta_j - \beta_{j-1}| \leq s_2.$$

The first constraint yields sparse estimates while the second encourages flatness of the coefficients $\beta_j$. Fused lasso tries to maintain grouping effects as well as sparsity of the coefficients, and has been successfully applied in some real applications, for example, the hot spot detection for comparative genomic hybridization (CGH) data in Tibshirani and Wang (2007), and

Jianhua Guo is Professor, School of Mathematics and Statistics, Key Laboratory for Applied Statistics of MOE, Changchun 130024, China (E-mail: jhguo@nenu.edu.cn). Jianchang Hu is PhD Student, (E-mail: hjianchang@gmail.com), and Bing-Yi Jing is Professor, (E-mail: majing@ust.hk), Hong Kong University of Science and Technology, Hong Kong, China. Zhen Zhang is Assistant Professor, Department of Mathematics, National University of Singapore, Singapore (E-mail: matzz@nus.edu.sg). We wish to thank the Editor, an Associate Editor, and two reviewers for their helpful comments which led to considerable improvements in the presentation of the article.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jasa.
Despite these successes, however, we would like to make two observations, which we believe that improvements can be made upon.

1. First of all, the fused lasso encourages local constancy of the coefficient profile within each group; see, for example, Figure 3 of Tibshirani et al. (2005). However, in some applications, the features in a group might vary smoothly, rather than being like a step function, for example, the SNP data in Section 4 later. In those situations, the fused lasso may not work as well as hoped, see Figure 1 later.

2. Second, the optimization problem for the fused lasso is a quadratic programming problem; for large $p$, as pointed out in Tibshirani and Wang (2007, sec. 2.4), the problem is difficult to solve and special care must be taken to avoid the use of $p^3$ storage elements. For large $p$, more efficient algorithms are needed. One of the promising algorithms is the coordinate descent algorithm, proposed by Friedman et al. (2007). The coordinate decent has been shown to work for a variety of penalized (convex and nonconvex alike) regressions, for example, elastic net, grouped lasso, bridge lasso, SCAD, and MCP, etc. The reason why it works is that it reduces the high-dimensional optimization problem to one-dimensional one, for which we can easily derive explicit solution at every step. It is proven to be faster than LARS for the lasso, particularly for large $p$. However, as shown in Tibshirani and Taylor (2011), the coordinate descent does not work in the fused lasso, partially due to the nondifferentiability of $L_1$-norm, but a modified version works.

In this article, we propose a new approach, referred to as “spline-lasso” or “spline-MCP” to overcome the above two issues. Borrowing the idea from the cubic spline smoothing literature, we impose an $L_2$ penalty on the discrete version of the second derivatives of $\beta$’s, as well as the lasso penalty on $\beta$’s. We will see later that the proposed method can capture smoothing changes in $\beta$’s within a group. Furthermore, the proposed method is very easy to implement computationally since it can be easily turned into usual lasso or MCP problem, thus using many available efficient algorithms such as LARS of Efron et al. (2004) and PLUS of Zhang (2010) when $p$ is not very large. If the dimension is very large, the coordinate decent algorithm could also be adopted and since for each step, there are closed-form solutions for the optimization, thus the algorithm is easy to implement and converges quickly.

It is worth mentioning some related work here. Hebiri and Van de Geer (2011) introduced the “smooth-lasso,” where the $L_1$-penalty $\sum_{j=1}^{p} |\beta_j - \beta_{j-1}|$ in the fused lasso is replaced by the $L_2$-penalty $\sum_{j=1}^{p} |\beta_j - \beta_{j-1}|^2$. The method is good at capturing the smooth features in a group, but it does not work as well as our proposed method under similar settings, see the simulation section later in the article. Huang et al. (2011) proposed the sparse Laplacian shrinkage (SLS) method for variable selection and estimation that explicitly incorporates the correlation patterns among predictors. The article is different from ours since it does not treat the case of ordered covariates (or features) as we do in this article.

The layout of the article is as follows. In Section 2, we introduce the spline-lasso and spline-MCP, and given some oracle inequalities. Simulation studies are given in Section 3. Some real application is given in Section 4. In Section 5, we give some discussions. Proofs will be postponed in Section 6.

2. METHODOLOGY AND MAIN RESULTS

We will first introduce the spline-lasso method, and prove some oracle inequalities. We will then introduce the spline-MCP, a more general version of the spline-lasso, and will also prove some oracle inequalities.

2.1 The Spline-Lasso

The spline-lasso procedure is defined as

$$
\hat{\beta}_{\text{spline}} = \arg \min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} (\Delta_j^2 \beta)^2 \right\},
$$

(1)
where
\[
\Delta_j \beta = : (\beta_{j+1} - \beta_j), \quad \text{and} \quad \Delta_j^{(2)} \beta =: \Delta_j \beta - \Delta_{j-1} \beta = \beta_{j+1} - 2\beta_j + \beta_{j-1},
\]
are the first- and second-order difference (or discrete versions of derivatives) of the coefficients \( \beta_j \). The first \( \lambda_1 \)-penalty in (1) is the usual lasso \( L_1 \)-penalty on \( \beta_j \)'s to encourage sparse coefficients. The second \( \lambda_2 \)-penalty in (1) mimics the cubic spline, that is penalizing the \( L_2 \) norm of discrete version of the second-order derivatives of the coefficients \( \beta_j \) to encourage smoothness of coefficients.

The spline-lasso is designed to overcome the two difficulties encountered by the fused lasso mentioned in the introduction. First, it tries to capture the smooth varying features by providing a smoother estimation than the fused lasso. Second, it is computationally easy since the \( L_2 \) penalty term can be easily merged with the least squares terms to transform the problem into another lasso problem (see the next proposition), so that we can make full use of the LARS algorithms proposed by Efron et al. (2004).

**Proposition 1.** Given dataset \((Y, X)\) and \((\lambda_1, \lambda_2)\), define an artificial dataset \((Y^*, X^*)\) by
\[
X_{(n+p-2)\times p}^* = (X, \sqrt{n}\lambda_2 L)^T, \quad Y_{(n+p-2)}^* = (Y, 0)^T,
\]
where \( L \) is a \((p - 2) \times p \) matrix with \( L_{i,i} = L_{i,i+1} = 1, L_{i,i+1} = -2 \) and \( L_{i,j} = 0 \) otherwise. Then
\[
\hat{\beta}^{\text{spline}} = \arg \min_{\beta} \left\{ \frac{1}{n} J^{*2} (Y^* - X^* \beta)^T (Y^* - X^* \beta) + \lambda \sum_{j=1}^p |\beta_j| \right\}.
\]

The next result is concerned with the oracle inequality for proposed spline-lasso \( \hat{\beta}^{\text{spline}} \) which gives \( L_1 \)- and \( L_2 \)-error bounds on the risk, where
\[
\|x\|_p = \left( \sum_{i=1}^n |x_i|^p \right)^{1/p}.
\]
We use \(|A|\) to denote the cardinality of the set \( A \). We first give the so-called restricted eigenvalue assumption, see Bickel, Ritov, and Tsybakov (2009). We will discuss more about on Assumption 1 after Theorem 1.

**Assumption 1.** Let \( \bar{L} = L^T X \) where \( L \) is given in Proposition 1, and \( K_n = X^T X + n\lambda_2 I \) and let \( \rho_n = 4\sqrt{|A|} + 4\lambda_2 \lambda_1^{-1} \|\bar{L}\|_2 \) where \( A \) is the nonsparse set \( A = \{ j : \beta_j \neq 0 \} \), and \( \lambda_1 \) and \( \lambda_2 \) are tuning parameters for the spline-lasso. There is a constant \( \phi_{\beta} > 0 \) such that, for any \( \Delta \in \mathbb{R}^p \) satisfying
\[
\sum_{j \notin \Theta} |\Delta_j| \leq \rho_n \sqrt{\sum_{j \in \Theta} \Delta_j^2}, \quad \Theta \subset \{1, 2, \ldots, p\},
\]
there is an \( \alpha_0 \) such that, for any \( \Delta \in \mathbb{R}^p \) satisfying
\[
\frac{1}{n} \Delta^T K_n \Delta \geq \phi_{\beta} \sum_{j \in \Theta} \Delta_j^2.
\]
We now present the oracle inequalities for the spline-lasso.

**Theorem 1.** Suppose that Assumption 1 holds with a set \( B \supset A \) such that \(|B| \leq c_f |A|\) for a given constant \( c_f \geq 1 \). Assume \( \varepsilon_i \)'s are normally distributed. Let \( \lambda_1 = 4\sqrt{2}\sigma \sqrt{n^{-1}} \log(p/\eta) \) where \( \eta \in (0, 1) \). Then, with probability greater than \( 1 - \eta \), we have
\[
\frac{1}{n} \|X \beta - X \hat{\beta}^{\text{spline}}\|_2^2 \leq \phi_{\beta}^{-1} \left( 2\lambda_1 \sqrt{|A|} + 2\lambda_2 \|\bar{L}\|_2 \right)^2,
\]
and
\[
\|\beta - \hat{\beta}^{\text{spline}}\|_1 \leq 2\phi_{\beta}^{-1}\lambda_1^{-1} \left( 2\lambda_1 \sqrt{|A|} + 2\lambda_2 \|\bar{L}\|_2 \right)^2.
\]
In addition, when \( \lambda_2 \neq 0 \),
\[
\|L(\beta - \hat{\beta}^{\text{spline}})\|_2 \leq \lambda_2^{-1/2}\phi_{\beta}^{-1/2} \left( 2\lambda_1 \sqrt{|A|} + 2\lambda_2 \|\bar{L}\|_2 \right).
\]
Assumption 1 is widely used in the literature and basically requires that the restriction of the generalized Gram matrix \( K_n \) to the columns in \( \Theta \) is invertible; when \( X^T X \) is invertible, the condition (2) always holds with \( \phi_{\beta} \) at least as large as the smallest eigenvalue of \( X^T X \). See Bickel, Ritov, and Tsybakov (2009), van de Geer and Buhlmann (2009), and Hebiri and Van de Geer (2011) for more details on this assumption.

The two parameters \((\lambda_1, \lambda_2)\) have different roles to play: \( \lambda_1 \) controls the sparsity of coefficients \( \beta_j \)'s and is preset in Theorem 1, while \( \lambda_2 \) controls the smoothness of the coefficients \( \beta_j \)'s as can be seen from the last inequality in Theorem 1. As long as the \( \lambda_2 \) is chosen such that \( \lambda_1 \sqrt{|A|} \) dominates the \( \lambda_2 \|\bar{L}\|_2 \), an increase in the level of \( \lambda_2 \) will lead to an estimation whose smoothness is closer to the true \( \beta \).

Note that the optimality properties are with respect to the same smoothness conditions. If there is no smoothness requirement, \( \lambda_2 \) can be set as zero, which would reduce to the case of lasso.

The oracle inequalities in Theorem 1 depend intricately on many factors working together. The results in Theorem 1 are not asymptotic in nature, in that they hold for all \( n \) and \( p \) under certain conditions.

**The Spline-Lasso as Bayes Estimates**

It is well-known that the lasso can be regarded as a Bayes estimator with double exponential priors on \( \beta_j \)'s. We now explain that the spline-lasso has a similar Bayesian interpretation. To illustrate idea, let us first look at a related but slightly simpler problem:
\[
\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^n \left( y_i - \beta_i^T x_i \right)^2 + \lambda \sum_{j=1}^p \left( \beta_{j+1} - \beta_j \right)^2 \right\}.
\]
It is easy to verify that this amounts to assuming a priori that \( \beta_1, \ldots, \beta_p \) has an Markovian structure with \( \beta_0 = 0 \) and the transition probability is normal:
\[
f(\beta_{j+1} | \beta_j) \propto e^{-\lambda(\beta_{j+1} - \beta_j)^2}, \quad \text{for } j = 0, 1, \ldots, p - 1.
\]
Similarly, for the spline-lasso (1), the penalty on the second-order differences \( \Delta_j^{(2)} \beta = \Delta_j \beta - \Delta_{j-1} \beta \) can be regarded as imposing an Markovian structure on \( \Delta_j \beta \) and assuming a priori the transition probability on \( \Delta_{j+1} \beta \) given \( \Delta_j \beta \) is normal.

2.2 The Spline-MCP

Zhang (2010) proposed to use MCP and showed that it has a number of advantages over the lasso. This motivates us to
extend the spline-lasso to the spline-MCP, which we hope will inherit some of the nice properties of MCP. Furthermore, the spline-MCP is computationally easy to implement since it can be readily transformed to MCP, where we can use the PLUS algorithm of Zhang (2010). Oracle inequalities will also be established. Later we will see that the spline-MCP gives spectacular performance in simulations and real application.

First we define the generalized spline-like estimator as the solution of the following optimization problem:

$$\hat{\beta}(\lambda, \gamma) = \arg\min_{b} M(b; \lambda, \gamma),$$

where

$$M(b; \lambda, \gamma) = \frac{1}{2n} \|Y - Xb\|^2 + \sum_{j=1}^{p} \rho(b_j; \lambda_1, \gamma) + \frac{1}{2} \lambda_2 b^T J b,$$

and $J$ is the same as defined in Assumption 1. This $M(b; \lambda, \gamma)$ is very similar to the objective function in Huang et al. (2011), but the intention behind their problem is quite different from ours.

The penalty function $\rho(|b_j; \lambda, \gamma)$ could be SCAD, MCP, and more general quadratic spline penalty functions. To solve for the solution, again as mentioned in Proposition 1, the original optimization problem could be transformed into an equivalent penalized least square problem with augmented dataset $(Y^*, X^*)$, which can then be solved by algorithm like PLUS in Zhang (2010).

Next, we consider the properties of the general spline-like estimators with concave penalty. We use MCP as an example and for other quadratic spline penalties, the results would be similar. The MCP is defined as

$$\rho(t; \lambda_1, \gamma) = \lambda_1 \int_{0}^{\|t\|}(1 - x/(\gamma \lambda_1))_+ dx.$$ 

Its first-order derivative is

$$\dot{\rho}(t; \lambda_1, \gamma) = \lambda_1 (1 - |t|/(\gamma \lambda_1))_+ \text{sgn}(t).$$

Therefore, larger values of $\gamma$ make $\rho$ less concave.

Denote the true regression coefficient by $\beta^o = (\beta^o_1, \ldots, \beta^o_p)^T$. Let $O = \{j : \beta^o_j \neq 0\}$ be the set of indices of nonzero coefficients. Let $d^o = |O|$ be the cardinality of $O$. Define

$$\hat{\beta}^o(\lambda) = \arg\min_{b} \left\{ \frac{1}{2n} \|Y - Xb\|^2 + \frac{1}{2} \lambda_2 b^T J b, \right. \left. b_j = 0, j \notin O \right\}.$$ 

This is the oracle spline OLS estimator on the set $O$.

Let $\Sigma = n^{-1} X^T X$. Let $c_{\min}(\lambda_2)$ be the smallest eigenvalue of $\Sigma(\lambda_2) = \Sigma + \lambda_2 J$.

We consider two separate cases: $\Sigma(\lambda_2)$ is positive definite, or singular. In both cases, we will show the sign consistency under different conditions.

**Case I: $\Sigma(\lambda_2)$ is Positive Definite**

In this case, the oracle spline OLS estimator $\hat{\beta}^o$ can be explicitly written as

$$\hat{\beta}^o_0 = (\Sigma_0 + \lambda_2 I_0)^{-1} X_0^T Y / n, \quad \hat{\beta}^o_\gamma = 0.$$ 

Its expectation $\beta^* = E \hat{\beta}^o$ satisfy $\beta^*_0 = 0$ and

$$\beta^*_\gamma = (\Sigma_0 + \lambda_2 I_0)^{-1} X_0^T X_0 \beta^o_\gamma / n = (\Sigma_0 + \lambda_2 I_0)^{-1} X_0^T X_0 \beta^o_\gamma / n = (\Sigma_0 + \lambda_2 I_0)^{-1} \Sigma_0 \beta^o_\gamma.$$ 

Define $\beta_* = \min(|\beta^*_j|, j \in O)$. In addition, we give two conditions.

**Condition A.** For a certain constant $\epsilon \in (0, 1/3)$,

$$\sup_{|u|_{\infty} = 1} P(u^T \epsilon > \alpha t) \leq e^{-t^2/2}, \quad \text{for } 0 < t \leq \sqrt{2 \log(p/\epsilon)}.$$ 

**Condition B.**

1. $c_{\min}(\lambda_2) > 1/\gamma$.

2. The penalties satisfy $\lambda_1 \geq \lambda_2 C_2 + \sigma \sqrt{2 \log((p - d^o)/\epsilon)}$.

3. $\Sigma(\lambda_2) \geq \lambda_2 \Sigma_0, \Sigma_0 \leq \Sigma(\lambda_2)$.

Condition A puts the sub-Gaussian constraint on the data. Condition B1 ensures that the overall optimization will be convex. Condition B2 puts a minimum level to the sparsity penalty, that is, the threshold should be large enough to kill off all small estimates around zero. Condition B3 ensures that the true nonzero $\beta^*$ is large enough to avoid being killed by the threshold level.

**Theorem 2.** Suppose Conditions A and B hold. Then

$$P(\{j : \hat{\beta}_j \neq 0\} \neq O \text{ or } \hat{\beta} \neq \hat{\beta}^o) \leq 3 \epsilon.$$ 

If $\lambda_2 \geq \lambda_2 C_1 + \max_j \sqrt{(2 n_j)/(\log((p - d^o)/\epsilon))}$ instead of Condition B(3), then

$$P(\text{sgn}(\hat{\beta}) \neq \text{sgn}(\beta^o) \text{ or } \hat{\beta} \neq \hat{\beta}^o) \leq 3 \epsilon.$$ 

Theorem 2 shows that under the assumed conditions, our proposed spline-MCP would enjoy sign consistency and it will be the same as the oracle estimator $\hat{\beta}^o$ with high probability.

**Case II: $\Sigma(\lambda_2)$ is Singular**

For high-dimensional data, most likely $\Sigma(\lambda_2) = \Sigma + \lambda_2 J$ is singular, then the above theorem is not applicable. To give the oracle property for the estimator, the sparse Reisz condition as in Zhang (2010) is needed.

Let $\tilde{X}$ be a matrix satisfying $\tilde{X}^T \tilde{X}/n = \Sigma(\lambda_2) = X^T X / n + \lambda_2 \tilde{J}$ and $\tilde{Y}$ be a vector satisfying $\tilde{X}^T \tilde{Y} = X^T Y$. Define

$$\tilde{M}(b; \lambda, \gamma) = \frac{1}{2n} \|\tilde{Y} - \tilde{X}b\|^2 + \sum_{j=1}^{p} \rho(b_j; \lambda_1, \gamma).$$ 

Then the minimizer of $\tilde{M}$ is the same as that of original optimization problem $M$.

We list another condition used in the next theorem.

**Condition C.**

1. For an integer $d^*$ and spectrum bounds $0 < c_*(\lambda_2) \leq c^*(\lambda_2) < \infty$, $\forall B$ with $|B \cup \bar{O}| \leq d^*$ and $\|u_B\|_2 = 1$,

$$0 < c_*(\lambda_2) \leq \frac{\tilde{u}_B^T \Sigma_0 \beta(\lambda_2) \tilde{u}_B}{\sqrt{d^*}} \leq c^*(\lambda_2) < \infty,$$ 

where $b = \{j : |\beta_j| \neq 0\}$. We will show that with the conditions of Condition C, we can apply the above theorem to $\tilde{M}$.
with \( d^* \geq d^0(K_s + 1) \), \( \gamma \geq c_4^{-1}(\lambda_2)\sqrt{3} + c_6(\lambda_2)/c_4(\lambda_2) \), and \( K_s = c_6(\lambda_2)/c_4(\lambda_2) - 1/2 \).

2. With \( C_2 = \| \sum B_O(\lambda_2)\sum^{-1}(\lambda_2)\sum C_O^{-1}(\lambda_2) - \sum B_C(\lambda_2,\lambda_2)\beta^0 \|_{\infty} \),

\[
\max \left\{ 1, \sqrt{c_2(\lambda_2)}\sqrt{\lambda_1/(\lambda_1 + 1)} \right\} \lambda_1 \geq \lambda_2 C_2 + \sigma_2 \sqrt{\log(p/e)} \max_j \| x_j \|_2/n.
\]

3. With \( \{ v_{ij}, j \in O \} \) being the diagonal elements of \( \sum^{-1}(\lambda_2)\sum C_O^{-1}(\lambda_2) \),

\[
\min_{j \in O} \left( |\beta_j^*| - \gamma(2\sqrt{c_2(\lambda_2)}\lambda_{ij}) (n/v_{ij})^{1/2} \right) \geq \sigma \sqrt{2 \log (d^*/\epsilon)}.
\]

Condition C1 is the sparse Reisz condition in our situation, which requires that there are submodels containing the true model that are nonsingular. Condition C2 places a minimal penalty level so that all small estimates are set to zero. Condition C3 puts a minimum level to the true \( \beta^* \). Therefore, the true nonzero \( \beta^* \) is large enough to avoid being killed in this case.

**Theorem 3.** Suppose Conditions A and C hold. Then

\[
P(\{ j : \hat{\beta}_j \neq 0 \} \neq 0 \text{ or } \hat{\beta} \neq \beta^0) \leq 3 \epsilon.
\]

If \( \beta_s \geq \lambda_2 C_1 + \gamma(2\sqrt{c_2(\lambda_2)}\lambda_{ij}) + \max_j \sqrt{2v_{ij}/n} \log (d^*/\epsilon) \) instead of Condition C(3), then

\[
P(\text{sgn}(\hat{\beta}) \neq \text{sgn}(\beta^0) \text{ or } \hat{\beta} \neq \beta^0) \leq 3 \epsilon.
\]

Again, Theorem 3 shows the sign consistency of our proposed spline-MCP estimator in this situation and it will also equal to the oracle estimator \( \beta^0 \) with high probability.

Finally, it is worth mentioning that a nonzero \( \lambda_2 \) will induce smoothness in coefficient estimates. The larger \( \lambda_2 \) is, the smoother the estimates. What is the right level of penalty level for \( \lambda_2 \)? In fact, the optimal level of \( \lambda_2 \) is related to theoretical work in nonparametric estimation literature for example, Tsybakov (2009), where the minimax rate could be obtained with assumptions on the smoothness of \( \beta \). On the other hand, in implementing the method, since we do not know the smoothness for the \( \beta \), so data-driven method like cross-validation will be adopted.

### 2.3 Second-Stage Thresholding After Spline-Lasso/MCP

It will be seen from the simulation studies later that the spline-lasso/MCP can indeed capture smoothly changing effects and have very good prediction accuracy. However, it may not always necessarily maintain sparsity as many small false features may have been left there. Incidentally, the same phenomenon is also shared by the smooth-lasso method. The reason for this is as follows. The tuning parameters are determined by cross-validation which was prediction-oriented, so more effort has been spent on capturing the smoothness of coefficients, resulting in small tuning parameters.

To fix this, we propose a second-stage “thresholding” procedure to the spline-lasso/MCP, whenever necessary, after the first-stage estimation to further improve variable selection ability, which proves very effective in practice.

The question is then how to choose the thresholding level \( T \). One possibility is to use cross-validation, however, this may be very computationally demanding. We now suggest a simple procedure. The idea is inspired by Donoho and Johnstone (1998), where they proposed to apply \( \pm \hat{\sigma} \sqrt{2 \log p} \) as threshold to wavelet coefficients with noise level \( \hat{\sigma} \) being estimated from the coefficients in the finest level. Similarly, we propose to adopt

\[
T = \pm \hat{\sigma} \sqrt{2 \log p}
\]

as the thresholding level to estimated coefficients, where \( \hat{\sigma} \) is the standard error of coefficients with small magnitude. One approach to determine the coefficients used to form \( \hat{\sigma} \) is by sorting the absolute values of the estimated \( \beta^\text{spline} \). After ordering the absolute values in an ascending order, the smaller half can be used to calculate the \( \hat{\sigma} \). This specific threshold relies on Gaussian-type errors originally. However, we also conducted simulation studies later involving other error distributions than Gaussian errors, and found that the above thresholding levels work out nicely as well. Finally, we remark that this level of thresholding may not be necessarily needed if the solution is already very sparse. One such example is the spline-MCP. In the later simulations, we found spline-MCP gives sparse solution, so there is no need to have an extra level of thresholding.

### 3. SIMULATION STUDIES

We now conduct simulation studies to compare performances of the spline-lasso and spline-MCP with other alternatives, for example, the fused lasso of Tibshirani et al. (2005) and the smooth lasso of Hebiri and Van de Geer (2011). The features \( X_1, \ldots, X_p \) are generated randomly from multivariate normal distribution \( N(0, \Sigma) \), and the response variable \( y \) is generated from

\[
y = X\beta + \epsilon.
\]

We fix \( n = 71 \) and \( p = 600 \) so that \( p \gg n \). The coefficients \( \beta \)'s are obtained from a continuous function consisting of two parts: one part being a combination of sine and constant function, and the other being a combination of two linear functions. Then it is evaluated at discrete points which gives the \( \beta \) as shown in Figures 1 and 2 by blue circles. There are \( k = 79 \) non-zero \( \beta \)'s with the rest being zero. All tuning parameters are determined by 10-fold cross-validation. We refer to Hastie, Tibshirani, and Friedman (2009, sec. 7.1) for a detailed description of the \( K \)-fold cross-validation. The cross-validation is conducted over a grid of parameters \( (\lambda_1, \lambda_2) \), and the pair with the smallest cross-validated prediction error is chosen.

Two different covariance structures \( \Sigma \) are considered:

- **Case 1:** \( \Sigma = I \) (no neighboring correlation).
- **Case 2:** \( \Sigma_{ij} = 0.5^{i-j} \) (relatively high neighboring correlation).

We will first show simulation results when the errors are normally distributed, that is, \( \epsilon \sim N(0, \sigma^2) \). Here, for each case, two different noise levels are considered. Specifically, we take \( \sigma = 1 \) as a low noise level and take \( \sigma = 3 \) as a high noise level.

**Performance Criterion**

The performance of different methods will be examined in two aspects:

1. **Variable Selection Accuracy:**
   - \( \hat{\beta} \neq 0 \) vs. \( \beta = 0 \)
   - \( \text{Percentage of True Positives (PTE)} \)
   - \( \text{Percentage of True Negatives (PNE)} \)

2. **Prediction Accuracy:**
   - \( \| y - \hat{y} \|_2 \) vs. \( \| y - \hat{y} \|_2 \)
   - **Mean Squared Error (MSE)**
   - **Root Mean Squared Error (RMSE)**

3. **Computational Efficiency:**
   - \( \text{Computation Time} \)
   - \( \text{Memory usage} \)
• **Prediction error.** We generate 1000 testing data for each method and use the fitted models from the training set to give predictions to the testing set. The mean square error (MSE) of the predicted value for the testing dataset then can be calculated as the $l_2$ norm of the difference between predicted values and the true simulated values.

• **Feature selection.** We compare sensitivity and specificity for different methods. Sensitivity (or true positive rate) measures the proportion of actual positives being correctly identified. Specificity (or true negative rate) measures the proportion of negatives being correctly identified.

**Simulation results.** For illustration, in Figures 1 and 2 we present the estimation results for the correlated features in Case II, by four different methods: fused-lasso, spline-MCP, spline-lasso, and spline-lasso with thresholding, the last of which was described in Section 2.3. Clearly, the spline-MCP gives the best estimation, followed by spline-lasso with thresholding. The other two did not clean out the noisy signals very well. The improvements of the spline-MCP and spline-lasso in capturing the smooth changing effects are clearly seen in Figures 1 and 2.

From Tables 1–4, we make some further observations.

1. Columns 2–5 of Tables 1–4 give results without thresholding. The spline-MCP and spline-lasso outperform the fused lasso in terms of estimation and prediction. This may not be surprising since both are designed to capture the smooth changes of features.

2. The last three columns of Tables 1–4 give results with thresholding. Clearly, thresholding does help improve performances for both the smooth and spline-lassos. The spline-lasso with thresholding gives performance very close to that of spline-MCP. However, the improvement is not significant for fused lasso. This is not surprising. Since the penalties in forming the fused lasso are both $l_1$ penalty, which makes the estimation sparser than the smooth and spline-lassos. Therefore, when forming the threshold, its magnitude is very small, which makes the thresholding not very effective. This observation also indicates that fused lasso can be another example where the thresholding may not be necessarily applied.

3. Comparing Cases 1–2 in Tables 1 and 3, the estimations are more accurate when there was correlation between features. This is because the second penalty was trying to use these information to help make estimation, especially for spline-MCP, spline-lasso, and smooth-lasso.

4. From Tables 2 and 4, without thresholding, the smooth-lasso and spline-lasso have very high sensitivity, but rather low specificity. When the thresholding is applied (the last two columns), specificities have been substantially improved with very little sacrifice on sensitivities.

More Simulation Results

We then consider the cases where noise term in the regression follows $t$-distribution with degree of freedom 5. We also adjust their variances to have $\sigma = 3$. For the second-stage thresholding, we still use the same formula to determine the threshold levels. The performance criterion are also kept the same. The following tables gives an summary to the simulation results.

From the above two tables, we can still see clearly that spline-MCP and spline-lasso after thresholding outperform other methods. Fused lasso seems to be better than spline-lasso without thresholding. This is actually reasonable since fused lasso provides sparser estimations than spline-lasso. When the noises follow $t_5$, the estimations for those zero $\beta$’s will be larger than those in the normal case. These false positives may badly affect the performance of the spline-lasso. However, after thresholding, we see that the performance of spline-lasso is very close to and sometimes even better than spline-MCP. This means that the estimation for those true signals is actually fine for the spline-lasso and the reason for spline-lasso to perform undesirable is largely due to those noisy terms, which again highlights the necessity of the second-stage thresholding.
Table 1. Estimation and prediction errors with low noise level. The testing MSE is the average MSE of predictions from fitted models based on 1000 testing data for this case

\[
\begin{array}{cccccccc}
\|\hat{\beta} - \beta\|^2 & Fused & Spline-MCP & Spline-Lasso & Smooth & Fused/Threshold. & Spline/Threshold. & Smooth/Threshold. \\
Case 1 & 2.94 & \textbf{1.27} & 1.81 & 3.49 & 2.93 & 1.65 & 3.12 \\
Case 2 & 2.22 & 1.09 & 0.88 & 1.49 & 2.22 & \textbf{0.79} & 1.28 \\
\end{array}
\]

Testing MSE

\[
\begin{array}{cccccccc}
Case 1 & 9.97 & \textbf{2.58} & 4.33 & 13.02 & 9.90 & 3.75 & 10.77 \\
Case 2 & 7.31 & 3.09 & 2.22 & 5.04 & 7.31 & \textbf{1.91} & 3.63 \\
\end{array}
\]

Note: bold formatting indicates the smallest estimation and prediction errors.

Table 2. Specificity and sensitivity when noise level is low

\[
\begin{array}{cccccccc}
Specificity & Fused & Spline-MCP & Spline-Lasso & Smooth & Fused/Threshold. & Spline/Threshold. & Smooth/Threshold. \\
Case 1 & 0.9309 & 0.9942 & 0.1574 & 0.1305 & 0.9923 & 0.9827 & 0.9827 \\
Case 2 & 0.9866 & 0.9923 & 0.4683 & 0.4088 & 0.9866 & 0.9789 & 0.9846 \\
\end{array}
\]

Sensitivity

\[
\begin{array}{cccccccc}
Case 1 & 0.9114 & 0.962 & 1.00 & 1.00 & 0.9114 & 0.962 & 0.962 \\
Case 2 & 0.9747 & 0.9873 & 1.00 & 1.00 & 0.9747 & 0.9873 & 0.9873 \\
\end{array}
\]

Table 3. Estimation and prediction errors when noise level is high. The testing MSE is the average MSE of predictions from fitted models based on 1000 testing data for this case

\[
\begin{array}{cccccccc}
\|\hat{\beta} - \beta\|^2 & Fused & Spline-MCP & Spline-Lasso & Smooth & Fused/Threshold. & Spline/Threshold. & Smooth/Threshold. \\
Case 1 & 3.95 & 3.66 & 3.21 & 5.48 & 3.39 & \textbf{1.87} & 4.05 \\
Case 2 & 2.40 & \textbf{1.40} & 2.08 & 3.51 & 2.39 & 1.55 & 2.66 \\
\end{array}
\]

Testing MSE

\[
\begin{array}{cccccccc}
Case 1 & 23.79 & 21.71 & 18.20 & 38.65 & 19.58 & 13.21 & 26.30 \\
Case 2 & 19.97 & \textbf{14.25} & 20.70 & 39.99 & 19.67 & 14.74 & 26.11 \\
\end{array}
\]

Note: bold formatting indicates the smallest estimation and prediction errors.

Table 4. Specificity and sensitivity when noise level is high

\[
\begin{array}{cccccccc}
Specificity & Fused & Spline-MCP & Spline-Lasso & Smooth & Fused/Threshold. & Spline/Threshold. & Smooth/Threshold. \\
Case 1 & 0.5912 & 0.9079 & 0.0096 & 0.0192 & 0.9750 & 0.9981 & 0.9942 \\
Case 2 & 0.9309 & 0.9981 & 0.1881 & 0.1382 & 0.9904 & 0.9846 & 0.9904 \\
\end{array}
\]

Sensitivity

\[
\begin{array}{cccccccc}
Case 1 & 1.00 & 0.9114 & 1.00 & 1.00 & 0.9494 & 0.9241 & 0.9494 \\
Case 2 & 0.9747 & 0.962 & 1.00 & 1.00 & 0.9747 & 0.9747 & 0.9367 \\
\end{array}
\]

Table 5. Estimation and prediction errors when noises follow \(t_5\). The testing MSE is the average MSE of predictions from fitted models based on 1000 testing data for this case

\[
\begin{array}{cccccccc}
\|\hat{\beta} - \beta\|^2 & Fused & Spline-MCP & Spline-Lasso & Smooth & Fused/Threshold. & Spline/Threshold. & Smooth/Threshold. \\
Case 1 & 3.74 & 3.59 & 3.11 & 4.58 & 3.39 & \textbf{2.23} & 3.66 \\
Case 2 & 2.34 & \textbf{1.27} & 2.34 & 3.46 & 2.22 & 1.58 & 2.57 \\
\end{array}
\]

Testing MSE

\[
\begin{array}{cccccccc}
Case 1 & 22.28 & 20.55 & 18.19 & 28.71 & 19.58 & \textbf{13.21} & 21.45 \\
Case 2 & 16.87 & \textbf{11.61} & 21.64 & 37.77 & 15.31 & 13.87 & 23.00 \\
\end{array}
\]

Note: bold formatting indicates the smallest estimation and prediction errors.
4. EMPIRICAL ANALYSIS

We apply our method to a dataset of Crohn’s Disease (CD) from WTCCC (Consortium 2007). The combined sample size is 5009 in the case-control study. We conduct analysis on Chromosome 16, which consists of 10827 SNPs after standard quality control strategy. The purpose is to identify disease-associated SNP clusters in the GWAS.

Although GWAS has identified many disease-susceptibility SNPs, these findings can only explain a small portion of genetic contributions to complex diseases (missing heritability). Since a causal SNP will have a small chance of being directly genotyped, the effects of its neighboring SNPs may be too weak to be detected due to the effect decay caused by imperfect linkage disequilibrium. Furthermore, it is still challenging to detect a causal SNP with a small effect even if it has been genotyped. To increase the statistical power when detecting disease-associated SNPs with relatively small effects, Yang et al. (2011) applied a weighted version of fused lasso and found that there are some SNPs with relatively small effects, Yang et al. (2011) applied a method are chosen in the same way as in Yang et al. (2011). That is, we fix $\lambda_1 = \sqrt{2\log(\pi/2.7)}$ to adjust the number of effective SNPs, and consider a wide range of decreasing $\lambda_2$. We solve for spline-MCP solution for each $\lambda_2$ and consider the variance of the residuals. Finally, we choose the largest $\lambda_2$ such that the variance of the residuals is $\leq 1$. In the analysis below, we control the estimated FDR at 5% level.

Adopting this parameter tuning procedure may produce some false positive results. We can consider the stability selection strategy (Meinshausen and Buhlmann 2010) to reduce the effect of parameter tuning. This stability selection procedure depends on the probability that certain signal is detected, also called the probability of detection, under model perturbation via subsampling. More specifically, we randomly sample half of the cases and controls for each subsampling round. Let $p_{b}$ denote the set of $p$ values for the $b$th subsampling. Then we apply our method on $-\log p_{b} - 1$ and obtain the support $s_{b}$. If $s_{b} = 1$, it indicates that $i$th SNP is detected in the $b$th subsampling. The probability of detecting the $i$th SNP is estimated by

$$\pi_{i} = \frac{\sum_{b=1}^{B} s_{b,i}}{B},$$

where $B$ is the total number of subsamplings. Here, we set $B = 100$. Then we can get a set of SNPs as $A_{\tau} = \{i : \pi_{i} \geq \tau\}$ with $\tau$ being the threshold. This threshold will be selected to control the estimated FDR at 5% level.

We found a total of 31 SNPs, among which 30 of them are close to each other and their estimation by spline-MCP is shown in Figure 3 where the $x$-axis gives the location and $y$-axis is the fitted value. From Figure 3, we see a smooth change in the estimation. More interestingly, we find two peaks in the estimation, which means the two centers could be more relevant to the disease than others.

The full results are listed in Table 7, where the first 30 SNPs correspond to those plotted in Figure 3 and their orders are also kept the same, that is, rs6500315 is the name of the very left point in Figure 3 and rs8047222 corresponds to the point next to rs6500315 and so on. The last one, rs4783163, is the single SNP located far away from the previous 30 SNPs but also identified as disease-associated by spline-MCP. The probability of detection reported is the probability that certain signal is detected under model perturbation via subsampling as described above.

A close look at these identified SNPs reveals that near the higher peak, from rs1861759 to rs1861758 and from rs1077861 to rs8060598, these are several expression quantitative trait

---

**Table 6. Specificity and sensitivity when noises follow $t_5$**

| Specificity | Fused | Spline-MCP | Spline-Lasso | Smooth | Fused/Thresh. | Spline/Thresh. | Smooth/Thresh. |
|-------------|-------|------------|-------------|--------|--------------|---------------|-------------|
| Case 1      | 0.17  | 0.88       | 0.05        | 0.06   | 0.96         | 0.99          | 0.99        |
| Case 2      | 0.65  | 0.99       | 0.04        | 0.09   | 0.99         | 0.99          | 0.99        |

| Sensitivity | Fused | Spline-MCP | Spline-Lasso | Smooth | Fused/Thresh. | Spline/Thresh. | Smooth/Thresh. |
|-------------|-------|------------|-------------|--------|--------------|---------------|-------------|
| Case 1      | 1.00  | 0.95       | 1.00        | 1.00   | 0.97         | 0.92          | 0.96        |
| Case 2      | 1.00  | 0.97       | 1.00        | 1.00   | 1.00         | 0.96          | 0.96        |

**Table 7. Diseases-associated SNPs identified by spline-MCP**

| SNP name     | Prob. of detection | SNP name     | Prob. of detection |
|--------------|--------------------|--------------|--------------------|
| rs6500315    | 0.86               | rs2076756    | 1.00               |
| rs8047222    | 0.67               | rs1077861    | 0.83               |
| rs7199150    | 0.54               | rs3135499    | 0.99               |
| rs7186163    | 0.81               | rs8060598    | 1.00               |
| rs8050932    | 0.75               | rs3785142    | 1.00               |
| rs1981760    | 0.68               | rs7342715    | 1.00               |
| rs9933594    | 0.77               | rs3135503    | 1.00               |
| rs4785448    | 0.81               | rs4785450    | 1.00               |
| rs8057341    | 0.83               | rs11076540   | 1.00               |
| rs17221417   | 1.00               | rs7205760    | 0.83               |
| rs17312836   | 1.00               | rs17224078   | 0.86               |
| rs2066843    | 1.00               | rs10852609   | 0.82               |
| rs1861759    | 0.99               | rs6500336    | 0.83               |
| rs751271     | 0.83               | rs9922324    | 0.63               |
| rs748855     | 0.99               | rs4783163    | 0.51               |
| rs1861758    | 1.00               |
loci (eQTLs), which are the SNPs that can regulate gene expression, of CARD15 which is a known gene associated with Crohn’s disease. Therefore, our methods can indeed help highlight important SNPs from less important ones by looking at the shape of the estimation. In addition, among identified SNPs, there are also some eQTLs, from rs1981760 to rs8057341, of SLIC1 which is another gene considered to be associated with Crohn’s disease. More related information can be found at http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/.

Comparing our findings with those in Yang et al. (2011), we are able to find all those SNPs in Chromosome 16 reported in their Table 2. Moreover, as pointed out above, our proposed method can also potentially differentiate important SNPs from those less important ones by looking at the shape of the curve.

5. DISCUSSION

In this article, we consider a high-dimensional linear regression problem, where the features are ordered in some meaningful way. When the coefficients are sparse and change smoothly, we propose a spline-lasso or spline-MCP method, which combines the lasso or MCP penalty and cubic spline estimation procedure (i.e., penalizing second-order derivatives of the coefficients). The estimation and prediction of the newly proposed methods were more accurate compared to some existing methods. Furthermore, whenever it is necessary, adopting a second-stage thresholding procedure can improve the variable selection as well as the prediction and estimation.

The spline-lasso or spline-MCP method tried to achieve feature selection with smooth structure on the coefficients. These objectives are often conflicting with each other. It turns out that spline-MCP works very well for this purpose. For other methods, trying to achieve several targets might work less effectively. In this case, one might divide the task in several stages, and focus on one objective at a time.

When the response $y_i$ is binary, for example, $\{0, 1\}$ representing two classes like “healthy” and “sick,” it would be better to use logistic regression rather than the linear model (1). The idea of the spline-lasso could be easily extended to this context, which we are currently exploring.

APPENDIX

We now give a sketch to the proofs of the main theorems. The full proofs can be found in the supplementary materials.

Proof of Theorem 1

We first give two lemmas.

Lemma A.1. Let $\eta \in (0, 1)$. Denote $A_{n,p} = \{\max_{j=1, \ldots, p} 2|V_j| \leq \lambda_1/2\}$ where $V_j = n^{-1} \sum_{i=1}^{n} x_{i,j} \varepsilon_i$. Let $\lambda_1$ be defined as in Theorem 1. Then

$$P \left( \max_{j=1, \ldots, p} 2|V_j| \leq \lambda_1/2 \right) \geq 1 - \eta.$$

Lemma A.2. On the event $A_{n,p} = \{\max_{j=1, \ldots, p} 2|V_j| \leq \lambda_1/2\}$, we have

$$\frac{1}{n} \| \hat{X} \beta - \hat{X} \hat{\beta}_{\text{spline}} \|_2^2 + \frac{\lambda_1}{2} \| \beta - \hat{\beta}_{\text{spline}} \|_1 \leq r_{\nu} \| \beta_B - \hat{\beta}_{\text{spline}} \|_2,$$

where $\hat{X} = (X, \sqrt{n} \tau_{L} L)^T$, $r_{\nu} = 2\lambda_1 \sqrt{\lambda_1} + 2\lambda_2 \| \hat{J} \beta \|_2$, and $B \supseteq A$.

We return to the proof of Theorem 1. From Lemmas A.1-A.2, we have $\| \beta - \hat{\beta}_{\text{spline}} \|_1 \leq \varphi_{\nu} \| \beta_B - \hat{\beta}_{\text{spline}} \|_2$, where $\varphi_{\nu} = 2r_{\nu}/\lambda_1 = 4\sqrt{\lambda_1} + 4\lambda_2 \lambda_1^{-1} \| \hat{J} \beta \|_2$. Then the vector $\beta - \hat{\beta}_{\text{spline}}$ satisfies Assumption 1. Therefore, we get

$$\frac{1}{n} \| \hat{X} \beta - \hat{X} \hat{\beta}_{\text{spline}} \|_2^2 \leq \frac{r_{\nu}}{\sqrt{n} \varphi_{\nu} } \| \hat{X} \beta - \hat{X} \hat{\beta}_{\text{spline}} \|_2,$$

which will easily lead to the first and the third part of the result

$$\frac{1}{n} \| X \beta - X \hat{\beta}_{\text{spline}} \|_2^2 \leq \phi_{\nu}^{-1} \left( 2\lambda_1 \sqrt{\lambda_1} + 2\lambda_2 \| \hat{J} \beta \|_2 \right)^2,$$

and when $\lambda_2 \neq 0$, we have $\| L(\beta - \hat{\beta}_{\text{spline}}) \|_2 \leq \lambda_2^{-1/2} \phi_{\nu}^{-1/2} (2\lambda_1 \sqrt{\lambda_1} + 2\lambda_2 \| \hat{J} \beta \|_2)$. Then the first part, together with Assumption 1, yields the second inequality.
Proof of Theorem 2

Let \( \hat{X} = \sqrt{n}(\Sigma + \lambda_2 I)^{-1/2} \). \( \tilde{Y} = \hat{X}^T Y \) and \( \tilde{M}(\mathbf{b}; \lambda, \gamma) = (2n)^{-1} \tilde{Y}^T \tilde{X}^T \tilde{X} \). Thus, \( \hat{\beta} \) is also the minimizer of \( \tilde{M}(\mathbf{b}; \lambda, \gamma) \). To show Theorem 2, it suffices to show \( P(\Omega_1' \cup \Omega_2') \leq 3\epsilon \), where

\[
\Omega_1 = \left\{ \|\hat{X}^T (\tilde{Y} - \hat{\beta}^* / n)\|_\infty \leq \lambda_1 \right\},
\Omega_2 = \left\{ \min_{j \in \mathcal{O}} \sgn(\beta_j^*) \hat{\beta}_j^* \geq \gamma \lambda_1 \right\}.
\]

Note \( \Omega_1 \) is related to the gradient in the linear model \( \hat{X} = \tilde{X}^* + \tilde{e}^* \). It can be shown that \( \Omega_1 \supseteq \{ \|X^T \tilde{e} / n\|_\infty < \lambda_1 - \lambda_2 C_2 \} \). If the event on the right-hand side holds, we can show

\[
\|\hat{X}^T (\tilde{Y} - \hat{\beta}^* / n)\|_\infty \leq \|X^T \tilde{e} / n\|_\infty + \|X^T E \tilde{e}^* / n\|_\infty \leq \lambda_1.
\]

Since \( \tilde{e}^* = E \tilde{\beta}^*, \tilde{\beta}^* = \Sigma^{-1}(\lambda_2 X^TY/n) \) can be written as \( \beta^*_0 + ((\tilde{v}_j / n)^2 u^T_e, j \in \mathcal{O})^T \), where \( \|u_j\|_2 = 1 \) and \( \{v_j, j \in \mathcal{O}\} \) are the diagonal elements of \( \Sigma_0^{-1}\Sigma_0 \Sigma_0^{-1} \). Thus, \( \Omega_2 \subseteq \bigcup_{j \in \mathcal{O}} \{ \min_{j \in \mathcal{O}} \sgn(\beta_j^*) \hat{\beta}_j^* \geq \gamma \lambda_1 \( \sqrt{2} (\sqrt{c_2} \lambda_1) \) \}. \Omega_1 = \{ \xi (\tilde{Y} - \tilde{\beta}^*; d^* - \mathcal{O}, \mathcal{O}, \lambda_2) \leq \lambda_1 \} \), where \( m \geq 1 \) and vector \( v \) in the range of \( \hat{X} \).

\[
x(\mathcal{O}; m, \mathcal{O}, \lambda_2) = \max \left\{ \frac{\|B \hat{\beta} - \hat{P}_B \mathbf{v}\|_2}{(mn)^{1/2}} : \mathcal{O} \subseteq B \subseteq \{1, \ldots, p\}, |B| = m + |\mathcal{O}| \right\},
\]

and \( \hat{P}_B = \hat{X}^T \hat{X} \hat{X}^{+} \). Similarly to Theorem 2, we have \( \Omega_1 \supseteq \{ \|X^T \tilde{e} / n\|_\infty < \lambda_1 - \lambda_2 C_2 \} \). In addition, a slight modification on the proof of Theorem 2 leads to

\[
\Omega_2' \subseteq \bigcup_{j \in \mathcal{O}} \{ \min_{j \in \mathcal{O}} \sgn(\beta_j^*) \hat{\beta}_j^* \geq \gamma (\sigma_2 (\sqrt{c_2} \lambda_1)) \} \geq \sigma \sqrt{2} \log (\|\mathcal{O}| / e)}\}
\]

Finally, it can be shown that \( \Omega_1 \supseteq \{ \|X^T \tilde{e} / n\|_\infty < \sqrt{c_2 (\lambda_2) \mathcal{K}_\mathcal{O} (\mathcal{K}_\mathcal{O} + 1) \lambda_1 - \lambda_2 C_2} \} \). Arguing as in the proof of Theorem 2, we have \( P(\Omega_1' \cup \Omega_2' \cup \Omega_2) \leq 2p(e/p) + |\mathcal{O}| / |\mathcal{O}| = 3\epsilon \). The proof is completed since \( \hat{\beta}_j^* \neq 0 \) for all \( j \in \mathcal{O} \) in \( \Omega_2 \).

**SUPPLEMENTARY MATERIAL**

The supplementary materials contain more detailed proofs to Theorems 1-3.

[Received November 2013. Revised December 2014.]

**REFERENCES**

Bickel, P., Ritov, Y., and Tsybakov, A. (2009), “Simultaneous Analysis of Lasso and Dantzig Selector,” *Annals of Statistics*, 37, 1705–1732. [290]

Candes, E., and Tao, T. (2007), “The Dantzig Selector: Statistical Estimation When p is Much Larger Than n,” *Annals of Statistics*, 35, 2313–2351. [288]

Consoritum, T. W. T. C. C. (2007), “Genome-Wide Association Study of 14, 000 Cases of Seven Common Diseases and 3000 Shared Controls,” *Nature*, 447, 661–678. [295]

Donoho, D. L., and Johnstone, I. M. (1998), “Minimax Estimation via Wavelet Shrinkage,” *Annals of Statistics*, 26, 879–921. [292]

Efroyn, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), “Least Angle Regression,” *Annals of Statistics*, 32, 407–499. [289,290]

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. [288]

Friedman, J., Hastie, T., Hofling, H., and Tibshirani, R. (2007), “Pathwise Coordinate Optimization,” *Annals of Applied Statistics*, 2, 302–332. [289]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer. [292]

Hebiri, M., and Van de Geer, S. (2011), “The Smooth-Lasso and Other \( \ell_1 + \ell_2 \)-penalized Methods,” *Electronic Journal of Statistics*, 5, 1184–1226. [289,289,292]

Huang, J., Ma, S., Li, H., and Zhang, C. H. (2011), “The Sparse Laplacian Shrinkage Estimator for High-Dimensional Regression,” *Annals of Statistics*, 39, 2021–2046. [289,291]

Knight, K., and Fu, W. (2000), “Asymptotics for Lasso-Type Estimators,” *Annals of Statistics*, 28, 1356–1378. [288]

Meinshausen, N., and Buhlmann, P. (2010), “Stability Selection,” *Journal of the Royal Statistical Society*, Series B, 72, 417–473. [295]

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), “Sparsity and Smoothness via the Fused Lasso,” *Journal of the Royal Statistical Society*, Series B, 67, 91–108. [288,289,292]

Tibshirani, R., and Taylor, J. (2011), “The Solution Path of the Generalized Lasso,” *Annals of Statistics*, 39, 1335–1371. [289]

Tibshirani, R., and Wang, P. (2007), “Spatial Smoothing and Hot Spot Detection for CGH Data using the Fused Lasso,” *Biostatistics*, 9, 18–29. [288,289]

Tsybakov, A. B. (2009), *Introduction to Nonparametric Estimation*, New York: Springer. [292]

van de Geer, S., and Buhlmann, P. (2009), “On the Conditions used to Prove Oracle Results for the Lasso,” *Electronic Journal of Statistics*, 3, 1360–1392.

Yang, C., Zhou, X. W., Wan, X., Yang, Q., Xue, H., and Yu, W. C. (2011), “Identifying Disease-Associated SNP Clusters via Contiguous Outlier Detection,” *Bioinformatics*, 27, 2578–2585. [289,289,296]

Yuan, M., and Lin, Y. (2006), “Model Selection and Estimation in Regression With Grouped Variables,” *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [288]

Zhang, C. H. (2010), “Nearly Unbiased Variable Selection Under Minimax Concave Penalty,” *Annals of Statistics*, 38, 894–942. [288,289,290,291,297]

Zou, H. (2006), “The Adaptive Lasso and Its Oracle Properties,” *Journal of the American Statistical Association*, 97, 210–221. [288]

Zou, H., and Hastie, T. (2005), “Regularization and Variable Selection via the Elastic Net,” *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [288]