THE SPARSE LAPLACIAN SHRINKAGE ESTIMATOR FOR HIGH-DIMENSIONAL REGRESSION

BY JIAN HUANG[1], SHUANGGE MA[2], HONGZHE LI[3] AND CUN-HUI ZHANG[4]

University of Iowa, Yale University, University of Pennsylvania and Rutgers University

We propose a new penalized method for variable selection and estimation that explicitly incorporates the correlation patterns among predictors. This method is based on a combination of the minimax concave penalty and Laplacian quadratic associated with a graph as the penalty function. We call it the sparse Laplacian shrinkage (SLS) method. The SLS uses the minimax concave penalty for encouraging sparsity and Laplacian quadratic penalty for promoting smoothness among coefficients associated with the correlated predictors. The SLS has a generalized grouping property with respect to the graph represented by the Laplacian quadratic. We show that the SLS possesses an oracle property in the sense that it is selection consistent and equal to the oracle Laplacian shrinkage estimator with high probability. This result holds in sparse, high-dimensional settings with $p \gg n$ under reasonable conditions. We derive a coordinate descent algorithm for computing the SLS estimates. Simulation studies are conducted to evaluate the performance of the SLS method and a real data example is used to illustrate its application.

1. Introduction. There has been much work on penalized methods for variable selection and estimation in high-dimensional regression models. Several important methods have been proposed. Examples include estimators based on the bridge penalty [Frank and Friedman (1993)], the ℓ_1 penalty or the least absolute shrinkage and selection operator [LASSO, Tibshirani (1996), Chen, Donoho and Saunders (1998)], the smoothly clipped absolute deviation (SCAD) penalty [Fan (1997), Fan and Li (2001)] and the minimum concave penalty [MCP, Zhang (2010)]. These methods are able to do estimation and automatic variable selection simultaneously and provide a computationally feasible way for variable selection in high-dimensional settings. Much progress has been made in understanding the

theoretical properties of these methods. Efficient algorithms have also been developed for implementing these methods.

A common feature of the methods mentioned above is the independence between the penalty and the correlation among predictors. This can lead to unsatisfactory selection results, especially in $p \gg n$ settings. For example, as pointed out by Zou and Hastie (2005), the LASSO tends to only select one variable among a group of highly correlated variables; and its prediction performance may not be as good as the ridge regression if there exists high correlation among predictors. To overcome these limitations, Zou and Hastie (2005) proposed the elastic net (Enet) method, which uses a combination of the ℓ_1 and ℓ_2 penalties. Selection properties of the Enet and adaptive Enet have also been studied by Jia and Yu (2010) and Zou and Zhang (2009). Bondell and Reich (2008) proposed the OSCAR (octagonal shrinkage and clustering algorithm for regression) approach, which uses a combination of the ℓ_1 norm and a pairwise ℓ_∞ norm for the coefficients. Huang et al. (2010a) proposed the Mnet method, which uses a combination of the MCP and ℓ_2 penalties. The Mnet estimator is equal to the oracle ridge estimator with high probability under certain conditions. These methods are effective in dealing with certain types of collinearity among predictors and has the useful grouping property of selecting and dropping highly correlated predictors together. Still, these combination penalties do not use any specific information on the correlation pattern among the predictors.

Li and Li (2008) proposed a network-constrained regularization procedure for variable selection and estimation in linear regression models, where the predictors are genomic data measured on genetic networks. Li and Li (2010) considered the general problem of regression analysis when predictors are measured on an undirected graph, which is assumed to be known a priori. They called their method a graph-constrained estimation procedure or GRACE. The GRACE penalty is a combination of the ℓ_1 penalty and a penalty that is the Laplacian quadratic associated with the graph. Because the GRACE uses the ℓ_1 penalty for selection and sparsity, it has the same drawbacks as the Enet discussed above. In addition, the full knowledge of the graphical structure for the predictors is usually not available, especially in high-dimensional problems. Daye and Jeng (2009) proposed the weighted fusion method, which also uses a combination of the ℓ_1 penalty and a quadratic form that can incorporate information among correlated variables for estimation and variable selection. Tutz and Ulbricht (2009) studied a form of correlation based penalty, which can be considered a special case of the general quadratic penalty. But this approach does not do variable selection. The authors proposed a blockwise boosting procedure in combination with the correlation based penalty for variable selection. Hebiri and van de Geer (2010) studied the theoretical properties of the smoothed-Lasso and other $\ell_1 + \ell_2$-penalized methods in $p \gg n$ models. Pan, Xie and Shen (2011) studied a grouped penalty based on the L_γ-norm for $\gamma > 1$ that smoothes the regression coefficients over a network. In particular, when $\gamma = 2$ and after appropriate rescaling of the regression coefficients, this group L_γ

penalty simplifies to the group Lasso [Yuan and Lin (2006)] with the nodes in the network as groups. This method is capable of group selection, but it does not do individual variable selection. Also, because the group L_γ penalty is convex for $\gamma > 1$, it does not lead to consistent variable selection, even at the group level.

We propose a new penalized method for variable selection and estimation in sparse, high-dimensional settings that takes into account certain correlation patterns among predictors. We consider a combination of the MCP and Laplacian quadratic as the penalty function. We call the proposed approach the sparse Laplacian shrinkage (SLS) method. The SLS uses the MCP to promote sparsity and Laplacian quadratic penalty to encourage smoothness among coefficients associated with the correlated predictors. An important advantage of the MCP over the ℓ_1 penalty is that it leads to estimators that are nearly unbiased and achieve selection consistency under weaker conditions [Zhang (2010)].

The contributions of this paper are as follows.

- First, unlike the existing methods that use an ℓ_1 penalty for selection and a ridge penalty or a general ℓ_2 penalty for dealing with correlated predictors, we use the MCP to achieve nearly unbiased selection and proposed a concrete class of quadratics, the Laplacians, for incorporating correlation patterns among predictors in a local fashion. In particular, we suggest to employ the approaches for network analysis for specifying the Laplacians. This provides an implementable strategy for incorporating correlation structures in high-dimensional data analysis.

- Second, we prove that the SLS estimator is sign consistent and equal to the oracle Laplacian shrinkage estimator under reasonable conditions. This result holds for a large class of Laplacian quadratics. An important aspect of this result is that it allows the number of predictors to be larger than the sample size. In contrast, the works of Daye and Jeng (2009) and Tutz and Ulbricht (2009) do not contain such results in $p \gg n$ models. The selection consistency result of Hebiri and van de Geer (2010) requires certain strong assumptions on the magnitude of the smallest regression coefficient (their Assumption C) and on the correlation between important and unimportant predictors (their Assumption D), in addition to a variant of the restricted eigenvalue condition (their Assumption B). In comparison, our assumption involving the magnitude of the regression coefficients is weaker and we use a sparse Riese condition instead of imposing restriction on the correlations among predictors. In addition, our selection results are stronger in that the SLS estimator is not only sign consistent, but also equal to the oracle Laplacian shrinkage estimator with high probability. In general, similar results are not available with the use of the ℓ_1 penalty.

- Third, we show that the SLS method is potentially capable of incorporating correlation structure in the analysis without incurring extra bias. The Enet and the more general $\ell_1 + \ell_2$ methods in general introduces extra bias due to the quadratic penalty, in addition to the bias resulting from the ℓ_1 penalty. To the

best of our knowledge, this point has not been discussed in the existing literature. We also demonstrate that the SLS has certain local smoothing property with respect to the graphical structure of the predictors.

- Fourth, unlike in the GRACE method, the SLS does not assume that the graphical structure for the predictors is known a priori. The SLS uses the existing data to construct the graph Laplacian or to augment partial knowledge of the graph structure.

- Fifth, our simulation studies demonstrate that the SLS method outperforms the ℓ_1 penalty plus a quadratic penalty approach as studied in Daye and Jeng (2009) and Hebiri and van de Geer (2010). In our simulation examples, the SLS in general has smaller empirical false discovery rates with comparable false negative rates. It also has smaller prediction errors.

This paper is organized as follows. In Section 2, we define the SLS estimator. In Section 3 we discuss ways to construct graph Laplacian, or equivalently, its corresponding adjacency matrix. In Section 4, we study the selection properties of the SLS estimators. In Section 5, we investigate the properties of Laplacian shrinkage. In Section 6, we describe a coordinate descent algorithm for computing the SLS estimators, present simulation results and an application of the SLS method to a microarray gene expression dataset. Discussions of the proposed method and results are given in Section 7. Proofs for the oracle properties of the SLS and other technical details are provided in the Appendix.

2. The sparse Laplacian shrinkage estimator. Consider the linear regression model

$$(2.1) \qquad \mathbf{y} = \sum_{j=1}^{p} \mathbf{x}_j \beta_j + \boldsymbol{\varepsilon}$$

with n observations and p potential predictors, where $\mathbf{y} = (y_1, \ldots, y_n)'$ is the vector of n response variables, $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})'$ is the jth predictor, β_j is the jth regression coefficient and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is the vector of random errors. Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ be the $n \times p$ design matrix. Throughout, we assume that the response and predictors are centered and the predictors are standardized so that $\sum_{i=1}^{n} x_{ij}^2 = n$, $j = 1, \ldots, p$. For $\lambda = (\lambda_1, \lambda_2)$ with $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$, we propose the penalized least squares criterion

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

$$(2.2)$$

$$+ \frac{1}{2}\lambda_2 \sum_{1 \leq j < k \leq p} |a_{jk}|(b_j - s_{jk}b_k)^2,$$

where $\| \cdot \|$ denotes the ℓ_2 norm, ρ is the MCP with penalty parameter λ_1 and regularization parameter γ, $|a_{jk}|$ measures the strength of the connection between $\mathbf{x}_j$

and $\mathbf{x}_k$, and $s_{jk} = \mathrm{sgn}(a_{jk})$ is the sign of a_{jk}, with $\mathrm{sgn}(t) = -1, 0$ or 1, respectively, for $t < 0, = 0$ or > 0. The two penalty terms in (2.2) play different roles. The first term promotes sparsity in the estimated model. The second term encourages smoothness of the estimated coefficients of the connected predictors. We can associate the quadratic form in this term with the Laplacian for a suitably defined undirected weighted graph for the predictors. See the description below. For any given (λ, γ), the SLS estimator is

$$(2.3) \qquad \hat{\boldsymbol{\beta}}(\lambda, \gamma) = \arg\min_{\mathbf{b}} M(\mathbf{b}; \lambda, \gamma).$$

The SLS uses the MCP, defined as

$$(2.4) \qquad \rho(t; \lambda_1, \gamma) = \lambda_1 \int_0^{|t|} \left(1 - x/(\gamma\lambda_1)\right)_+ dx,$$

where for any $a \in R$, a_+ is the nonnegative part of a, that is, $a_+ = a 1_{\{a \geq 0\}}$. The MCP can be easily understood by considering its derivative,

$$(2.5) \qquad \dot{\rho}(t; \lambda_1, \gamma) = \lambda_1 \left(1 - |t|/(\gamma\lambda_1)\right)_+ \mathrm{sgn}(t).$$

We observe that the MCP begins by applying the same level of penalization as the ℓ_1 penalty, but continuously reduces that level to 0 for $|t| > \gamma\lambda$. The regularization parameter γ controls the degree of concavity. Larger values of γ make ρ less concave. By sliding the value of γ from 1 to ∞, the MCP provides a continuum of penalties with the hard-threshold penalty as $\gamma \to 1+$ and the convex ℓ_1 penalty at $\gamma = \infty$. Detailed discussion of MCP can be found in Zhang (2010).

The SLS also allows the use of different penalties than the MCP for ρ, including the SCAD [Fan (1997), Fan and Li (2001)] and other quadratic splines. Because the MCP minimizes the maximum concavity measure and has the simplest form among nearly unbiased penalties in this family, we choose it as the default penalty for the SLS. Further discussion of the MCP and its comparison with the LASSO and SCAD can be found in Zhang (2010) and Mazumder, Friedman and Hastie (2009).

We express the nonnegative quadratic form in the second penalty term in (2.2) using a positive semi-definite matrix L, which satisfies

$$\mathbf{b}' L \mathbf{b} = \sum_{1 \leq j < k \leq p} |a_{jk}| (b_j - s_{jk} b_k)^2 \qquad \forall \mathbf{b} \in \mathbb{R}^p.$$

For simplicity, we confine our discussion to the symmetric case where $a_{kj} = a_{jk}, 1 \leq j < k \leq p$. Since the diagonal elements a_{jj} do not appear in the quadratic form, we can define them any way we like for convenience. Let $A = (a_{jk}, 1 \leq j, k \leq p)$ and $D = \mathrm{diag}(d_1, \dots, d_p)$, where $d_j = \sum_{k=1}^p |a_{jk}|$. We have $\sum_{1 \leq j < k \leq p} |a_{jk}| (b_j - s_{jk} b_k)^2 = \mathbf{b}'(D - A)\mathbf{b}$. Therefore, $L = D - A$. This matrix is associated with a labeled weighted graph $\mathcal{G} = (V, \mathcal{E})$ with vertex set $V = \{1, \dots, p\}$ and edge set $\mathcal{E} = \{(j, k): (j, k) \in V \times V\}$. Here the $|a_{jk}|$ is the

weight of edge (j, k) and d_j is the degree of vertex j. The d_j is also called the connectivity of vertex j. The matrix L is called the Laplacian of $\mathcal{G}$ and A its signed adjacency matrix [Chung (1997)]. The edge (j, k) is labeled with the "+" or "−" sign, but its weight $|a_{jk}|$ is always nonnegative. We use a labeled graph to accommodate the case where two predictors can have a nonzero adjacency coefficient but are negatively correlated. Note that the usual adjacency matrix can be considered a special case of signed adjacency matrix when all $a_{jk} \geq 0$. For simplicity, we will use the term adjacency matrix below.

We usually require that the adjacency matrix to be sparse in the sense that many of its entries are zero or nearly zero. With a sparse adjacency matrix, the main characteristic of the shrinkage induced by the Laplacian penalty is that it occurs locally for the coefficients associated with the predictors connected in the graph. Intuitively, this can be seen by writing

$$\lambda_2 \sum_{1 \leq j < k \leq p} |a_{jk}|(b_j - s_{jk}b_k)^2 = \frac{1}{2}\lambda_2 \sum_{(j,k):\, a_{jk} \neq 0} |a_{jk}|(b_j - s_{jk}b_k)^2.$$

Thus for $\lambda_2 > 0$, the Laplacian penalty shrinks $b_j - s_{jk}b_k$ toward zero for $a_{jk} \neq 0$. This can also be considered as a type of local smoothing on the graph $\mathcal{G}$ associated with the adjacency matrix A. In comparison, the shrinkage induced by the ridge penalty used in the Enet is global in that it shrinks all the coefficients toward zero, regardless of the correlation structure among the predictors. We will discuss the Laplacian shrinkage in more detail in Section 5.

Using the matrix notation, the SLS criterion (2.2) can be written as

$$(2.6) \quad M(\mathbf{b}; \lambda, \gamma) = \frac{1}{2n}\|\mathbf{y} - X\mathbf{b}\|^2 + \sum_{j=1}^{p} \rho(|b_j|; \lambda_1, \gamma) + \frac{1}{2}\lambda_2 \mathbf{b}'(D - A)\mathbf{b}.$$

Here the Laplacian is not normalized, meaning that the weight d_j is not standardized to 1. In problems where predictors should be treated without preference with respect to connectivity, we can first normalized the Laplacian $L^* = I_p - A^*$ with $A^* = D^{-1/2}AD^{-1/2}$ and use the criterion

$$M^*(\mathbf{b}; \lambda, \gamma) = \frac{1}{2n}\|\mathbf{y} - X\mathbf{b}\|^2 + \sum_{j=1}^{p} \rho(|b_j|; \lambda_1, \gamma) + \frac{1}{2}\lambda_2 \mathbf{b}'(\mathbf{I}_p - A^*)\mathbf{b}.$$

Technically, a normalized Laplacian L^* can be considered a special case of a general L. We only consider the SLS estimator based on the criterion (2.6) when studying its properties. In network analysis of gene expression data, genes with large connectivity also tend to have important biological functions [Zhang and Horvath (2005)]. Therefore, it is prudent to provide more protection for such genes in the selection process.

3. Construction of adjacency matrix. In this section, we describe several simple forms of adjacency measures proposed by Zhang and Horvath (2005), which have have been successfully used in network analysis of gene expression data. The adjacency measure is often defined based on the notion of dissimilarity or similarity.

(i) A basic and widely used dissimilarity measure is the Euclidean distance. Based on this distance, we can define adjacency coefficient as $a_{jk} = \phi(\|\mathbf{x}_j - \mathbf{x}_k\|/\sqrt{n})$, where $\phi : [0, \infty) \mapsto [0, \infty)$. A simple adjacency function is the threshold function $\phi(x) = 1\{x \leq 2r\}$. Then

$$(3.1) \qquad a_{jk} = \begin{cases} 1, & \text{if } \|\mathbf{x}_j - \mathbf{x}_k\|/\sqrt{n} \leq 2r, \\ 0, & \text{if } \|\mathbf{x}_j - \mathbf{x}_k\|/\sqrt{n} > 2r. \end{cases}$$

It is convenient to express a_{jk} in terms of the Pearson's correlation coefficient r_{jk} between $\mathbf{x}_j$ and $\mathbf{x}_k$, where $r_{jk} = \mathbf{x}_j'\mathbf{x}_k/(\|\mathbf{x}_j\|\|\mathbf{x}_k\|)$. For predictors that are standardized with $\|\mathbf{x}_j\|^2 = n$, $1 \leq j \leq p$, we have $\|\mathbf{x}_j - \mathbf{x}_k\|^2/n = 2 - 2r_{jk}$. Thus in terms of correlation coefficients, we can write $a_{jk} = 1\{r_{jk} > r\}$. We determine the value of r based on the Fisher transformation $z_{jk} = 0.5 \log((1 + r_{jk})/(1 - r_{jk}))$. If the correlation between $\mathbf{x}_j$ and $\mathbf{x}_k$ is zero, $\sqrt{n-3}z_{jk}$ is approximately distributed as $N(0, 1)$. We can use this to determine a threshold c for $\sqrt{n-3}z_{jk}$. The corresponding threshold for r_{jk} is $r = (\exp(2c/\sqrt{n-3}) - 1)/(\exp(2c/\sqrt{n-3}) + 1)$.

We note that here we use the Fisher transformation to change the scale of the correlation coefficients from $[-1, 1]$ to the normal scale for determining the threshold value r, so that the adjacency matrix is relatively sparse. We are not trying to test the significance of correlation coefficients.

(ii) The adjacency coefficient in (3.1) is defined based on a dissimilarity measure. Adjacency coefficient can also be defined based on similarity measures. An often used similarity measure is Pearson's correlation coefficient r_{jk}. Other correlation measures such as Spearman's correlation can also be used. Let

$$s_{jk} = \text{sgn}(r_{jk}) \quad \text{and} \quad a_{jk} = s_{jk}1\{|r_{jk}| > r\}.$$

Here r can be determined using the Fisher transformation as above.

(iii) With the power adjacency function considered in Zhang and Horvath (2005),

$$a_{jk} = \max(0, r_{jk})^\alpha \quad \text{and} \quad s_{jk} = 1.$$

Here $\alpha > 0$ and can be determined by, for example, the scale-free topology criterion.

(iv) A variation of the above power adjacency function is

$$a_{jk} = |r_{jk}|^\alpha \quad \text{and} \quad s_{jk} = \text{sgn}(r_{jk}).$$

For the adjacency matrices given above, (i) and (ii) use dichotomized measures, whereas (iii) and (iv) use continuous measures. Under (i) and (iii), two covariates are either positively or not connected/correlated. In contrast, under (ii) and (iv), two covariates are allowed to be negatively connected/correlated.

There are many other ways for constructing an adjacency matrix. For example, a popular adjacency measure in cluster analysis is $a_{jk} = \exp(-\|\mathbf{x}_j - \mathbf{x}_k\|^2/n\tau^2)$ for $\tau > 0$. The resulting adjacency matrix $A = [a_{jk}]$ is the Gram matrix associated with the Gaussian kernel. For discrete covariates, the Pearson correlation coefficient can still be used as a measure of correlation or association between two discrete predictors or between a discrete predictor and a continuous one. For example, for single nucleotide polymorphism data, Pearson's correlation coefficient is often used as a measure of linkage disequilibrium (i.e., association) between two markers. Other measures, such as odds ratio or measure of association based on contingency table can also be used for r_{jk}.

We note that how to construct the adjacency matrix is problem specific. Different applications may require different adjacency matrices. Since construction of adjacency matrix is not the focus of the present paper, we will only consider the use of the four adjacency matrices described above in our numerical studies in Section 6.

4. Oracle properties. In this section, we study the theoretical properties of the SLS estimator. Let the true value of the regression coefficient be $\boldsymbol{\beta}^o = (\beta_1^o, \ldots, \beta_p^o)'$. Denote $\mathcal{O} = \{j : \beta_j^o \neq 0\}$, which is the set of indices of nonzero coefficients. Let $d^o = |\mathcal{O}|$ be the cardinality of $\mathcal{O}$. Define

$$(4.1) \qquad \hat{\boldsymbol{\beta}}^o(\lambda_2) = \arg\min_{\mathbf{b}} \left\{ \frac{1}{2n}\|\mathbf{y} - X\mathbf{b}\|^2 + \frac{1}{2}\lambda_2 \mathbf{b}'L\mathbf{b}, \, b_j = 0, \, j \notin \mathcal{O} \right\}.$$

This is the oracle Laplacian shrinkage estimator on the set $\mathcal{O}$. Theorems 1 and 2 below provide sufficient conditions under which $P(\mathrm{sgn}(\hat{\boldsymbol{\beta}}) \neq \mathrm{sgn}(\boldsymbol{\beta}^o)$ or $\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^o) \to 0$. Thus, under those conditions, the SLS estimator is sign consistent and equal to $\hat{\boldsymbol{\beta}}^o$ with high probability.

We need the following notation in stating our results. Let $\Sigma = n^{-1}X'X$. For any $A \cup B \subseteq \{1, \ldots, p\}$, vectors $\mathbf{v}$, the design matrix X and $V = (v_{ij})_{p \times p}$, define

$$\mathbf{v}_B = (v_j, j \in B)', \qquad X_B = (\mathbf{x}_j, j \in B),$$

$$V_{A,B} = (v_{ij}, i \in A, j \in B)_{|A| \times |B|}, \qquad V_B = V_{B,B}.$$

For example, $\Sigma_B = X_B' X_B/n$ and $\Sigma_{\mathcal{O}}(\lambda_2) = \Sigma_{\mathcal{O}} + \lambda_2 L_{\mathcal{O}}$. Let $|B|$ denote the cardinality of B. Let $c_{\min}(\lambda_2)$ be the smallest eigenvalue of $\Sigma + \lambda_2 L$. We use the following constants to bound the bias of the Laplacian:

$$(4.2) \qquad \begin{aligned} C_1 &= \|\Sigma_{\mathcal{O}}^{-1}(\lambda_2)L_{\mathcal{O}}\boldsymbol{\beta}_{\mathcal{O}}^o\|_\infty, \\ C_2 &= \|\{\Sigma_{\mathcal{O}^c,\mathcal{O}}(\lambda_2)\Sigma_{\mathcal{O}}^{-1}(\lambda_2)L_{\mathcal{O}} - L_{\mathcal{O}^c,\mathcal{O}}\}\boldsymbol{\beta}_{\mathcal{O}}^o\|_\infty. \end{aligned}$$

We make the following sub-Gaussian assumption on the error terms in (2.1).

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

$$\sup_{\|\mathbf{u}\|=1} P\{\mathbf{u}'\boldsymbol{\varepsilon} > \sigma t\} \leq e^{-t^2/2}, \qquad 0 < t \leq \sqrt{2\log(p/\epsilon)}.$$

4.1. *Convex penalized loss.* We first consider the case where $\Sigma(\lambda_2) = \Sigma + \lambda_2 L$ is positive definite. Since (4.1) is the minimizer of the Laplacian restricted to the support $\mathcal{O}$, it can be explicitly written as

$$(4.3) \qquad \hat{\boldsymbol{\beta}}_{\mathcal{O}}^o = (\Sigma_{\mathcal{O}} + \lambda_2 L_{\mathcal{O}})^{-1} X_{\mathcal{O}}' \mathbf{y}/n, \qquad \hat{\boldsymbol{\beta}}_{\mathcal{O}^c}^o = 0,$$

provided that $\Sigma_{\mathcal{O}}(\lambda_2)$ is invertible. Its expectation $\boldsymbol{\beta}^* = E\hat{\boldsymbol{\beta}}^o$, considered as a target of the SLS estimator, must satisfy

$$(4.4) \qquad \boldsymbol{\beta}_{\mathcal{O}}^* = (\Sigma_{\mathcal{O}} + \lambda_2 L_{\mathcal{O}})^{-1} \Sigma_{\mathcal{O}} \boldsymbol{\beta}^o, \qquad \boldsymbol{\beta}_{\mathcal{O}^c}^* = 0.$$

CONDITION (B). (i) $c_{\min}(\lambda_2) > 1/\gamma$ with $\rho(t; \lambda_1, \gamma)$ in (2.2).
(ii) The penalty levels satisfy

$$\lambda_1 \geq \lambda_2 C_2 + \sigma\sqrt{2\log((p-d^o)/\epsilon)} \max_{j \leq p} \|\mathbf{x}_j\|/n$$

with C_2 in (4.2).
(iii) With $\{v_j, j \in \mathcal{O}\}$ being the diagonal elements of $\Sigma_{\mathcal{O}}^{-1}(\lambda_2)\Sigma_{\mathcal{O}}\{\Sigma_{\mathcal{O}}^{-1}(\lambda_2)\}$,

$$\min_{j \in \mathcal{O}}\{|\beta_j^*|(n/v_j)^{1/2}\} \geq \sigma\sqrt{2\log(d^o/\epsilon)}.$$

Define $\beta_* = \min\{|\beta_j^o|, j \in \mathcal{O}\}$. If $\mathcal{O}$ is an empty set, that is, when all the regression coefficients are zero, we set $\beta_* = \infty$.

THEOREM 1. *Suppose Conditions* (A) *and* (B) *hold. Then*

$$(4.5) \qquad P(\{j : \hat{\beta}_j \neq 0\} \neq \mathcal{O} \text{ or } \hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^o) \leq 3\epsilon.$$

If $\beta_* \geq \lambda_2 C_1 + \max_j \sqrt{(2v_j/n)\log(d^o/\epsilon)}$ *instead of Condition* (B)(iii), *then*

$$(4.6) \qquad P(\text{sgn}(\hat{\boldsymbol{\beta}}) \neq \text{sgn}(\boldsymbol{\beta}^o) \text{ or } \hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^o) \leq 3\epsilon.$$

Here note that p, d^o, γ *and* $c_{\min}(\lambda_2)$ *are all allowed to depend on* n.

The probability bound on the selection error in Theorem 1 is nonasymptotic. If the conditions of Theorem 1 hold with $\epsilon \to 0$, then (4.5) implies selection consistency of the SLS estimator and (4.6) implies sign consistency. The conditions are mild. Condition (A) concerns the tail probabilities of the error distribution and is satisfied if the errors are normally distributed. Condition (B)(i) ensures that the SLS criterion is strictly convex so that the solution is unique. The oracle estimator $\hat{\beta}^o$ is biased due to the Laplacian shrinkage. Condition (B)(ii) requires a penalty

level λ_1 to prevent this bias and noise to cause false selection of variables in $\mathcal{O}^c$. Condition (B)(iii) requires that the nonzero coefficients not be too small in order for the SLS estimator to be able to distinguish nonzero from zero coefficients.

In Theorem 1, we only require $c_{\min}(\lambda_2) > 0$, or equivalently, $\Sigma + \lambda_2 L$ to be positive definite. The matrix Σ can be singular. This can be seen as follows. The adjacency matrix partitions the graph into disconnected cliques V_g, $1 \leq g \leq J$, for some $J \geq 1$. Let node j_g be a (representative) member of V_g. A node k belongs to the same clique V_g iff (if and only if) $a_{j_g k_1} a_{k_1 k_2} \cdots a_{k_m k} \neq 0$ through a certain chain $j_g \rightarrow k_1 \rightarrow k_2 \rightarrow \cdots \rightarrow k_m \rightarrow k$. Define $\bar{\mathbf{x}}_g = \sum_{k \in V_g} a_{j_g k_1} a_{k_1 k_2} \cdots a_{k_m k} \mathbf{x}_k / |V_g|$, where $|V_g|$ is the cardinality of V_g. The matrix $\Sigma + \lambda_2 L$ is positive definite iff $\mathbf{b}'\Sigma \mathbf{b} = \mathbf{b}'L\mathbf{b} = 0$ implies $\mathbf{b} = 0$. Since $\mathbf{b}'L\mathbf{b} = 0$ implies $\sum_{k \in V_g} b_k \mathbf{x}_k = b_{j_g}|V_g|\bar{\mathbf{x}}_g$, $\Sigma + \lambda_2 L$ is positive definite iff the vectors $\bar{\mathbf{x}}_g$ are linearly independent. This does not require $n \geq p$. In other words, Theorem 1 is applicable to $p > n$ problems as long as the vectors $\bar{\mathbf{x}}_g$ are linearly independent.

4.2. *The nonconvex case.* When $\Sigma(\lambda_2) = \Sigma + \lambda_2 L$ is singular, Theorem 1 is not applicable. In this case, further conditions are required for the oracle property to hold. The key condition needed is the sparse Reisz condition, or SRC [Zhang and Huang (2008)], in (4.9) below. It restricts the spectrum of diagonal subblocks of $\Sigma(\lambda_2)$ up to a certain dimension.

Let $\widetilde{X} = \widetilde{X}(\lambda_2)$ be a matrix satisfying $\widetilde{X}'\widetilde{X}/n = \Sigma(\lambda_2) = X'X/n + \lambda_2 L$ and $\tilde{\mathbf{y}} = \tilde{\mathbf{y}}(\lambda_2)$ be a vector satisfying $\widetilde{X}'\tilde{\mathbf{y}} = X'\mathbf{y}$. Define

$$(4.7) \qquad \widetilde{M}(\mathbf{b}; \lambda, \gamma) = \frac{1}{2n}\|\tilde{\mathbf{y}} - \widetilde{X}\mathbf{b}\|^2 + \sum_{j=1}^{p} \rho(|b_j|; \lambda_1, \gamma).$$

Since $M(\mathbf{b}; \lambda, \gamma) - \widetilde{M}(\mathbf{b}; \lambda, \gamma) = (\|\mathbf{y}\|^2 - \|\tilde{\mathbf{y}}\|^2)/(2n)$, the two penalized loss functions have the same set of local minimizers. For the penalized loss (4.7) with the data $(\widetilde{X}, \tilde{\mathbf{y}})$, let

$$(4.8) \qquad \hat{\boldsymbol{\beta}}(\lambda) = \boldsymbol{\delta}(\widetilde{X}(\lambda_2), \tilde{\mathbf{y}}(\lambda_2), \lambda_1),$$

where the map $\boldsymbol{\delta}(X, \mathbf{y}, \lambda_1) \in \mathbb{R}^p$ defines the MC+ estimator [Zhang (2010)] with data $(X, \mathbf{y})$ and penalty level λ_1. It was shown in Zhang (2010) that $\boldsymbol{\delta}(X, \mathbf{y}, \lambda_1)$ depends on $(X, \mathbf{y})$ only through $X'\mathbf{y}/n$ and $X'X/n$, so that different choices of $\widetilde{X}$ and $\tilde{\mathbf{y}}$ are allowed. One way is to pick $\tilde{\mathbf{y}} = (\mathbf{y}', 0)'$ and $\widetilde{X} = \text{diag}(X, (n\lambda_2 L)^{1/2})$. Another way is to pick $\widetilde{X}'\widetilde{X}/n = \Sigma(\lambda_2)$ and $\tilde{\mathbf{y}} = (\widetilde{X}')^{\dagger}X'\mathbf{y}$ of smaller dimensions, where $(\widetilde{X}')^{\dagger}$ is the Moore–Penrose inverse of $\widetilde{X}'$.

CONDITION (C). (i) For an integer d^* and spectrum bounds $0 < c_*(\lambda_2) \leq c^*(\lambda_2) < \infty$,

$$0 < c_*(\lambda_2) \leq \mathbf{u}_B' \Sigma_B(\lambda_2) \mathbf{u}_B \leq c^*(\lambda_2) < \infty$$

$$(4.9)$$

$$\forall B \text{ with } |B \cup \mathcal{O}| \leq d^*, \|\mathbf{u}_B\| = 1,$$

with $d^* \geq d^o(K_* + 1)$, $\gamma \geq c_*^{-1}(\lambda_2)\sqrt{4 + c_*(\lambda_2)/c^*(\lambda_2)}$ in (2.2), and $K_* = c^*(\lambda_2)/c_*(\lambda_2) - 1/2$.

(ii) With $C_2 = \|\{\Sigma_{B,\mathcal{O}}(\lambda_2)\Sigma_{\mathcal{O}}^{-1}(\lambda_2)L_{\mathcal{O}} - L_{B,\mathcal{O}}\}\boldsymbol{\beta}_{\mathcal{O}}^o\|_\infty$,

$$\max\{1, \sqrt{c_*(\lambda_2)K_*/(K_* + 1)}\}\lambda_1 \geq \lambda_2 C_2 + \sigma\sqrt{2\log(p/\epsilon)}\max_{j \leq p}\|\mathbf{x}_j\|/n.$$

(iii) With $\{v_j, j \in \mathcal{O}\}$ being the diagonal elements of $\Sigma_{\mathcal{O}}^{-1}(\lambda_2)\Sigma_{\mathcal{O}}\{\Sigma_{\mathcal{O}}^{-1}(\lambda_2)\}$,

$$\min_{j \in \mathcal{O}}\{|\beta_j^*| - \gamma(2\sqrt{c^*(\lambda_2)}\lambda_1)\}(n/v_j)^{1/2} \geq \sigma\sqrt{2\log(d^o/\epsilon)}.$$

THEOREM 2. (i) *Suppose Conditions* (A) *and* (C) *hold. Let* $\hat{\boldsymbol{\beta}}(\lambda)$ *be as in* (4.8). *Then*

$$(4.10) \qquad P(\{j : \hat{\beta}_j \neq 0\} \neq \mathcal{O} \text{ or } \hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^o) \leq 3\epsilon.$$

If $\beta_* \geq \lambda_2 C_1 + \gamma(2\sqrt{c^*(\lambda_2)}\lambda_1) + \max_j\sqrt{(2v_j/n)\log(d^o/\epsilon)}$ *instead of Condition* (C)(iii), *then*

$$(4.11) \qquad P(\text{sgn}(\hat{\boldsymbol{\beta}}) \neq \text{sgn}(\boldsymbol{\beta}^o) \text{ or } \hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^o) \leq 3\epsilon.$$

Here note that p, γ, d^o, d^*, K_*, ϵ, $c_*(\lambda_2)$ *and* $c^*(\lambda_2)$ *are all allowed to depend on* n, *including the case* $c_*(\lambda_2) \to 0$ *as long as the conditions hold as stated.*

(ii) *The statements in* (i) *also hold for all local minimizers* $\hat{\beta}$ *of* (2.6) *or* (4.7) *satisfying* $\#\{j \notin \mathcal{O} : \hat{\beta}_j \neq 0\} + d^o \leq d^*$.

If the conditions of Theorem 2 hold with $\epsilon \to 0$, then (4.10) implies selection consistency of the SLS estimator and (4.11) implies sign consistency.

Condition (C), designed to handle the noncovexity of the penalized loss, is a weaker version of Condition (B) in the sense of allowing singular $\Sigma(\lambda_2)$. The SRC (4.9), depending on X or $\widetilde{X}$ only through the regularized Gram matrix $\widetilde{X}'\widetilde{X}/n = \Sigma(\lambda_2) = \Sigma + \lambda_2 L$, ensures that the model is identifiable in a lower d^*-dimensional space. When $p > n$, the smallest singular value of X is always zero. However, the requirement $c_*(\lambda_2) > 0$ only concerns $d^* \times d^*$ diagonal submatrices of $\Sigma(\lambda_2)$, not the Gram matrix Σ of the design matrix X. We can have $p \gg n$ but still require $d^*/d^o \geq K_* + 1$ as in (4.9). Since p, d^0, γ, d^*, K_*, $c_*(\lambda_2)$ and $c^*(\lambda_2)$ can depend on n, we allow the case $c_*(\lambda_2) \to 0$ as long as Conditions (A) and (C) hold as stated. Thus, we allow $p \gg n$ but require that the model is sparse, in the sense that the number of nonzero coefficients d^o is smaller than $d^*/(1 + K_*)$. For example, if $c_*(\lambda_2) \asymp O(n^{-\alpha})$ for a small $\alpha > 0$ and $c^*(\lambda_2) \asymp O(1)$, then we require $\gamma \asymp O(n^{3\alpha/2})$ or greater, $K^* \asymp O(n^\alpha)$ and $d^*/d^o \asymp O(n^\alpha)$ or greater. So all these quantities can depend on n, as long as the other requirements are met in Condition (C).

By examining the Conditions (C)(ii) and (C)(iii), for standardized predictors with $\|\mathbf{x}_j\| = \sqrt{n}$, we can have $\log(p/\epsilon) = o(n)$ or $p = \epsilon\exp(o(n))$ as long as

Condition (C)(ii) is satisfied. As in Zhang (2010), under a somewhat stronger version of Condition (C), Theorem 2 can be extended to quadratic spline concave penalties satisfying $\rho(t; \lambda_1, \gamma) = \lambda_1^2 \rho(t/\lambda; \gamma)$ with a penalty function satisfying $(\partial/\partial t)\rho(t; \gamma) = 1$ at $t = 0+$ and 0 for $t > \gamma$.

Also, comparing our results with the selection consistency results of Hebiri and van de Geer (2010) on the smoothed $\ell_1 + \ell_2$-penalized methods, our conditions tend to be weaker. Notably, Hebiri and van de Geer (2010) require an condition on the Gram matrix which assumes that the correlations between the truly relevant variables and those which are not are small. No such assumption is required for our selection consistency results. In addition, our selection results are stronger in the sense that the SLS estimator is not only sign consistent, but also equal to the oracle Laplacian shrinkage estimator with high probability. In general, similar results are not available with the use of the ℓ_1 penalty for sparsity.

Theorem 2 shows that the SLS estimator automatically adapts to the sparseness of the p-dimensional model and the denseness of a *true* submodel. From a sparse p-model, it correctly selects the true underlying model $\mathcal{O}$. This underlying model is a dense model in the sense that all its coefficients are nonzero. In this dense model, the SLS estimator behaves like the oracle Laplacian shrinkage estimator in (4.1). As in the convex penalized loss setting, here the results do not require a correct specification of a population correlation structure of the predictors.

4.3. *Unbiased Laplacian and variance reduction.* There are two natural questions concerning the SLS. First, what are the benefits from introducing the Laplacian penalty? Second, what kind of Laplacian L constitutes a reasonable choice? Since the SLS estimator is equal to the oracle Laplacian estimator with high probability by Theorem 1 or 2, these questions can be answered by examining the oracle Laplacian shrinkage estimator (4.1), whose nonzero part is

$$\hat{\boldsymbol{\beta}}_{\mathcal{O}}^{o}(\lambda_2) = \Sigma_{\mathcal{O}}^{-1}(\lambda_2) X_{\mathcal{O}}' \mathbf{y}/n.$$

Without the Laplacian, that is, when $\lambda_2 = 0$, it becomes the least squares (LS) estimator

$$\hat{\boldsymbol{\beta}}_{\mathcal{O}}^{o}(0) = \Sigma_{\mathcal{O}}^{-1} X_{\mathcal{O}}' \mathbf{y}/n.$$

If some of the predictors in $\{\mathbf{x}_j, j \in \mathcal{O}\}$ are highly correlated or $|\mathcal{O}| \geq n$, the LS estimator $\hat{\boldsymbol{\beta}}_{\mathcal{O}}^{o}(0)$ is not stable or unique. In comparison, as discussed below Theorem 1, $\Sigma_{\mathcal{O}}(\lambda_2) = \Sigma_{\mathcal{O}} + \lambda_2 L_{\mathcal{O}}$ can be a full rank matrix under a reasonable condition, even if the predictors in $\{\mathbf{x}_j, j \in \mathcal{O}\}$ are highly correlated or $|\mathcal{O}| \geq n$.

For the second question, we examine the bias of $\hat{\boldsymbol{\beta}}_{\mathcal{O}}^{o}(\lambda_2)$. Since the bias of the target vector (4.4) is $\boldsymbol{\beta}_{\mathcal{O}}^{o} - \boldsymbol{\beta}_{\mathcal{O}}^{*}(\lambda_2) = \lambda_2 \Sigma_{\mathcal{O}}^{-1}(\lambda_2) L_{\mathcal{O}} \boldsymbol{\beta}_{\mathcal{O}}^{o}$, $\hat{\boldsymbol{\beta}}_{\mathcal{O}}^{o}(\lambda_2)$ is unbiased iff $L_{\mathcal{O}} \boldsymbol{\beta}_{\mathcal{O}}^{o} = 0$. Therefore, in terms of bias reduction, a Laplacian L is most appropriate if the condition $L_{\mathcal{O}} \boldsymbol{\beta}_{\mathcal{O}}^{o} = 0$ is satisfied. We shall say that a Laplacian L is unbiased if $L_{\mathcal{O}} \boldsymbol{\beta}_{\mathcal{O}}^{o} = 0$. It follows from the discussion at the end of Section 4.1 that

$L_{\mathcal{O}}\boldsymbol{\beta}_{\mathcal{O}}^o = 0$ if $\beta_k^o = \beta_{j_g}^o a_{j_g k_1} a_{k_1 k_2} \cdots a_{k_m k}$, where j_g is a representative member of the clique $V_g \cap \mathcal{O}$ and $\{k_1, \ldots, k_m, k\} \subseteq V_g \cap \mathcal{O}$.

With an unbiased Laplacian, the mean square error of $\hat{\boldsymbol{\beta}}_{\mathcal{O}}^o(\lambda_2)$ is

$$E\|\hat{\boldsymbol{\beta}}_{\mathcal{O}}^o(\lambda_2) - \boldsymbol{\beta}_{\mathcal{O}}^o\|^2 = \frac{\sigma^2}{n}\operatorname{trace}(\Sigma_{\mathcal{O}}^{-1}(\lambda_2)\Sigma_{\mathcal{O}}\Sigma_{\mathcal{O}}^{-1}(\lambda_2)).$$

The mean square error of $\hat{\boldsymbol{\beta}}_{\mathcal{O}}(0)$ is

$$E\|\hat{\boldsymbol{\beta}}_{\mathcal{O}}^o(0) - \boldsymbol{\beta}_{\mathcal{O}}^o\|^2 = \frac{\sigma^2}{n}\operatorname{trace}(\Sigma_{\mathcal{O}}^{-1}).$$

We always have $E\|\hat{\boldsymbol{\beta}}_{\mathcal{O}}^o(\lambda_2) - \boldsymbol{\beta}_{\mathcal{O}}^o\|^2 < E\|\hat{\boldsymbol{\beta}}_{\mathcal{O}}^o(0) - \boldsymbol{\beta}_{\mathcal{O}}^o\|^2$ for $\lambda_2 > 0$. Therefore, an unbiased Laplacian reduces variance without incurring any bias on the estimator.

5. Laplacian shrinkage. The results in Section 4 show that the SLS estimator is equal to the oracle Laplacian shrinkage estimator with probability tending to one under certain conditions. In addition, an unbiased Laplacian reduces variance but does not increase bias. Therefore, to study the shrinkage effect of the Laplacian penalty on $\hat{\boldsymbol{\beta}}$, we can consider the oracle estimator $\hat{\boldsymbol{\beta}}_{\mathcal{O}}^o$. To simplify the notation and without causing confusion, in this section, we study some other basic properties of the Laplacian shrinkage and compare it with the ridge shrinkage. The Laplacian shrinkage estimator is defined as

$$(5.1) \quad \tilde{\boldsymbol{\beta}}(\lambda_2) = \arg\min_{\mathbf{b}}\left\{G(\mathbf{b};\lambda_2) \equiv \frac{1}{2n}\|\mathbf{y} - X\mathbf{b}\|^2 + \frac{1}{2}\lambda_2 \mathbf{b}'L\mathbf{b}, \mathbf{b} \in \mathbb{R}^q\right\}.$$

The following proposition shows that the Laplacian penalty shrinks a coefficient toward the center of all the coefficients connected to it.

PROPOSITION 1. *Let* $\tilde{\mathbf{r}} = \mathbf{y} - X\tilde{\boldsymbol{\beta}}$.
(i)

$$\lambda_2 \max_{1 \le j \le q} d_j|\tilde{\beta}_j - \mathbf{a}_j'\tilde{\boldsymbol{\beta}}/d_j| \le \|\tilde{\mathbf{r}}\| \le \|\mathbf{y}\|.$$

(ii)

$$\lambda_2|d_j\tilde{\beta}_j - \mathbf{a}_j'\tilde{\boldsymbol{\beta}} - (d_k\tilde{\beta}_k - \mathbf{a}_k'\tilde{\boldsymbol{\beta}})| \le \frac{1}{n}\|\mathbf{x}_j - \mathbf{x}_k\|\|\mathbf{y}\|.$$

Note that $\mathbf{a}_j'\tilde{\boldsymbol{\beta}}/d_j = \sum_{k=1}^q a_{jk}\tilde{\beta}_k/d_j = \sum_{k=1}^q \operatorname{sgn}(a_{jk})|a_{jk}|\tilde{\beta}_k/d_j$ is a signed weighted average of the $\tilde{\beta}_k$'s connected to $\tilde{\beta}_j$, since $d_j = \sum_k |a_{jk}|$. Part (i) of Proposition 1 provides an upper bound on the difference between $\tilde{\beta}_j$ and the center of all the coefficients connected to it. When $\|\tilde{\mathbf{r}}\|/(\lambda_2 d_j) \to 0$, this difference converges to zero. For standardized $d_j = 1$, part (ii) implies that the difference between the centered $\tilde{\beta}_j$ and $\tilde{\beta}_k$ converges to zero when $\|\mathbf{x}_j - \mathbf{x}_k\|\|\mathbf{y}\|/(\lambda_2 n) \to 0$.

When there are certain local structures in the adjacency matrix A, shrinkage occurs at the local level. As an example, we consider the adjacency matrix based on partition of the predictors into $2r$-balls defined in (3.1). Correspondingly, the index set $\{1, \ldots, q\}$ is divided into disjoint neighborhoods/cliques $V_1, \ldots, V_J$. We consider the normalized Laplacian $L = I_q - A$, where I_q is a $q \times q$ identity matrix and $A = \text{diag}(A_1, \ldots, A_J)$ with $A_g = v_g^{-1}\mathbf{1}'_g\mathbf{1}$. Here $v_g = |V_g|, 1 \le g \le J$. Let $\mathbf{b}_g = (b_j, j \in V_g)'$. We can write the objective function as

$$(5.2) \qquad G(\mathbf{b}; \lambda_2) = \frac{1}{2n}\|\mathbf{y} - X\mathbf{b}\|^2 + \frac{1}{2}\lambda_2 \sum_{g=1}^{J} \mathbf{b}'_g(\mathbf{I}_g - v_g^{-1}\mathbf{1}'_g\mathbf{1}_g)\mathbf{b}_g.$$

For the Laplacian shrinkage estimator based on this criterion, we have the following grouping properties.

PROPOSITION 2.   (i) *For any* $j, k \in V_g, 1 \le g \le J$,

$$\lambda_2|\tilde{\beta}_j - \tilde{\beta}_k| \le \frac{1}{n}\|\mathbf{x}_j - \mathbf{x}_k\| \cdot \|\mathbf{y}\|, \qquad j, k \in V_g.$$

(ii) *Let* $\bar{\beta}_g$ *be the average of the estimates in* V_g. *For any* $j \in V_g$ *and* $k \in V_h$, $g \ne h$,

$$\lambda_2|\tilde{\beta}_j - \bar{\beta}_g - (\tilde{\beta}_k - \bar{\beta}_h)| \le \frac{1}{n}\|\mathbf{x}_j - \mathbf{x}_k\| \cdot \|\mathbf{y}\|, \qquad j \in V_g, k \in V_h.$$

This proposition characterizes the smoothing effect and grouping property of the Laplacian penalty in (5.2). Consider the case $\|\mathbf{y}\|^2/n = O(1)$. Part (i) implies that, for j and k in the same neighborhood and $\lambda_2 > 0$, the difference $\tilde{\beta}_j - \tilde{\beta}_k \to 0$ if $\|\mathbf{x}_j - \mathbf{x}_k\|/(\lambda_2 n^{1/2}) \to 0$. Part (ii) implies that, for j and k in different neighborhoods and $\lambda_2 > 0$, the difference between the centered $\tilde{\beta}_j$ and $\tilde{\beta}_k$ converges to zero if $\|\mathbf{x}_j - \mathbf{x}_k\|/(\lambda_2 n^{1/2}) \to 0$.

We now compare the Laplacian shrinkage and ridge shrinkage. The discussion at the end of Section 4 about the requirement for the unbiasedness of Laplacian can be put in a wider context when a general positive definite or semidefinite matrix Q is used in the place of L. This wider context includes the Laplacian shrinkage and ridge shrinkage as special cases. Specifically, let

$$\hat{\boldsymbol{\beta}}_Q(\lambda, \gamma) = \arg\min_{\mathbf{b}} \frac{1}{2n}\|\mathbf{y} - X\mathbf{b}\|^2 + \sum_{j=1}^{p} \rho(|b_j|; \lambda_1, \gamma) + \frac{1}{2}\lambda_2\mathbf{b}'Q\mathbf{b}.$$

For $Q = I_p$, $\hat{\boldsymbol{\beta}}_Q$ becomes the Mnet estimator [Huang et al. (2010a)]. With some modifications on the conditions in Theorem 1 or Theorem 2, it can be shown that $\hat{\boldsymbol{\beta}}_Q$ is equal to the oracle estimator defined as

$$\hat{\boldsymbol{\beta}}_Q^o(\lambda_2) = \arg\min_{\mathbf{b}} \left\{ \frac{1}{2n}\|\mathbf{y} - X\mathbf{b}\|^2 + \frac{1}{2}\mathbf{b}'Q\mathbf{b}, b_j = 0, j \notin \mathcal{O} \right\}.$$

Then in a way similar to the discussion in Section 4, $\hat{\boldsymbol{\beta}}_Q$ is nearly unbiased iff $Q_\mathcal{O}\boldsymbol{\beta}_\mathcal{O}^o = 0$. Therefore, for $\|\boldsymbol{\beta}_\mathcal{O}^o\| \neq 0$, $Q_\mathcal{O}$ must be a rank deficient matrix, which in turn implies that Q must be rank deficient. Note that any Laplacian L is rank deficient. This rank deficiency requirement excludes the ridge penalty with $Q = I_p$. For the ridge penalty to yield an unbiased estimator, it must hold that $\|\boldsymbol{\beta}^o\| = 0$ in the underlying model.

We now give a simple example that illustrates the basic characteristics of Laplacian shrinkage and its differences from ridge shrinkage.

EXAMPLE 5.1. Consider a linear regression model with two predictors satisfying $\|\mathbf{x}_j\|^2 = n$, $j = 1, 2$. The Laplacian shrinkage and ridge estimators are defined as

$$(\hat{b}_{L1}(\lambda_2), \hat{b}_{L2}(\lambda_2)) = \arg\min_{b_1, b_2} \frac{1}{2n} \sum_{i=1}^n (y_i - x_{i1}b_1 - x_{i2}b_2)^2 + \frac{1}{2}\lambda_2(b_1 - b_2)^2$$

and

$$(\hat{b}_{R1}(\lambda_2), \hat{b}_{R2}(\lambda_2)) = \arg\min_{b1, b_2} \frac{1}{2n} \sum_{i=1}^n (y_i - x_{i1}b_1 - x_{i2}b_2)^2 + \frac{1}{2}\lambda_2(b_1^2 + b_2^2).$$

Denote $r_1 = \mathrm{cor}(\mathbf{x}_1, \mathbf{y})$, $r_2 = \mathrm{cor}(\mathbf{x}_2, \mathbf{y})$ and $r_{12} = \mathrm{cor}(\mathbf{x}_1, \mathbf{x}_2)$. The Laplacian shrinkage estimates are

$$\hat{b}_{L1}(\lambda_2) = \frac{(1 + \lambda_2)r_1 - (r_{12} - \lambda_2)r_2}{(1 + \lambda_2)^2 - (r_{12} - \lambda_2)^2}, \qquad \hat{b}_{L2}(\lambda_2) = \frac{(1 + \lambda_2)r_2 - (r_{12} - \lambda_2)r_1}{(1 + \lambda_2)^2 - (r_{12} - \lambda_2)^2}.$$

Let

$$\hat{b}_{\mathrm{ols}1} = \frac{r_1 - r_{12}r_2}{1 - r_{12}^2}, \qquad \hat{b}_{\mathrm{ols}2} = \frac{r_2 - r_{12}r_1}{1 - r_{12}^2}, \qquad \hat{b}_L(\infty) = \frac{r_1 + r_2}{2(1 + r_{12})},$$

where $(\hat{b}_{\mathrm{ols}1}, \hat{b}_{\mathrm{ols}2})$ is the ordinary least squares (OLS) estimator for the bivariate regression, $\hat{b}_L(\infty)$ is the OLS estimator that assumes the two coefficients are equal, that is, it minimizes $\sum_{i=1}^n (y_i - (x_{i1} + x_{i2})b)^2$. Let $w_L = (2\lambda_2)/(1 - r_{12} + 2\lambda_2)$. After some simple algebra, we have

$$\hat{b}_{L1}(\lambda_2) = (1 - w_L)\hat{b}_{\mathrm{ols}1} + w_L\hat{b}_L(\infty)$$

and

$$\hat{b}_{L2}(\lambda_2) = (1 - w_L)\hat{b}_{\mathrm{ols}2} + w_L\hat{b}_L(\infty).$$

Thus, for any fixed λ_2, $\hat{b}_L(\lambda_2)$ is a weighted average of $\hat{b}_{\mathrm{ols}}$ and $\hat{b}_L(\infty)$ with the weights depending on λ_2. When $\lambda_2 \to \infty$, $\hat{b}_{L1} \to \hat{b}_L(\infty)$ and $\hat{b}_{L2} \to \hat{b}_L(\infty)$. Therefore, the Laplacian penalty shrinks the OLS estimates toward a common value, which is the OLS estimate assuming equal regression coefficients.

Now consider the ridge regression estimator. We have

$$\hat{b}_{R1}(\lambda_2) = \frac{(1+\lambda_2)r_1 - r_{12}r_2}{(1+\lambda_2)^2 - r_{12}^2} \quad \text{and} \quad \hat{b}_{R2}(\lambda_2) = \frac{(1+\lambda_2)r_2 - r_{12}r_1}{(1+\lambda_2)^2 - r_{12}^2}.$$

The ridge estimator converges to zero as $\lambda_2 \to \infty$. For it to converge to a nontrivial solution, we need to rescale it by a factor of $1 + \lambda_2$. Let $w_R = \lambda/(1+\lambda - r_{12}^2)$. Let $\hat{b}_{u1} = r_1$ and $\hat{b}_{u2} = r_2$. Because $n^{-1}\sum_{i=1}^{n} x_{i1}^2 = 1$ and $n^{-1}\sum_{i=1}^{n} x_{i2}^2 = 1$, r_1 and r_2 are also the OLS estimators of univariate regressions of $\mathbf{y}$ on $\mathbf{x}_1$ and $\mathbf{y}$ on $\mathbf{x}_2$, respectively. We can write

$$(1+\lambda_2)\hat{b}_{R1}(\lambda_2) = c_{\lambda_2}(1-w_R)\hat{b}_{\text{ols}1} + c_\lambda w_R \hat{b}_{u1},$$

$$(1+\lambda_2)\hat{b}_{R2}(\lambda_2) = c_{\lambda_2}(1-w_R)\hat{b}_{\text{ols}2} + c_\lambda w_R \hat{b}_{u2},$$

where $c_{\lambda_2} = \{(1+\lambda_2)^2 - (1+\lambda)r_{12}^2\}/\{(1+\lambda_2)^2 - r_{12}^2\}$. Note that $c_{\lambda_2} \approx 1$. Thus, $(1+\lambda_2)\hat{b}_R$ is a weighted average of the OLS and the univariate regression estimators. The ridge penalty shrinks the (rescaled) ridge estimates toward individual univariate regression estimates.

6. Simulation studies. We use a coordinate descent algorithm to compute the SLS estimate. This algorithm optimizes a target function with respect to a single parameter at a time and iteratively cycles through all parameters until convergence. This algorithm was originally proposed for criterions with convex penalties such as LASSO [Fu (1998), Genkin, Lewis and Madigan (2004), Friedman et al. (2007), Wu and Lange (2008)]. It has been proposed to calculate the MCP estimates [Breheny and Huang (2011)]. Detailed steps of this algorithm for computing the SLS estimates can be found in the technical report accompanying this paper [Huang et al. (2010b)].

In simulation studies, we consider the following ways of defining the adjacency measure. (N.1) $a_{jk} = I(r_{jk} > r)$ and $s_{jk} = 1$. Here the cutoff r is computed as 3.09 using the approach described in Section 3 with a p-value of 10^{-3}; (N.2) $a_{jk} = I(|r_{jk}| > r)$ and $s_{jk} = \text{sgn}(r_{jk})$. Here the cutoff r is computed as 3.29 using the approach described in Section 3 with a p-value of 10^{-3}; (N.3) $a_{jk} = \max(0, r_{jk})^\alpha$ and $s_{jk} = 1$. We set $\alpha = 6$, which satisfies the scale-free topology criteria [Zhang and Horvath (2005)]; (N.4) $a_{jk} = r_{jk}^\alpha$ and $s_{jk} = \text{sgn}(r_{jk})$. We set $\alpha = 6$.

The penalty levels λ_1 and λ_2 are selected using V-fold cross validation. In our numerical study, we set $V = 5$. To reduce computational cost, we search over the discrete grid of $2^{\cdots, -1, -0.5, 0, 0.5, \cdots}$. For comparison, we also consider the MCP estimate and the approach proposed in Daye and Jeng (2009); referred to as D–J hereafter. Both the SLS and MCP involve the regularization parameter γ. For MCP, Zhang (2010) suggested using $\gamma = 2/(1 - \max_{j \neq k} |x_j' x_k|/n)$ for standardized covariates. The average γ value of this choice is 2.69 in his simulation studies. The

simulation studies in Breheny and Huang (2011) suggest that $\gamma = 3$ is a reasonable choice. We have experimented with different γ values and reached the same conclusion. Therefore, we set $\gamma = 3$.

We set $n = 100$ and $p = 500$. Among the 500 covariates, there are 100 clusters, each with size 5. We consider two different correlation structures. (I) Covariates in different clusters are independent, whereas covariates i and j within the same cluster have correlation coefficients $\rho^{|i-j|}$; and (II) covariates i and j have correlation coefficients $\rho^{|i-j|}$. Under structure I, zero and nonzero effects are independent, whereas under structure II, they are correlated. Covariates have marginal normal distributions with mean zero and variance one. We consider different levels of correlation with $\rho = 0.1, 0.5, 0.9$. Among the 500 covariates, the first 25 (5 clusters) have nonzero regression coefficients. We consider the following scenarios for nonzero coefficients: (a) all the nonzero coefficients are equal to 0.5; and (b) the nonzero coefficients are randomly generated from the uniform distribution on [0.25, 0.75]. In (a), the Laplacian matrices satisfy the unbiasedness property $L\beta^o = 0$ discussed in Section 4. We have experienced with other levels of nonzero regression coefficients and reached similar conclusions.

We examine the accuracy of identifying nonzero covariate effects and the prediction performance. For this purpose, for each simulated dataset, we simulate an independent testing dataset with sample size 100. We conduct cross validation (for tuning parameter selection) and estimation using the training set only. We then make prediction for subjects in the testing set and compute the PMSE (prediction mean squared error).

We simulate 500 replicates and present the summary statistics in Table 1. We can see that the MCP performs satisfactorily when the correlation is small. However, when the correlation is high, it may miss a considerable number of true positives and have large prediction errors. The D–J approach, which can also accommodate the correlation structure, is able to identify all the true positives. However, it also identifies a large number of false positives, causing by the over-selection of the Lasso penalty. The proposed SLS approach outperforms the MCP and D–J methods in the sense that it has smaller empirical false discovery rates with comparable false negative rates. It also has significantly smaller prediction errors.

6.1. *Application to a microarray study.* In the study reported in Scheetz et al. (2006), F1 animals were intercrossed and 120 twelve-week-old male offspring were selected for tissue harvesting from the eyes and microarray analysis using the Affymetric GeneChip Rat Genome 230 2.0 Array. The intensity values were normalized using the RMA [robust multi-chip averaging, Bolstad et al. (2003), Irizarry et al. (2003)] method to obtain summary expression values for each probe set. Gene expression levels were analyzed on a logarithmic scale. For the probe sets on the array, we first excluded those that were not expressed in the eye or that lacked sufficient variation. The definition of expressed was based on the empirical distribution of RMA normalized values. For a probe set to be considered expressed,

TABLE 1

Simulation study: *median based on* 500 *replicates. In each cell, the three numbers are positive findings, true positives and PMSE* ×100, *respectively*

Coefficient	ρ	MCP	D–J N.1	N.2	N.3	N.4	SLS N.1	N.2	N.3	N.4
						Correlation structure I				
0.5	0.1	27 25 41.33	61 25 125.34	53 25 46.64	55 25 60.14	59 25 51.24	27 25 40.53	27 25 39.84	26 25 41.74	27 25 39.34
	0.5	28 25 54.10	51 25 66.38	67 25 66.84	72 25 56.22	63 25 53.43	27 25 37.71	28 25 39.18	28 25 33.87	27 25 36.00
	0.9	22 15 137.52	66 25 55.51	55 25 56.94	61 25 49.22	74 25 51.41	29 25 48.89	28 25 49.96	29 25 45.16	27 25 41.49
U[0.25,	0.1	37 25 52.24	72 25 54.28	61 25 88.00	59 25 70.00	78 25 60.51	33 25 51.80	36 25 52.19	30 25 53.03	30 25 52.22
0.75]	0.5	29 24 65.12	66 25 78.76	54 25 72.34	63 25 63.55	57 25 66.33	28 25 42.24	28 25 43.96	27 24 54.72	28 24 58.77
	0.9	17 13 152.42	67 25 63.43	62 25 57.30	50 25 53.88	74 25 57.98	29 25 47.73	29 25 49.14	27 25 48.49	28 25 50.83
						Correlation structure II				
0.5	0.1	26 25 38.22	62 25 121.69	58 25 117.10	63 25 127.34	72 25 122.34	27 25 40.33	27 25 40.65	27 25 41.49	27 25 37.40
	0.5	29 25 53.01	52 25 55.99	49 25 62.04	66 25 62.70	65 25 64.41	27 25 36.97	28 25 39.47	28 25 38.53	27 25 39.53
	0.9	15 13 140.69	48 25 55.75	34 25 56.71	32 25 60.27	38 25 59.78	29 25 66.79	29 25 60.52	29 25 57.91	30 25 60.19
U[0.25,	0.1	37 25 54.31	77 25 60.02	72 25 66.14	74 25 78.32	66 25 74.50	29 25 50.05	32 25 51.34	37 25 50.74	29 25 49.47
0.75]	0.5	27 24 57.66	74 25 61.71	66 25 67.54	75 25 62.01	74 25 66.91	28 25 44.92	28 25 46.65	28 25 41.35	28 25 41.17
	0.9	14 13 136.49	33 25 61.50	35 25 55.08	34 25 54.54	38 25 60.67	29 25 56.87	29 25 57.03	30 25 53.28	30 25 56.79

the maximum expression value observed for that probe among the 120 F2 rats was required to be greater than the 25th percentile of the entire set of RMA expression values. For a probe to be considered "sufficiently variable," it had to exhibit at least 2-fold variation in expression level among the 120 F2 animals.

We are interested in finding the genes whose expression are most variable and correlated with that of gene TRIM32. This gene was recently found to cause Bardet–Biedl syndrome [Chiang et al. (2006)], which is a genetically heterogeneous disease of multiple organ systems including the retina. One approach to find the genes related to TRIM32 is to use regression analysis. Since it is expected that the number of genes associated with gene TRIM32 is small and since we are mainly interested in genes whose expression values across samples are most variable, we conduct the following initial screening. We compute the variances of gene expressions and select the top 1,000. We then standardize gene expressions to have zero mean and unit variance.

We analyze data using the MCP, D–J, and proposed approach. In cross validation, we set $V = 5$. The numbers of genes identified are MCP: 23, D–J: 31 (N.1), 41 (N.2), 34 (N.3), 30 (N.4), SLS: 25 (N.1), 26 (N.2), 16 (N.3) and 17 (N.4), respectively. More detailed results are available from the authors. Different approaches and different ways of defining the adjacency measure lead to the identification of different genes. As expected, the SLS identifies shorter lists of genes than the D–J, which may lead to more parsimonious models and more focused hypothesis for confirmation. As the proposed approach pays special attention to the correlation among genes, we also compute the median of the absolute values of correlations among the identified genes, which are MCP: 0.171, D–J: 0.201 (N.1), 0.207 (N.2), 0.215 (N.3), 0.206 (N.4), SLS: 0.247 (N.1), 0.208 (N.2), 0.228 (N.3), 0.212 (N.4). The D–J and SLS, which incorporate correlation in the penalty, identify genes that are more strongly correlated than the MCP. The SLS identified genes have slightly higher correlations than those identified by D–J.

Unlike in simulation study, we are not able to evaluate true and false positives. This limitation is shared by most existing studies. We use the following V-fold ($V = 5$) cross validation based approach to evaluate prediction. (a) Randomly split data into V-subsets with equal sizes; (b) Remove one subset from data; (c) Conduct cross validation and estimation using the rest $V - 1$ subsets; (d) Make prediction for the one removed subset; (e) Repeat Steps (b)–(d) over all subsets and compute the prediction error. The sums of squared prediction errors are MCP: 1.876; D–J: 1.951 (N.1), 1.694 (N.2), 1.534 (N.3) and 1.528 (N.4); SLS: 1.842 (N.1), 1.687 (N.2), 1.378 (N.3) and 1.441 (N.4), respectively. The SLS has smaller cross validated prediction errors, which may indirectly suggest better selection properties.

7. Discussion. In this article, we propose the SLS method for variable selection and estimation in high-dimensional data analysis. The most important feature of the SLS is that it explicitly incorporates the graph/network structure in predictors into the variable selection procedure through the Laplacian quadratic. It

provides a systematic framework for connecting penalized methods for consistent variable selection and those for network and correlation analysis. As can be seen from the methodological development, the application of the SLS variable selection is relatively independent of the graph/network construction. Thus, although graph/network construction is of significant importance, it is not the focus of this study and not thoroughly pursued.

An important feature of the SLS method is that it incorporates the correlation patterns of the predictors into variable selection through the Laplacian quadratic. We have considered two simple approaches for determining the Laplacian based on dissimilarity and similarity measures. Our simulation studies demonstrate that incorporating correlation patterns improves selection results and prediction performance. Our theoretical results on the selection properties of the SLS are applicable to a general class of Laplacians and do not require the underlying graph for the predictors to be correctly specified.

We provide sufficient conditions under which the SLS estimator possesses an oracle property, meaning that it is sign consistent and equal to the oracle Laplacian shrinkage estimator with high probability. We also study the grouping properties of the SLS estimator. Our results show that the SLS is adaptive to the sparseness of the original p-dimensional model with $p \gg n$ and the denseness of the underlying d^o-dimensional model, where $d^o < n$ is the number of nonzero coefficients. The asymptotic rates of the penalty parameters are derived. However, as in many recent studies, it is not clear whether the penalty parameters selected using cross validation or other procedures can match the asymptotic rate. This is an important and challenging problem that requires further investigation, but is beyond the scope of the current paper. Our numerical study shows a satisfactory finite-sample performance of the SLS. Particularly, we note that the cross validation selected tuning parameters seem sufficient for our simulated data. We are only able to experiment with four different adjacency measures. It is not our intention to draw conclusions on different ways of defining adjacency. More adjacency measures are hence not explored.

We have focused on the linear regression model in this article. However, the SLS method can be applied to general linear regression models. Specifically, for general linear models, the SLS criterion can be formulated as

$$\frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, b_0 + \sum_j x_{ij} b_j\right) + \sum_{j=1}^{p} \rho(|b_j|; \lambda_1, \gamma)$$
$$+ \frac{1}{2} \lambda_2 \sum_{1 \le j < k \le p} |a_{jk}| (b_j - s_{jk} b_k)^2,$$

where ℓ is a given loss function. For instance, for generalized linear models such as logistic regression, we can take ℓ to be the negative log-likelihood function. For Cox regression, we can use the negative partial likelihood as the loss function.

Computationally, for loss functions other than least squares, the coordinate descent algorithm can be applied iteratively to quadratic approximations to the loss function. However, further work is needed to study theoretical properties of the SLS estimators for general linear models.

There is a large literature on the analysis of network data and much work has also been done on estimating sparse covariance matrices in high-dimensional settings. See, for example, Zhang and Horvath (2005), Chung and Lu (2006), Meinshausen and Bühlmann (2006), Yuan and Lin (2007), Friedman, Hastie and Tibshirani (2008), Fan, Feng and Wu (2009), among others. It would be useful to study ways to incorporate these methods and results into the proposed SLS approach. In some problems such as genomic data analysis, partial external information may also be available on the graphical structure of some genes used as predictors in the model. It would be interesting to consider approaches for combining external information on the graphical structure with existing data in constructing the Laplacian quadratic penalty.

APPENDIX

In this Appendix, we give proofs of Theorems 1 and 2 and Propositions 1 and 2.

PROOF OF THEOREM 1. Since $c_{\min}(\lambda_2) > 1/\gamma$, the criterion (2.2) is strictly convex and its minimizer is unique. Let $\widetilde{X} = \widetilde{X}(\lambda_2) = \sqrt{n}(\Sigma + \lambda_2 L)^{1/2}$, $\tilde{\mathbf{y}} = \tilde{\mathbf{y}}(\lambda_2) = \widetilde{X}^{-1} X' \mathbf{y}$ and

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

Since $\widetilde{X}'(\widetilde{X}/n, \tilde{\mathbf{y}}) = (\Sigma + \lambda_2 L, X' \mathbf{y})$, $M(\mathbf{b}; \lambda, \gamma) - \widetilde{M}(\mathbf{b}; \lambda, \gamma) = (\|\mathbf{y}\|^2 - \|\tilde{\mathbf{y}}\|^2)/(2n)$ does not depend on $\mathbf{b}$. Thus, $\hat{\boldsymbol{\beta}}$ is the minimizer of $\widetilde{M}(\mathbf{b}; \lambda, \gamma)$.

Since $|\hat{\beta}_j^o| \geq \gamma \lambda_1$ gives $\rho'(|\hat{\beta}_j^o|; \lambda_1) = 0$, the KKT conditions hold for $\widetilde{M}(\mathbf{b}; \lambda, \gamma)$ at $\hat{\boldsymbol{\beta}}(\lambda) = \hat{\boldsymbol{\beta}}^o(\lambda)$ in the intersection of the events

$$(A.1) \quad \Omega_1 = \{\|\widetilde{X}'_{\mathcal{O}^c}(\tilde{\mathbf{y}} - \widetilde{X}\hat{\boldsymbol{\beta}}^o)/n\|_\infty \leq \lambda_1\}, \qquad \Omega_2 = \Big\{\min_{j \in \mathcal{O}} \operatorname{sgn}(\beta_j^*) \hat{\beta}_j^o \geq \gamma \lambda_1\Big\}.$$

Let $\tilde{\boldsymbol{\varepsilon}}^* = \tilde{\mathbf{y}} - \widetilde{X}\boldsymbol{\beta}^* = \tilde{\boldsymbol{\varepsilon}} + E\tilde{\boldsymbol{\varepsilon}}^*$ with $\tilde{\boldsymbol{\varepsilon}} = \tilde{\mathbf{y}} - E\tilde{\mathbf{y}}$. Since $\widetilde{X}'\tilde{\mathbf{y}} = X'\tilde{\mathbf{y}}$ and both $\boldsymbol{\beta}^o$ and $\boldsymbol{\beta}^*$ are supported in $\mathcal{O}$,

$$(A.2) \quad \begin{aligned} \widetilde{X}'_B E\tilde{\boldsymbol{\varepsilon}}^*/n &= X'_B X \boldsymbol{\beta}^o/n - \widetilde{X}'_B \widetilde{X}\boldsymbol{\beta}^*/n \\ &= \Sigma_{B,\mathcal{O}}\boldsymbol{\beta}_{\mathcal{O}}^o - \Sigma_{B,\mathcal{O}}(\lambda_2)\Sigma_{\mathcal{O}}^{-1}(\lambda_2)\Sigma_{\mathcal{O}}\boldsymbol{\beta}_{\mathcal{O}}^o \\ &= \lambda_2\{\Sigma_{B,\mathcal{O}}(\lambda_2)\Sigma_{\mathcal{O}}^{-1}(\lambda_2)L_{\mathcal{O}} - L_{B,\mathcal{O}}\}\boldsymbol{\beta}_{\mathcal{O}}^o, \end{aligned}$$

which describes the effect of the bias of $\hat{\boldsymbol{\beta}}^o$ on the gradient in the linear model $\tilde{\mathbf{y}} = \widetilde{X}\boldsymbol{\beta}^* + \tilde{\boldsymbol{\varepsilon}}^*$. Since $\widetilde{X}'_{\mathcal{O}} E\tilde{\boldsymbol{\varepsilon}}^*/n = 0$, we have $\|\widetilde{X}' E\tilde{\boldsymbol{\varepsilon}}^*/n\|_\infty = \lambda_2 C_2$.

Since $\widetilde{X}'\widetilde{\boldsymbol{\varepsilon}} = \widetilde{X}'\widetilde{\mathbf{y}} - E\widetilde{X}'\widetilde{\mathbf{y}} = X'\mathbf{y} - EX'\mathbf{y} = X'\boldsymbol{\varepsilon}$, (A.2) gives

(A.3) $\Omega_1 \subseteq \{\|X'_{\mathcal{O}^c}\boldsymbol{\varepsilon}/n\|_\infty < \lambda_1 - \lambda_2 C_2\}.$

Since $\boldsymbol{\beta}^* = E\hat{\boldsymbol{\beta}}^o$, $\hat{\boldsymbol{\beta}}^o_{\mathcal{O}} = \Sigma_{\mathcal{O}}^{-1}(\lambda_2)X'_{\mathcal{O}}\mathbf{y}/n$ can be written as $\boldsymbol{\beta}^*_{\mathcal{O}} + ((v_j/n)^{1/2}\mathbf{u}'_j\boldsymbol{\varepsilon},$
$j \in \mathcal{O})'$, where $\|\mathbf{u}_j\| = 1$ and $\{v_j, j \in \mathcal{O}\}$ are the diagonal elements of $\Sigma_{\mathcal{O}}^{-1}(\lambda_2) \times$
$\Sigma_{\mathcal{O}}\{\Sigma_{\mathcal{O}}^{-1}(\lambda_2)\}$. Thus,

(A.4) $\Omega_2^c \subseteq \bigcup_{j \in \mathcal{O}} \{\mathrm{sgn}(\beta_j^*)\mathbf{u}'_j\boldsymbol{\varepsilon} \geq (n/v_j)^{1/2}|\beta_j^*| \geq \sigma\sqrt{2\log(|\mathcal{O}|/\epsilon)}\}.$

Since $\lambda_1 \geq \lambda_2 C_2 + \sigma\sqrt{2\log(p/\epsilon)}\max_{j\leq p}\|\mathbf{x}_j\|/n$, the sub-Gaussian Condition (A) yields

$$1 - P\{\Omega_1 \cap \Omega_2\} \leq P\Big\{\|X'_{\mathcal{O}^c}\boldsymbol{\varepsilon}/n\|_\infty > \sigma\sqrt{2\log((p-|\mathcal{O}|)/\epsilon)}\max_{j\leq p}\|\mathbf{x}_j\|/n\Big\}$$

$$+ \sum_{j \in \mathcal{O}} P\{\mathrm{sgn}(\beta_j^*)\mathbf{u}'_j\boldsymbol{\varepsilon} \geq \sigma\sqrt{2\log(|\mathcal{O}|/\epsilon)}\}$$

$$\leq 2|\mathcal{O}^c|\epsilon/(p-|\mathcal{O}|) + |\mathcal{O}|\epsilon/|\mathcal{O}| = 3\epsilon.$$

The proof of (4.5) is complete, since $\hat{\beta}_j^o \neq 0$ for all $j \in \mathcal{O}$ in Ω_2.

For the proof of (4.6), we have $\|\boldsymbol{\beta}^*_{\mathcal{O}} - \boldsymbol{\beta}^o_{\mathcal{O}}\|_\infty = \lambda_2 C_1$ due to

(A.5) $\boldsymbol{\beta}^*_{\mathcal{O}} - \boldsymbol{\beta}^o_{\mathcal{O}} = \Sigma_{\mathcal{O}}^{-1}(\lambda_2)\Sigma_{\mathcal{O}}\boldsymbol{\beta}^o_{\mathcal{O}} - \boldsymbol{\beta}^o_{\mathcal{O}} = -\lambda_2\Sigma_{\mathcal{O}}^{-1}(\lambda_2)L_{\mathcal{O}}\boldsymbol{\beta}^o_{\mathcal{O}}.$

It follows that the condition on β_* implies Condition (B)(iii) with $\mathrm{sgn}(\boldsymbol{\beta}^*_{\mathcal{O}}) = \mathrm{sgn}(\boldsymbol{\beta}^o_{\mathcal{O}}) = \mathrm{sgn}(\hat{\boldsymbol{\beta}}^o_{\mathcal{O}})$ in Ω_2. $\quad\square$

PROOF OF THEOREM 2.   For $m \geq 1$ and vectors $\mathbf{u}$ in the range of $\widetilde{X}$, define

(A.6)
$$\tilde{\zeta}(\mathbf{v}; m, \mathcal{O}, \lambda_2)$$
$$= \max\Big\{\frac{\|(\widetilde{P}_B - \widetilde{P}_{\mathcal{O}})\mathbf{v}\|_2}{(mn)^{1/2}} : \mathcal{O} \subseteq B \subseteq \{1, \ldots, p\}, |B| = m + |\mathcal{O}|\Big\},$$

where $\widetilde{P}_B = \widetilde{X}_B(\widetilde{X}'_B\widetilde{X}_B)^{-1}\widetilde{X}'_B$. Here $\tilde{\zeta}$ depends on λ_2 through $\widetilde{P}$. Since $\hat{\boldsymbol{\beta}}(\lambda)$ is the MC+ estimator based on data $(\widetilde{X}, \widetilde{\mathbf{y}})$ at penalty level λ_1 and (4.9) holds for $\Sigma(\lambda_2) = \widetilde{X}'\widetilde{X}/n$, the proof of Theorem 5 in Zhang (2010) gives $\hat{\boldsymbol{\beta}}(\lambda) = \hat{\boldsymbol{\beta}}^o(\lambda)$ in the event $\Omega = \bigcap_{j=1}^3 \Omega_j$, where $\Omega_1 = \{\|\widetilde{X}'_{\mathcal{O}^c}(\widetilde{\mathbf{y}} - \widetilde{X}\hat{\boldsymbol{\beta}}^o)/n\|_\infty \leq \lambda_1\}$ is as in (A.1) and

$$\Omega_2 = \Big\{\min_{j \in \mathcal{O}}\mathrm{sgn}(\beta_j^*)\hat{\beta}_j^o > \gamma(2\sqrt{c^*}\lambda_1)\Big\},$$

$$\Omega_3 = \{\zeta(\widetilde{\mathbf{y}} - \widetilde{X}\boldsymbol{\beta}^*; d^* - |\mathcal{O}|, \mathcal{O}, \lambda_2) \leq \lambda_1\}.$$

Note that $(\lambda_{1,\epsilon}, \lambda_{2,\epsilon}, \lambda_{3,\epsilon}, \alpha)$ in Zhang (2010) is identified with $(\lambda_1, 2\sqrt{c^*}\lambda_1, \lambda_1, 1/2)$ here.

Let $\tilde{\boldsymbol{\varepsilon}}^* = \tilde{\mathbf{y}} - \widetilde{X}\boldsymbol{\beta}^* = \tilde{\boldsymbol{\varepsilon}} + E\tilde{\boldsymbol{\varepsilon}}^*$ with $\tilde{\boldsymbol{\varepsilon}} = \tilde{\mathbf{y}} - E\tilde{\mathbf{y}}$. Since $\widetilde{X}'\tilde{\mathbf{y}} = X'\mathbf{y}$, (A.2) still holds with $\|\widetilde{X}'E\tilde{\boldsymbol{\varepsilon}}^*/n\|_\infty = \lambda_2 C_2$. Since $\widetilde{X}'\tilde{\boldsymbol{\varepsilon}} = X'\mathbf{y} - EX'\mathbf{y} = X'\boldsymbol{\varepsilon}$, (A.2) still gives (A.3). A slight modification of the argument for (A.4) yields

(A.7)
$$\Omega_2^c \subseteq \bigcup_{j \in \mathcal{O}} \{\mathrm{sgn}(\beta_j^*)\mathbf{u}_j'\boldsymbol{\varepsilon} \geq (n/v_j)^{1/2}(|\beta_j^*| - \gamma(2\sqrt{c^*}\lambda_1))$$
$$\geq \sigma\sqrt{2\log(|\mathcal{O}|/\epsilon)}\}.$$

For $|B| \leq d^*$, we have $\|\widetilde{P}_B E\tilde{\boldsymbol{\varepsilon}}^*\|/\sqrt{n} = \|\Sigma_B^{-1/2}(\lambda_2)\widetilde{X}_B'E\tilde{\boldsymbol{\varepsilon}}^*\|/n \leq \|\widetilde{X}_B'E\tilde{\boldsymbol{\varepsilon}}^*/n\|_\infty\sqrt{|B|/c_*(\lambda_2)}$ and $\|\widetilde{P}_B\tilde{\boldsymbol{\varepsilon}}\|/\sqrt{n} = \|\Sigma_B^{-1/2}(\lambda_2)\widetilde{X}_B'\tilde{\boldsymbol{\varepsilon}}\|/n \leq \|X_B'\boldsymbol{\varepsilon}/n\|_\infty \times \sqrt{|B|/c_*(\lambda_2)}$. Thus, by (A.6)

$$\zeta(\tilde{\mathbf{y}} - \widetilde{X}\boldsymbol{\beta}^*; d^* - |\mathcal{O}|, \mathcal{O}, \lambda_2) = \zeta(\tilde{\boldsymbol{\varepsilon}} + E\tilde{\boldsymbol{\varepsilon}}^*; d^* - |\mathcal{O}|, \mathcal{O}, \lambda_2)$$
$$\leq \frac{(\|X'\boldsymbol{\varepsilon}/n\|_\infty + \lambda_2 C_2)\sqrt{d^*}}{\sqrt{(d^* - |\mathcal{O}|)c_*(\lambda_2)}}.$$

Since $|\mathcal{O}| \leq d^*/(K_* + 1)$, this gives

(A.8) $$\Omega_3 \subseteq \{\|X'\boldsymbol{\varepsilon}/n\|_\infty < \sqrt{c_*(\lambda_2)K_*/(K_* + 1)}\lambda_1 - \lambda_2 C_2\}.$$

Since $\max\{1, \sqrt{c_*(\lambda_2)K_*/(K_* + 1)}\}\lambda_1 \geq \lambda_2 C_2 + \sigma\sqrt{2\log(p/\epsilon)}\max_{j \leq p}\|\mathbf{x}_j\|/n$, (A.3), (A.7), (A.8) and Condition (A) imply

$$1 - P\{\Omega_1 \cap \Omega_3\} + P\{\Omega_2^c\}$$
$$\leq P\Big\{\|X'\boldsymbol{\varepsilon}/n\|_\infty > \sigma\sqrt{2\log(p/\epsilon)}\max_{j \leq p}\|\mathbf{x}_j\|/n\Big\}$$
$$+ \sum_{j \in \mathcal{O}} P\{\mathrm{sgn}(\beta_j^*)\mathbf{u}_j'\boldsymbol{\varepsilon} \geq \sigma\sqrt{2\log(|\mathcal{O}|/\epsilon)}\}$$
$$\leq 2p(\epsilon/p) + |\mathcal{O}|\epsilon/|\mathcal{O}| = 3\epsilon.$$

The proof of (4.10) is complete, since $\hat{\beta}_j^o \neq 0$ for all $j \in \mathcal{O}$ in Ω_2. We omit the proof of (4.11) since it is identical to that of (4.6). $\square$

PROOF OF PROPOSITION 1. The $\tilde{\boldsymbol{\beta}}$ satisfies

(A.9) $$-\frac{1}{n}\mathbf{x}_j'(\mathbf{y} - X\tilde{\boldsymbol{\beta}}) + \lambda_2(d_j\tilde{\beta}_j - \mathbf{a}_j'\tilde{\boldsymbol{\beta}}) = 0, \qquad 1 \leq j \leq q.$$

Therefore, by Cauchy–Schwarz and using $\|\mathbf{x}_j\|^2 = n$, we have

$$\lambda_2 \max_{1 \leq j \leq q} |d_j\tilde{\beta}_j - \mathbf{a}_j'\tilde{\boldsymbol{\beta}}| \leq \frac{1}{n}\max_{1 \leq j \leq q} |\mathbf{x}_j'(\mathbf{y} - X\tilde{\boldsymbol{\beta}})| \leq \frac{1}{\sqrt{n}}\|\tilde{\mathbf{r}}\|.$$

Now because $G(\tilde{\boldsymbol{\beta}}; \lambda_2) \leq G(\mathbf{0}; \lambda_2)$, we have $\|\tilde{\mathbf{r}}\| \leq \|\mathbf{y}\|$. This proves part (i).

For part (ii), note that we have

$$\lambda_2\big(d_j\tilde{\beta}_j - \mathbf{a}_j'\tilde{\boldsymbol{\beta}} - (d_k\tilde{\beta}_k - \mathbf{a}_k'\tilde{\boldsymbol{\beta}})\big) = \frac{1}{n}(\mathbf{x}_j - \mathbf{x}_k)'\tilde{\mathbf{r}}.$$

Thus

$$\lambda_2|d_j\tilde{\beta}_j - \mathbf{a}_j'\tilde{\boldsymbol{\beta}} - (d_k\tilde{\beta}_k - \mathbf{a}_k'\tilde{\boldsymbol{\beta}})| \le \frac{1}{n}\|\mathbf{x}_j - \mathbf{x}_k\|\,\|\tilde{\mathbf{r}}\|.$$

Part (ii) follows. $\square$

PROOF OF PROPOSITION 2. The $\tilde{\boldsymbol{\beta}}$ must satisfy

$$(A.10) \qquad -\frac{1}{n}\mathbf{x}_j'(\mathbf{y} - X\tilde{\boldsymbol{\beta}}) + \lambda_2(\tilde{\beta}_j - v_g^{-1}\mathbf{1}_g'\tilde{\boldsymbol{\beta}}_g) = 0, \qquad j \in V_g, 1 \le g \le J.$$

Taking the difference between the jth and kth equations in (A.10) for $j, k \in V_g$, we get

$$\lambda_2(\tilde{\beta}_j - \tilde{\beta}_k) = \frac{1}{n}(\mathbf{x}_j - \mathbf{x}_k)'(\mathbf{y} - X\tilde{\boldsymbol{\beta}}), \qquad j, k \in V_g.$$

Therefore,

$$\lambda_2|\tilde{\beta}_j - \tilde{\beta}_k| \le \frac{1}{n}\|\mathbf{x}_j - \mathbf{x}_k\| \cdot \|\mathbf{y} - X\tilde{\boldsymbol{\beta}}\|, \qquad j, k \in V_g.$$

Part (i) follows from this inequality.

Define $\bar{\beta}_g = v_g^{-1}\mathbf{1}_g'\tilde{\boldsymbol{\beta}}_g$. This is the average of the elements in $\tilde{\boldsymbol{\beta}}_g$. For any $j \in V_g$ and $k \in V_h, g \ne h$, we have

$$\lambda_2\big(\tilde{\beta}_j - \bar{\beta}_g - (\tilde{\beta}_k - \bar{\beta}_h)\big) = \frac{1}{n}(\mathbf{x}_j - \mathbf{x}_k)'(\mathbf{y} - X\tilde{\boldsymbol{\beta}}), \qquad j \in V_g, k \in V_h.$$

Thus, part (ii) follows. This completes the proof of Proposition 2. $\square$

REFERENCES

BOLSTAD, B. M., IRIZARRY, R. A., ASTRAND, M. and SPEED, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19** 185–193.

BONDELL, H. D. and REICH, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **64** 115–123, 322–323. MR2422825

BREHENY, P. and HUANG, J. (2011). Coordinate descent algorithms for nonconvex penalized regression methods. *Ann. Appl. Stat.* **5** 232–253.

CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** 33–61. MR1639094

CHIANG, A. P., BECK, J. S., YEN, H. J., TAYEH, M. K., SCHEETZ, T. E., SWIDERSKI, R., NISHIMURA, D., BRAUN, T. A., KIM, K. Y., HUANG, J., ELBEDOUR, K., CARMI, R., SLUSARSKI, D. C., CASAVANT, T. L., STONE, E. M. and SHEFFIELD, V. C. (2006). Homozygosity mapping with SNP arrays identifies a novel gene for Bardet–Biedl Syndrome (BBS10). *Proc. Natl. Acad. Sci. USA* **103** 6287–6292.

CHUNG, F. R. K. (1997). *Spectral Graph Theory. CBMS Regional Conference Series in Mathematics* **92**. Conf. Board Math. Sci., Washington, DC. MR1421568

CHUNG, F. and LU, L. (2006). *Complex Graphs and Networks. CBMS Regional Conference Series in Mathematics* **107**. Conf. Board Math. Sci., Washington, DC. MR2248695

DAYE, Z. J. and JENG, X. J. (2009). Shrinkage and model selection with correlated variables via weighted fusion. *Comput. Statist. Data Anal.* **53** 1284–1298. MR2657091

FAN, J. (1997). Comments on "Wavelets in statistics: A review" by A. Antoniadis. *J. Italian Statist. Assoc.* **6** 131–138.

FAN, J., FENG, Y. and WU, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *Ann. Appl. Stat.* **3** 521–541. MR2750671

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FRANK, I. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35** 109–148.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatist.* **9** 432–441.

FRIEDMAN, J., HASTIE, T., HÖFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1** 302–332. MR2415737

FU, W. J. (1998). Penalized regressions: The bridge versus the lasso. *J. Comput. Graph. Statist.* **7** 397–416. MR1646710

GENKIN, A., LEWIS, D. D. and MADIGAN, D. (2004). Large-scale Bayesian logistic regression for text categorization. Technical report, DIMACS, Rutgers Univ.

HEBIRI, M. and VAN DE GEER, S. (2010). The smooth-Lasso and other $\ell_1 + \ell_2$-penalized methods. Preprint. Available at http://arxiv4.library.cornell.edu/PS_cache/arxiv/pdf/1003/1003.4885v1.pdf.

HUANG, J., BREHENY, P., MA, S. and ZHANG, C. H. (2010a). The Mnet method for variable selection. Technical Report # 402, Dept. Statistics and Actuarial Science, Univ. Iowa.

HUANG, J., MA, S., LI, H. and ZHANG, C. H. (2010b). The sparse Laplacian shrinkage estimator for high-dimensional regression. Technical Report # 403, Dept. Statistics and Actuarial Science, Univ. Iowa.

IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U. and SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatist.* **4** 249–264.

JIA, J. and YU, B. (2010). On model selection consistency of the elastic net when $p \gg n$. *Statist. Sinica* **20** 595–611. MR2682632

LI, C. and LI, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24** 1175–1182.

LI, C. and LI, H. (2010). Variable selection and regression analysis for covariates with graphical structure. *Ann. Appl. Stat.* **4** 1498–1516.

MAZUMDER, R., FRIEDMAN, J. and HASTIE, T. (2009). SparseNet: Coordinate descent with nonconvex penalties. Technical report, Dept. Statistics, Stanford Univ.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

PAN, W., XIE, B. and SHEN, X. (2011). Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*. To appear.

SCHEETZ, T. E., KIM, K. Y. A., SWIDERSKI, R. E., PHILP, A. R., BRAUN, T. A., KNUDTSON, K. L., DORRANCE, A. M., DIBONA, G. F., HUANG, J., CASAVANT, T. L., SHEFFIELD, V. C. and STONE, E. M. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proc. Natl. Acad. Sci. USA* **103** 14429–14434.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242

TUTZ, G. and ULBRICHT, J. (2009). Penalized regression with correlation-based penalty. *Stat. Comput.* **19** 239–253. MR2516217

WU, T. T. and LANGE, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.* **2** 224–244. MR2415601

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

ZHANG, B. and HORVATH, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4** 45 pp. (electronic). MR2170433

ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 301–320. MR2137327

ZOU, H. and ZHANG, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.* **37** 1733–1751. MR2533470

J. HUANG
DEPARTMENT OF STATISTICS
  AND ACTUARIAL SCIENCE, 241 SH
UNIVERSITY OF IOWA
IOWA CITY, IOWA 52242
USA
E-MAIL: jian-huang@uiowa.edu

H. LI
DEPARTMENT OF BIOSTATISTICS
  AND EPIDEMIOLOGY
UNIVERSITY OF PENNSYLVANIA SCHOOL
  OF MEDICINE
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: hongzhe@upenn.edu

S. MA
DIVISION OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06520
USA
E-MAIL: shuangge.ma@yale.edu

C.-H. ZHANG
DEPARTMENT OF STATISTICS
  AND BIOSTATISTICS
RUTGERS UNIVERSITY
PISCATAWAY, NEW JERSEY 08854
USA
E-MAIL: cunhui@stat.rutgers.edu