The Sparse Laplacian Shrinkage Estimator for High-Dimensional Regression

[1]Jian Huang, [2]Shuangge Ma, [3]Hongzhe Li and [4]Cun-Hui Zhang

[1]University of Iowa, [2]Yale University, [3]University of Pennsylvania and [4]Rutgers University

May 2010

The University of Iowa

Department of Statistics and Actuarial Science

Technical Report No. 403

The Sparse Laplacian Shrinkage Estimator for High-Dimensional Regression

[1]Jian Huang, [2]Shuangge Ma, [3]Hongzhe Li and [4]Cun-Hui Zhang

[1]University of Iowa, [2]Yale University, [3]University of Pennsylvania and [4]Rutgers University

Abstract 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. In a special case, it has a similar grouping property as the elastic net method. 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 data example is used to illustrate its application.

KEY WORD: Graphical structure; minimax concave penalty; penalized regression; high-dimensional data; variable selection; oracle property.

Short title. Sparse Laplacian shrinkage estimator

AMS 2000 subject classification. Primary 62J05, 62J07; secondary 62H20, 60F12

Corresponding Author: Jian Huang. Email: jian-huang@uiowa.edu

1 Introduction

Consider the linear regression model

$$\boldsymbol{y} = \sum_{j=1}^{p} \boldsymbol{x}_j \beta_j + \boldsymbol{\varepsilon} \tag{1.1}$$

with n observations and p potential predictors, where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is the vector of n response variables, $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})'$ is the jth predictor, β_j is the j regression coefficient and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is the vector of random errors. We consider the problem of variable selection and estimation in (1.1) in sparse, high-dimensional settings when the predictors have certain correlation patterns.

Our motivation comes from genomic applications where an important and ubiquitous problem is to identify genetic determinants affecting a certain phenotype or clinical outcome. For example, in microarray gene expression profiling studies, genes in the same pathway or functional group tend to have highly correlated expressions. Co-regulated genes have similar or compensating effects on the outcome variable. It is important to take into account the correlation pattern in gene expressions when selecting genes and pathways that are related to the clinical outcome. In genome wide association studies using dense genetic markers, it is useful to consider correlation patterns among genetic markers due to linkage disequilibrium in identifying regions that may harbor disease related mutations.

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.

However, these methods do not take into account correlations among predictors. This can lead to unsatisfactory selection results 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 2009 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. (2010) proposed the Mnet method, which uses a combination of the MCP and ℓ_2 penalties. The Mnet method is equal to the oracle ridge estimator with high probability under certain conditions. These methods are effective in dealing with certain types of colinearity among predictors and has the useful grouping property of selecting and dropping highly correlated predictors together. However, they do not use any specific information on correlation patterns 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 (2009) 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 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 (2009) 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 2005) with the nodes in the network as groups. This method is capable of grouped variable selection recognizing grouping effects, but it does not provide 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 that uses a combination of the MCP and Laplacian quadratic as the penalty. 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 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 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 result is 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 an 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.

- Finally, 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 Hebri and 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 microarry 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

Let $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ 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

$$M(\boldsymbol{b}; \lambda, \gamma) = \frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{b}\|^2 + \sum_{j=1}^{p} \rho(|b_j|; \lambda_1, \gamma) + \frac{1}{2}\lambda_2 \sum_{1 \leq j < k \leq p} |a_{jk}|(b_j - s_{jk}b_k)^2, \tag{2.1}$$

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 connection between $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$, and $s_{jk} = \text{sgn}(a_{jk})$ is the sign of a_{jk}, i.e., $s_{jk} = -1, 0,$ or 1 if $a_{jk} < 0, = 0,$ or > 0. The two penalty terms in (2.1) 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 $\hat{\boldsymbol{\beta}}(\lambda, \gamma) = \text{argmin}_{\boldsymbol{b}} M(\boldsymbol{b}; \lambda, \gamma)$.

2.1 The rational for using the MCP

Although other penalties can be used as ρ in the SLS criterion (2.1), we proposed to use the MCP defined as

$$\rho(t; \lambda_1, \gamma) = \lambda_1 \int_0^{|t|} (1 - x/(\gamma\lambda_1))_+ dx, \tag{2.2}$$

where for any $a \in R$, a_+ is the nonnegative part of a, i.e., $a_+ = a1_{\{a \geq 0\}}$. The MCP can be easily understood by considering its derivative,

$$\dot{\rho}(t; \lambda_1, \gamma) = \lambda_1(1 - |t|/(\gamma\lambda_1))_+ \text{sgn}(t). \tag{2.3}$$

We can see that the MCP begins by applying the same rate of penalization as the ℓ_1 penalty, but continuously reduces that penalization until, when $|t| > \gamma\lambda$, the rate of penalization drops to 0. 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 the hard-thresholding penalty as $\gamma \to 1+$ and the convex ℓ_1 penalty at $\gamma = \infty$.

As discussed in Zhang (2010), the MCP belongs to the family of quadratic spline penalties satisfying the requirements that they have the sparsity and continuity properties discussed in Fan and Li (2001). This family also includes the ℓ_1 and SCAD penalties. The ℓ_1 penalty is the only member in this family that is convex, but it yields biased estimates. The MCP minimizes the maximum concavity measure defined in Zhang (2010). It has the simplest form in this family that results in an estimator that is nearly unbiased, sparse and continuous. Further discussions on the advantages of the MCP over other popular penalties can be found in Mazumder, Friedman and Hastie (2009).

2.2 The Laplacian and signed adjacency matrices

We express the nonnegative quadratic form in the second penalty term in (2.1) using a positive semi-definite matrix L, which satisfies

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

Define $a_{kj} = a_{jk}, 1 \le j < k \le p$. The diagonal elements a_{jj}'s do not appear in the quadratic form. We can define them any way that is most convenient for a specific situation. For now, we simply assume their values are given. Let $A = (a_{jk}, 1 \le j, k \le p)$ and $D = \text{diag}(d_1, \ldots, d_p)$, where $d_j = \sum_{k=1}^p |a_{jk}|$. We have $\sum_{1\le j<k\le p} |a_{jk}|(b_j - s_{jk}b_k)^2 = \boldsymbol{b}'(D - A)\boldsymbol{b}$. Therefore, $L = D - A$. This matrix is associated with a labeled weighted graph $\mathcal{G} = (V, \mathcal{E}, w)$ with vertex set $V = \{1, \ldots, p\}$ and edge set $\mathcal{E} = \{(j, k) : (j, k) \in V \times V\}$. Here $|a_{jk}|$ is the weight of edge (j, k) and d_j the degree of vertex j. Here 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 adjacent coefficient but are negatively correlated. Note that the usual adjacency matrix can be considered a special case of signed adjacency matrix when all $a_{jk} \ge 0$. 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

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

Thus for $\lambda_2 > 0$, the Laplacian penalty shrinks $b_j - s_{jk}b_k$ towards zero for $a_{jk} \neq 0$. This can also be considered 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 towards 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 (2.1) can be written as

$$M(\boldsymbol{b};\lambda,\gamma) = \frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{b}\|^2 + \sum_{j=1}^{p} \rho(|b_j|;\lambda_1,\gamma) + \frac{1}{2}\lambda_2 \boldsymbol{b}'(D-A)\boldsymbol{b}. \tag{2.4}$$

Here the Laplacian is not normalized, meaning that the weight d_j is not standardized to 1. This criterion favors the predictors with a larger weight. To see this, make the transformation $\boldsymbol{b}^* = D^{1/2}\boldsymbol{b}$, $X^* = XD^{-1/2}$ and define $A^* = D^{-1/2}AD^{-1/2}$. We have

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

The penalty in ρ for the jth predictor is proportional to $1/\sqrt{d_j}$. Therefore, predictors with larger d_j will be more likely to be selected. Because d_j measures the connectivity of $\boldsymbol{x}_j$, this is desirable in certain applications. For example, 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.

However, in problems where predictors should be treated without preference with respect to connectivity, we can first normalized the Laplacian $L^* = I_p - A^*$ and use the criterion

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

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 (2.4) when studying its properties.

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 $a_{jk} = \phi(\|\boldsymbol{x}_j - \boldsymbol{x}_k\|/\sqrt{n})$, where $\phi : [0, \infty) \mapsto [0, \infty)$. A simple adjacency function is the threshold function $\phi(x) = 1\{x \leq 2r\}$. Then

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

It is convenient to express a_{jk} in terms of the Person's correlation coefficient r_{jk} between $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$, where $r_{jk} = \boldsymbol{x}_j'\boldsymbol{x}_k/(\|\boldsymbol{x}_j\| \, \|\boldsymbol{x}_k\|)$. For predictors that are standardized with $\|\boldsymbol{x}_j\|^2 = n, 1 \leq j \leq p$, we have $\|\boldsymbol{x}_j - \boldsymbol{x}_k\|^2/n = 2 - 2r_{jk}$. 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 $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$ 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 defined based on similarity measures. An often used similar-

ity measure is Pearson's correlation coefficient r_{jk}. Other correlation measures such as Spearman's correlation can also be used. Let

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

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

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

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

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

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

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 $a_{jk} = \exp(-\|\boldsymbol{x}_j - \boldsymbol{x}_k\|^2/n\tau^2)$ for $\tau > 0$. The resulting adjacency matrix $A = [a_{jk}]$ is the Gram matrix associated with the Gaussian kernel. 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. Define

$$\hat{\boldsymbol{\beta}}^o(\lambda_2) = \underset{\boldsymbol{b}}{\text{argmin}}\{\frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{b}\|^2 + \frac{1}{2}\lambda_2 \boldsymbol{b}' L \boldsymbol{b}, b_j = 0, j \notin \mathcal{O}\}. \tag{4.1}$$

This is the oracle Laplacian shrinkage estimator on the set $\mathcal{O}$. Theorems 1 and 2 below provide sufficient conditions under which $\mathrm{P}(\mathrm{sgn}(\hat{\boldsymbol{\beta}}) \neq \mathrm{sgn}(\boldsymbol{\beta}^o)$ 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 $\boldsymbol{v}$, the design matrix X and $V = (v_{ij})_{p\times p}$, define

$$\boldsymbol{v}_B = (v_j, j \in B)',\ X_B = (\boldsymbol{x}_j, j \in B),\ V_{A,B} = (v_{ij}, i \in A, j \in B)_{|A|\times|B|},\ V_B = V_{B,B}.$$

For example, $\Sigma_B = X_B'X_B/n$ and $\Sigma_{\mathcal{O}}(\lambda_2) = \Sigma_{\mathcal{O}} + \lambda_2 L_{\mathcal{O}}$. Let $|B|$ denotes 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:

$$C_1 = \|\Sigma_{\mathcal{O}}^{-1}(\lambda_2)L_{\mathcal{O}}\boldsymbol{\beta}^o_{\mathcal{O}}\|_\infty,\ C_2 = \|\{\Sigma_{\mathcal{O}^c,\mathcal{O}}(\lambda_2)\Sigma_{\mathcal{O}}^{-1}(\lambda_2)L_{\mathcal{O}} - L_{\mathcal{O}^c,\mathcal{O}}\}\boldsymbol{\beta}^o_{\mathcal{O}}\|_\infty. \tag{4.2}$$

We make the following sub-Gausian assumption on the error terms in (1.1).

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

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

4.1 Convex penalized loss

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

$$\hat{\boldsymbol{\beta}}^o_{\mathcal{O}} = (\Sigma_{\mathcal{O}} + \lambda_2 L_{\mathcal{O}})^{-1}X'_{\mathcal{O}}\boldsymbol{y}/n, \quad \hat{\boldsymbol{\beta}}^o_{\mathcal{O}^c} = 0, \tag{4.3}$$

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

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

Condition (B): (i) *$c_{\min}(\lambda_2) > 1/\gamma$ with $\rho(t; \lambda_1, \gamma)$ in (2.1).* (ii) *The penalty levels satisfy*

$$\lambda_1 \geq \lambda_2 C_2 + \sigma\sqrt{2\log((p - |\mathcal{O}|)/\epsilon)}\max_{j\leq p}\|\boldsymbol{x}_j\|/n$$

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

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

Define $\beta_* = \min\{|\beta_j|, j \in \mathcal{O}\}$. 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,*

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

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

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

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. Condition (B) (ii) requires that λ_2 is at most a multiple of λ_1. This is to ensure the bias introduced by the Laplacian shrinkage does not interfere with selection. 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 $J \geq 1$. Nodes j and k belong to the same clique if $a_{jk_1}a_{k_1k_2}\cdots a_{k_mk} \neq 0$ through a certain chain $j \to k_1 \to k_2 \to \cdots \to k_m \to k$. Let $\bar{\boldsymbol{x}}_g = \sum_{j\in V_g} \boldsymbol{x}_j/|V_g|$, where $|V_g|$ is the cardinality of V_g. The matrix $\Sigma + \lambda_2 L$ is positive definite if and only if $\boldsymbol{b}'\Sigma\boldsymbol{b} = \boldsymbol{b}'L\boldsymbol{b} = 0$ implies $\boldsymbol{b} = 0$, which is true if and only if the vectors $\bar{\boldsymbol{x}}_g$ 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 $\bar{\boldsymbol{x}}_g$ 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 $\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{y}}(\lambda_2) = (\widetilde{X}')^\dagger X'\boldsymbol{y}$, where $(\widetilde{X}')^\dagger$ is the Moore-Penrose inverse of $\widetilde{X}'$. Define

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

Since $\boldsymbol{b}'\Sigma(\lambda_2)\boldsymbol{b} = 0$ implies $\boldsymbol{b}'X'\boldsymbol{y} = 0$, we have $\widetilde{X}'\tilde{\boldsymbol{y}} = X'\boldsymbol{y}$. It follows that $M(\boldsymbol{b};\lambda,\gamma) - \widetilde{M}(\boldsymbol{b};\lambda,\gamma) = (\|\boldsymbol{y}\|^2 - \|\tilde{\boldsymbol{y}}\|^2)/(2n)$. Thus, the two penalized loss functions have the same set of local minimizers. Since (4.7) is the penalized loss with data $(\widetilde{X}, \tilde{\boldsymbol{y}})$, we define the estimator

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

where the map $\boldsymbol{\delta}(X, \boldsymbol{y}, \lambda_1) \in \mathbb{R}^p$ defines the MC+ estimator (Zhang, 2010) with data $(X, \boldsymbol{y})$ and penalty level λ_1. For the computation of $\hat{\boldsymbol{\beta}}(\lambda)$, $\tilde{\boldsymbol{y}}$ can be any solution of $\widetilde{X}'\tilde{\boldsymbol{y}} = X'\boldsymbol{y}$.

Condition (C): (i) *For an integer d^* and spectrum bounds $0 < c_*(\lambda_2) \le c^*(\lambda_2) < \infty$,*

$$0 < c_*(\lambda_2) \le \boldsymbol{u}_B'\Sigma_B(\lambda_2)\boldsymbol{u}_B \le c^*(\lambda_2) < \infty,\ \forall B \text{ with } |B| \le d^*, \|\boldsymbol{u}_B\| = 1, \tag{4.9}$$

with $d^* \ge d^o(K_*+1)$ and $\gamma \ge c_*^{-1}(\lambda_2)\sqrt{4 + c_*(\lambda_2)/c^*(\lambda_2)}$ in (2.1), where $K_* = c^*(\lambda_2)/c_*(\lambda_2) - (1/2)$. (ii) *With $C_2 = \|\{\Sigma_{B,\mathcal{O}}(\lambda_2)\Sigma_{\mathcal{O}}^{-1}(\lambda_2)L_{\mathcal{O}} - L_{B,\mathcal{O}}\}\boldsymbol{\beta}^o_{\mathcal{O}}\|_\infty$,*

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

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

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

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

$$\mathrm{P}\Big(\{j : \hat{\beta}_j \ne 0\} \ne \mathcal{O} \text{ or } \hat{\boldsymbol{\beta}} \ne \hat{\boldsymbol{\beta}}^o\Big) \le 3\epsilon. \tag{4.10}$$

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

$$\mathrm{P}\Big(sgn(\hat{\boldsymbol{\beta}}) \ne sgn(\boldsymbol{\beta}^o) \text{ or } \hat{\boldsymbol{\beta}} \ne \hat{\boldsymbol{\beta}}^o\Big) \le 3\epsilon. \tag{4.11}$$

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) is a stronger version of Condition (B) designed to handle the noncovexity of the penalized loss. If X satisfies the SRC (4.9), so does $\widetilde{X}$. Thus a stronger and perhaps more natural condition is to assume that X satisfies the SRC. However, (4.9) is used here for more generality. The SRC is required to ensure 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 the regularized Gram matrix $\Sigma(\lambda_2) = \Sigma + \lambda_2 L$, not the Gram matrix Σ of the design matrix X. We can have $p \gg n$ but still require $d^o \leq d^*/(1 + K_*)$ as in (4.9). Note that d^* and K_* can depend on n. Thus here we allow $p \gg n$ but require that the model is spares, in the sense that the number of nonzero coefficients is smaller than $d^*/(1 + K_*)$. Furthermore, we note that the results in Theorem 2 hold for a local minimizer defined as in (4.8).

Theorem 2 shows that the SLS estimator automatically adapts to the sparseness and denseness of the model. From the original sparse p-dimensional model, it correctly selects the true underlying model. 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). Also, as in the convex penalized loss setting, here the results do not require that the underlying correlation structure of the predictors is correctly specified.

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

$$\hat{\boldsymbol{\beta}}^o_{\mathcal{O}}(\lambda_2) = \Sigma^{-1}_{\mathcal{O}}(\lambda_2) X'_{\mathcal{O}} \boldsymbol{y}/n.$$

Without the Laplacian, i.e., when $\lambda_2 = 0$, it becomes the least squares (LS) estimator,

$$\hat{\boldsymbol{\beta}}^o_{\mathcal{O}}(0) = \Sigma^{-1}_{\mathcal{O}} X'_{\mathcal{O}} \boldsymbol{y}/n.$$

If some of the predictors in $\{\boldsymbol{x}_j, j \in \mathcal{O}\}$ are highly correlated or $|\mathcal{O}| \geq n$, the LS estimator $\hat{\boldsymbol{\beta}}^o_{\mathcal{O}}(0)$ is not be 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 $\{\boldsymbol{x}_j, j \in \mathcal{O}\}$ are highly correlated or $|\mathcal{O}| \geq n$.

For the second question, we examine the bias of $\hat{\boldsymbol{\beta}}^o_{\mathcal{O}}(\lambda_2)$. Since the bias of the target vector (4.4) is $\boldsymbol{\beta}^o_{\mathcal{O}} - \boldsymbol{\beta}^*_{\mathcal{O}}(\lambda_2) = \lambda_2 \Sigma^{-1}_{\mathcal{O}}(\lambda_2) L_{\mathcal{O}} \boldsymbol{\beta}^o_{\mathcal{O}}$, $\hat{\boldsymbol{\beta}}^o_{\mathcal{O}}(\lambda_2)$ is unbiased if and only if $L_{\mathcal{O}} \boldsymbol{\beta}^o_{\mathcal{O}} = 0$. Therefore, in terms of bias reduction, a Laplacian L is most appropriate if the condition $L_{\mathcal{O}} \boldsymbol{\beta}^o_{\mathcal{O}} = 0$ is satisfied. We shall say that a Laplacian L is unbiased if $L_{\mathcal{O}} \boldsymbol{\beta}^o_{\mathcal{O}} = 0$.

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

$$E\|\hat{\boldsymbol{\beta}}^o_{\mathcal{O}}(\lambda_2) - \boldsymbol{\beta}^o_{\mathcal{O}}\|^2 = \sigma^2 \mathrm{trace}(\Sigma^{-1}_{\mathcal{O}}(\lambda_2) \Sigma_{\mathcal{O}} \Sigma^{-1}_{\mathcal{O}}(\lambda_2)).$$

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

$$E\|\hat{\boldsymbol{\beta}}^o_{\mathcal{O}}(0) - \boldsymbol{\beta}^o_{\mathcal{O}}\|^2 = \sigma^2 \mathrm{trace}(\Sigma^{-1}_{\mathcal{O}}).$$

We always have $E\|\hat{\boldsymbol{\beta}}^o_{\mathcal{O}}(\lambda_2) - \boldsymbol{\beta}^o_{\mathcal{O}}\|^2 < E\|\hat{\boldsymbol{\beta}}^o_{\mathcal{O}}(0) - \boldsymbol{\beta}^o_{\mathcal{O}}\|^2$ 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 $\hat{\boldsymbol{\beta}}^o_{\mathcal{O}}$. 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

$$\tilde{\boldsymbol{\beta}}(\lambda_2) = \underset{\boldsymbol{b}}{\mathrm{argmin}} \{G(\boldsymbol{b}; \lambda_2) \equiv \frac{1}{2n} \|\boldsymbol{y} - X\boldsymbol{b}\|^2 + \frac{1}{2} \lambda_2 \boldsymbol{b}' L \boldsymbol{b}, \ \boldsymbol{b} \in \mathbb{R}^q\}. \tag{5.1}$$

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

Proposition 1 *Let* $\tilde{\boldsymbol{r}} = \boldsymbol{y} - X\tilde{\boldsymbol{\beta}}$.

(i)

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

(ii)

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

Note that $\boldsymbol{a}_j'\tilde{\boldsymbol{\beta}}/d_j = \sum_{k=1}^q a_{jk}\tilde{\beta}_k/d_j$ is a weighted average of the $\tilde{\beta}_k$'s connected to $\hat{\beta}_j$, 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 $\lambda_2 \to \infty$, this difference converges to zero. Part (ii) implies that the difference between the centered $\tilde{\beta}_j$ and $\tilde{\beta}_k$ converges to zero if $\|\boldsymbol{x}_j - \boldsymbol{x}_k\| \to 0$.

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 sets $V_1, \ldots, V_J$. We consider the normalized Laplacian $L = I_q - A$, where I_q is a $q \times q$ identity matrix and $A = \mathrm{diag}(A_1, \ldots, A_J)$ with $A_g = v_g^{-1}\mathbf{1}_g'\mathbf{1}$. Here $v_g = |V_g|, 1 \le g \le J$. Let $\boldsymbol{b}_g = (b_j, j \in V_g)'$. We can write the objective function as

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

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

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

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

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

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

This proposition characterizes the smoothing effect and grouping property of the Laplacian penalty in (5.2). 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 $\|\boldsymbol{x}_j - \boldsymbol{x}_k\| \to 0$. 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 $\|\boldsymbol{x}_j - \boldsymbol{x}_k\| \to 0$.

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

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

The Mnet estimator is also a special case of $\hat{\boldsymbol{\beta}}_Q$ for $Q = I_p$ (Huang et al. 2010). 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

$$\hat{\boldsymbol{\beta}}^o_Q(\lambda_2) = \underset{\boldsymbol{b}}{\operatorname{argmin}}\{\frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{b}\|^2 + \frac{1}{2}\boldsymbol{b}' Q \boldsymbol{b}, b_j = 0, j \notin \mathcal{O}\}.$$

Then in a way similar to the discussion in Section 4, $\hat{\boldsymbol{\beta}}_Q$ is nearly unbiased if and only if $Q_{\mathcal{O}}\boldsymbol{\beta}^o_{\mathcal{O}} = 0$. Therefore, for $\|\boldsymbol{\beta}^o_{\mathcal{O}}\| \neq 0$, $Q_{\mathcal{O}}$ 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 lead to 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. The Laplacian shrinkage and ridge estimators are defined as

$$(\hat{b}_{L1}(\lambda_2), \hat{b}_{L2}(\lambda_2)) = \underset{b1,b_2}{\operatorname{argmin}} \frac{1}{2n}\sum_{i=1}^{n}(y_i - x_{i1}b_1 - x_{i2}b_2)^2 + \frac{1}{2}\lambda_2(b_1 - b_2)^2,$$

and

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

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

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

Let

$$\hat{b}_{ols1} = \frac{r_1 - r_{12}r_2}{1-r_{12}^2}, \quad \hat{b}_{ols2} = \frac{r_2 - r_{12}r_1}{1-r_{12}^2}, \quad \hat{b}_L(\infty) = \frac{r_1 + r_2}{2(1+r_{12})},$$

where $(\hat{b}_{ols1}, \hat{b}_{ols2})$ 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

$$\hat{b}_{L1}(\lambda_2) = (1-w_L)\hat{b}_{ols1} + w_L\hat{b}_L(\infty) \text{ and } \hat{b}_{L2}(\lambda_2) = (1-w_L)\hat{b}_{ols2} + w_L\hat{b}_L(\infty).$$

Thus for any fixed λ_2, $\hat{b}_L(\lambda_2)$ is a weighted average of $\hat{b}_{ols}$ 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 towards a common value, which is the OLS estimate assuming equal regression coefficients.

Now consider the ridge regression estimator. We have

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

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 $\boldsymbol{y}$ on $\boldsymbol{x}_1$ and $\boldsymbol{y}$ on $\boldsymbol{x}_2$, respectively. We can write

$$(1+\lambda_2)\hat{b}_{R1}(\lambda_2) = c_{\lambda_2}(1-w_R)\hat{b}_{ols1} + c_\lambda w_R \hat{b}_{u1},$$

$$(1+\lambda_2)\hat{b}_{R2}(\lambda_2) = c_{\lambda_2}(1-w_R)\hat{b}_{ols2} + c_\lambda w_R \hat{b}_{u2},$$

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 towards individual univariate regression estimates.

6 Numerical Studies

6.1 Computational algorithm

For fixed (λ_2, γ), the PLUS algorithm of Zhang (2010) can be used to compute a path of (4.8) as a function of λ_1. This can be done using $(\widetilde{X}, \tilde{\boldsymbol{y}})$ as the input in the *plus* package in R. However, for $p \gg n$, the dimension of $\widetilde{X}$ is at least $\text{rank}(\Sigma + \lambda_2 L) \times p$, so that $\widetilde{X}$ could be much larger than X in dimension. The current *plus* package needs to be updated to incorporate the computation of SLS with large $\text{rank}(\Sigma + \lambda_2 L)$.

Here we adopt 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 et al. 2004; Friedman et al. 2007; Wu and Lange 2007). It has been proposed to calculate the MCP estimates (Breheny and Huang 2009).

Suppose we have current values of $\tilde{\beta}_k$ for $k \neq j$ and want to minimize (2.1) with respect to β_j to obtain its current value $\tilde{\beta}_j$. Define

$$M_j(\beta_j; \lambda) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \sum_{k \neq j} x_{ik} \tilde{\beta}_k - x_{ij} \beta_j)^2 + \rho(|\beta_j|; \lambda_1, \gamma) + \frac{1}{2} \lambda_2 d_j \beta_j^2 - \frac{1}{2} (\lambda_2 \sum_{k \neq j} a_{jk} \tilde{\beta}_k) \beta_j.$$

Let $\tilde{r}_{ij} = y_i - \sum_{k \neq j} x_{ik} \tilde{\beta}_k$, $\tilde{z}_j = n^{-1} \sum_{i=1}^{n} x_{ij} \tilde{r}_{ij}$ and $\tilde{\xi}_j = 2^{-1} \lambda_2 \sum_{k \neq j} a_{jk} \tilde{\beta}_k$. Some algebra shows that

$$\tilde{\beta}_j \equiv \underset{\beta_j}{\text{argmin}}\, M_j(\beta_j; \lambda) = \underset{\beta_j}{\text{argmin}} \Big\{ \frac{1}{2} (\tilde{z}_j - \tilde{\xi}_j - \beta_j)^2 + \frac{1}{2} \lambda_2 d_j \beta_j^2 + \rho(|\beta_j|; \lambda_1, \gamma) \Big\}.$$

This is an univariate MCP penalized estimation problem. When $\gamma(1 + \lambda_2 d_j) > 1$, its solution is

$$\tilde{\beta}_j = \begin{cases} \text{sgn}(\tilde{z}_j - \tilde{\xi}_j) \frac{\gamma(|\tilde{z}_j - \tilde{\xi}_j| - \lambda_1)_+}{\gamma(1 + \lambda_2 d_j) - 1} & \text{if } |\tilde{z}_j - \tilde{\xi}_j| \leq \gamma \lambda_1 (1 + \lambda_2 d_j), \\ \frac{\tilde{z}_j - \tilde{\xi}_j}{1 + \lambda_2 d_j} & \text{if } |\tilde{z}_j - \tilde{\xi}_j| > \gamma \lambda_1 (1 + \lambda_2 d_j). \end{cases} \tag{6.1}$$

Define $\hat{y}_i = \sum_{j=1}^{n} x_{ij} \tilde{\beta}_j$ and $\hat{r}_i = y_i - \hat{y}_i$, where $\hat{y}_i$ is the current fitted value and $\hat{r}_i$ is the current residual. Denote $\hat{\boldsymbol{r}} = (\hat{r}_1, \ldots, \hat{r}_n)'$ and let $\tilde{\boldsymbol{\beta}}^{(s)} = (\tilde{\beta}_1^{(s)}, \ldots, \tilde{\beta}_p^{(s)})'$ be the current estimate. The coordinate descent algorithm consists of the following steps.

(1) Calculate $\tilde{z}_j = n^{-1}\boldsymbol{x}_j'\hat{\boldsymbol{r}} + \tilde{\beta}_j^{(s)}$ and $\tilde{\xi}_j = 2^{-1}\lambda_2 \sum_{k\neq j} a_{jk}\tilde{\beta}_k^{(s)}$.

(2) Update $\tilde{\beta}_j^{(s+1)}$ using (6.1).

(3) Update $\hat{\boldsymbol{r}} \leftarrow \hat{\boldsymbol{r}} - (\tilde{\beta}_j^{(s+1)} - \tilde{\beta}_j^{(s)})\boldsymbol{x}_j$.

The last step ensures that $\hat{\boldsymbol{r}}$ always holds the current values of the residuals. This algorithm always converges to a local minimum (Mazumder et al. 2009). Therefore, in the case of convex penalized loss in Section 4.1, the $\hat{\boldsymbol{\beta}}$ computed using this algorithm satisfies Theorem 1. For fixed $\{\gamma, \lambda_2\}$, Theorem 2 holds when $\hat{\boldsymbol{\beta}}$ is defined as in (4.8). This is the case when the PLUS algorithm is used. If we let the step size converge to zero, the path of a coordinate descent algorithm will converge to the PLUS path if λ_1 is adjusted in the iteration in a fashion similar to that in the boosted Lasso (Zhao and Yu 2004).

6.2 Simulation studies

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 determined as 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 determined as 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 $\alpha = 6$.

Penalty parameters λ_1, λ_2 are selected using V-fold cross validation. 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 γ in MCP estimates, 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 (2009) suggest that $\gamma = 3$ is a reasonable choice. We

have experimented with different γ values and reached the same conclusion. Therefore, we fix the value of γ at 3.

We set the sample size $n = 100$ and the number of covariates $p = 500$. Among the 500 covariates, there are 100 clusters, each with size 5. Covariates in different clusters are independent, whereas covariates i and j within the same cluster have correlation coefficients $\rho^{|i-j|}$. 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 1; (b) all the nonzero coefficients are equal to 0.5; (c) the nonzero coefficients are randomly generated from the uniform distribution on $[0.5, 1.5]$; and (d) the nonzero coefficients are randomly generated from the uniform distribution on $[0.25, 0.75]$. Scenarios (a) and (b) are the ideal cases where correlated covariates have the same effects, whereas scenarios (c) and (d) represent the more realistic scenario where nonzero coefficients have the same sign but are not equal. In (a) and (b), the Laplacian matrices satisfy the unbiasedness property $L\beta^o = 0$ discussed in Section 4.

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 smaller prediction errors.

			D-J				SLS			
Coefficient	ρ	MCP	N.1	N.2	N.3	N.4	N.1	N.2	N.3	N.4
1	0.1	25 25 57.89	74 25 95.62	61 25 86.36	84 25 67.61	64 25 77.26	25 25 59.57	25 25 59.18	25 25 57.30	25 25 55.28
	0.5	25 25 69.51	76 25 114.1	75 25 115.96	74 25 94.97	64 25 84.69	25 25 61.20	25 25 60.78	26 25 57.15	25 25 54.75
	0.9	23 16 412.9	44 25 128.18	44 25 134.12	77 25 106.17	72 25 112.3	27 25 105.12	26 25 100.64	29 25 89.23	32 25 94.27
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 54.1	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[.5, 1.5]	0.1	26 25 59.62	63 25 86.93	62 25 71.39	57 25 90.84	63 25 94.42	27 25 61.51	27 25 57.33	26 25 57.51	26 25 62.15
	0.5	25 25 108.71	67 25 116.01	62 25 88.86	78 25 62.11	78 25 83.43	27 25 67.73	27 25 65.11	25 25 92.82	25 25 92.50
	0.9	16 14 488.9	41 25 132.52	45 25 80.85	75 25 72.94	64 25 86.63	26 25 127.20	26 25 113.34	25 25 98.58	25 25 113.98
U[.25, .75]	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
	0.5	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 57.3	50 25 53.88	74 25 57.98	29 25 47.73	29 25 49.14	27 25 48.49	28 25 50.83

Table 1: Simulation study: median based on 500 replicates. In each cell, the three numbers are positive findings, true positives, and PMSE $\times 100$, respectively.

6.3 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 2003, Irizzary 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, 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 are small and since that we 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 1000. We then standardize gene expressions to have zero mean and unit variance.

In Table 2, we show the number of genes identified under different adjacency measures. It is clear that different ways of accounting for the adjacency structure has a significant impact on gene identification results. For comparison, we also consider the MCP and D-J approach. The MCP identifies 23 genes, which have 9, 8, 6 and 6 overlapped genes with the proposed approach under N.1–N.4. Under N.1–N.4 adjacency measures respectively, the D-J approach identifies 31 (20), 41 (23), 34 (15) and 30 (14) genes, where the numbers in the "()" are the number of overlapped genes with the proposed approach.

Table 2: Data analysis: number of overlapped genes identified under different adjacency measures.

	N.1	N.2	N.3	N.4
N.1	25	18	11	12
N.2		23	13	14
N.3			16	16
N.4				17

With real data, we are unable to assess gene identification accuracy. Instead, we consider the V-fold cross validation based prediction evaluation, which consists of the following steps: (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); the proposed approach: 1.842 (N.1), 1.687 (N.2), 1.378 (N.3) and 1.441 (N.4), respectively.

With this dataset, the SLS with adjacency measure N.3 outperforms the other approaches. It identifies a smaller set of genes which correspond to a smaller model with a more focused hypothesis to test. In addition, it has the smallest cross validated prediction error.

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 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 the 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

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

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.

8 Appendix

In the 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 (2.1) is strictly convex and its minimizer is unique. Let $\widetilde{X} = \widetilde{X}(\lambda_2) = \sqrt{n}(\Sigma + \lambda_2 L)^{1/2}$, $\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{y}}(\lambda_2) = \widetilde{X}^{-1}X'\boldsymbol{y}$ and

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

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

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

$$\Omega_1 = \big\{\big\|\widetilde{X}'_{\mathcal{O}^c}(\tilde{\boldsymbol{y}} - \widetilde{X}\hat{\boldsymbol{\beta}}^o)/n\big\|_\infty \leq \lambda_1\big\},\ \Omega_2 = \big\{\min_{j\in\mathcal{O}} \mathrm{sgn}(\beta_j^*)\hat{\beta}_j^o \geq \gamma\lambda_1\big\}. \tag{8.1}$$

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

$$\widetilde{X}'_B E\tilde{\boldsymbol{\varepsilon}}^*/n \ = \ X'_B X\boldsymbol{\beta}^o/n - \widetilde{X}'_B\widetilde{X}\boldsymbol{\beta}^*/n$$

$$
\begin{aligned}
&= \Sigma_{B,\mathcal{O}}\boldsymbol{\beta}^o_{\mathcal{O}} - \Sigma_{B,\mathcal{O}}(\lambda_2)\Sigma^{-1}_{\mathcal{O}}(\lambda_2)\Sigma_{\mathcal{O}}\boldsymbol{\beta}^o_{\mathcal{O}} \\
&= \lambda_2\{\Sigma_{B,\mathcal{O}}(\lambda_2)\Sigma^{-1}_{\mathcal{O}}(\lambda_2)L_{\mathcal{O}} - L_{B,\mathcal{O}}\}\boldsymbol{\beta}^o_{\mathcal{O}}, \qquad (8.2)
\end{aligned}
$$

which describes the effect of the bias of $\hat{\boldsymbol{\beta}}^o$ on the gradient in the linear model $\tilde{\boldsymbol{y}} = \widetilde{X}\boldsymbol{\beta}^* + \tilde{\boldsymbol{\varepsilon}}^*$. 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 $\widetilde{X}'\tilde{\boldsymbol{\varepsilon}} = \widetilde{X}'\tilde{\boldsymbol{y}} - E\widetilde{X}'\tilde{\boldsymbol{y}} = X'\boldsymbol{y} - EX'\boldsymbol{y} = X'\boldsymbol{\varepsilon}$, (8.2) gives

$$
\Omega_1 \subseteq \big\{\|X'_{\mathcal{O}^c}\boldsymbol{\varepsilon}/n\|_\infty < \lambda_1 - \lambda_2 C_2\big\}. \qquad (8.3)
$$

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

$$
\Omega_2^c \subseteq \cup_{j\in\mathcal{O}}\Big\{\mathrm{sgn}(\beta^*_j)\boldsymbol{u}'_j\boldsymbol{\varepsilon} \ge (n/v_j)^{1/2}|\beta^*_j| \ge \sigma\sqrt{2\log(|\mathcal{O}|/\epsilon)}\Big\}. \qquad (8.4)
$$

Since $\lambda_1 \ge \lambda_2 C_2 + \sigma\sqrt{2\log(p/\epsilon)}\max_{j\le p}\|\boldsymbol{x}_j\|/n$, the sub-Gausian condition (A) yields

$$
\begin{aligned}
1 - P\big\{\Omega_1 \cap \Omega_2\big\} &\le P\big\{\|X'_{\mathcal{O}^c}\boldsymbol{\varepsilon}/n\|_\infty > \sigma\sqrt{2\log((p-|\mathcal{O}|)/\epsilon)}\max_{j\le p}\|\boldsymbol{x}_j\|/n\big\} \\
&\quad + \sum_{j\in\mathcal{O}} P\big\{\mathrm{sgn}(\beta^*_j)\boldsymbol{u}'_j\boldsymbol{\varepsilon} \ge \sigma\sqrt{2\log(|\mathcal{O}|/\epsilon)}\big\} \\
&\le 2|\mathcal{O}^c|\epsilon/(p-|\mathcal{O}|) + |\mathcal{O}|\epsilon/|\mathcal{O}| = 3\epsilon.
\end{aligned}
$$

The proof of (4.5) is complete, since $\hat{\beta}^o_j \ne 0$ 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

$$
\boldsymbol{\beta}^*_{\mathcal{O}} - \boldsymbol{\beta}^o_{\mathcal{O}} = \Sigma^{-1}_{\mathcal{O}}(\lambda_2)\Sigma_{\mathcal{O}}\boldsymbol{\beta}^o_{\mathcal{O}} - \boldsymbol{\beta}^o_{\mathcal{O}} = -\lambda_2\Sigma^{-1}_{\mathcal{O}}(\lambda_2)L_{\mathcal{O}}\boldsymbol{\beta}^o_{\mathcal{O}}. \qquad (8.5)
$$

It follows that the condition on β_* implies Conditon (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. □

Proof of Theorem 2. For $m \ge 1$ and vectors $\boldsymbol{u}$ in the range of $\widetilde{X}$, define

$$
\tilde{\zeta}(\boldsymbol{v}; m, \mathcal{O}, \lambda_2) = \max\left\{\frac{\|(\widetilde{P}_B - \widetilde{P}_{\mathcal{O}})\boldsymbol{v}\|_2}{(mn)^{1/2}} : \mathcal{O} \subseteq B \subseteq \{1,\dots p\},\ |B| = m + |\mathcal{O}|\right\}, \qquad (8.6)
$$

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 $(\widetilde{X}, \tilde{\boldsymbol{y}})$ 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 $\Omega = \cap_{j=1}^3\Omega_j$, where $\Omega_1 = \Big\{\|\widetilde{X}'_{\mathcal{O}^c}(\tilde{\boldsymbol{y}} - \widetilde{X}\hat{\boldsymbol{\beta}}^o)/n\|_\infty \le \lambda_1\Big\}$ is as in (8.1) and

$$
\Omega_2 = \Big\{\min_{j\in\mathcal{O}}\mathrm{sgn}(\beta^*_j)\hat{\beta}^o_j > \gamma(2\sqrt{c^*}\lambda_1)\Big\},\quad \Omega_3 = \Big\{\zeta(\tilde{\boldsymbol{y}} - \widetilde{X}\boldsymbol{\beta}^*; d^* - |\mathcal{O}|, \mathcal{O}, \lambda_2) \le \lambda_1\Big\}.
$$

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 $\tilde{\boldsymbol{\varepsilon}}^* = \tilde{\boldsymbol{y}} - \widetilde{X}\boldsymbol{\beta}^* = \tilde{\boldsymbol{\varepsilon}} + E\tilde{\boldsymbol{\varepsilon}}^*$ with $\tilde{\boldsymbol{\varepsilon}} = \tilde{\boldsymbol{y}} - E\tilde{\boldsymbol{y}}$. Since $\widetilde{X}'\tilde{\boldsymbol{y}} = X'\boldsymbol{y}$, (8.2) still holds with $\|\widetilde{X}'E\tilde{\boldsymbol{\varepsilon}}^*/n\|_\infty = \lambda_2 C_2$. Since $\widetilde{X}'\tilde{\boldsymbol{\varepsilon}} = X'\boldsymbol{y} - EX'\boldsymbol{y} = X'\boldsymbol{\varepsilon}$, (8.2) still gives (8.3). A slight modification of the argument for (8.4) yields

$$\Omega_2^c \subseteq \cup_{j\in\mathcal{O}}\Big\{\mathrm{sgn}(\beta_j^*)\boldsymbol{u}_j'\boldsymbol{\varepsilon} \geq (n/v_j)^{1/2}\big(|\beta_j^*| - \gamma(2\sqrt{c^*}\lambda_1)\big) \geq \sigma\sqrt{2\log(|\mathcal{O}|/\epsilon)}\Big\}. \tag{8.7}$$

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 $\|\widetilde{P}_B\tilde{\boldsymbol{\varepsilon}}\|/\sqrt{n} = \|\Sigma_B^{-1/2}(\lambda_2)\widetilde{X}_B'\tilde{\boldsymbol{\varepsilon}}\|/n \leq \|X_B'\boldsymbol{\varepsilon}/n\|_\infty\sqrt{|B|/c_*(\lambda_2)}$. Thus, by (8.6)

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

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

$$\Omega_3 \subseteq \big\{\|X'\boldsymbol{\varepsilon}/n\|_\infty < \sqrt{c_*(\lambda_2)K_*/(K_*+1)}\lambda_1 - \lambda_2 C_2\big\}. \tag{8.8}$$

Since $\max\{1, \sqrt{c_*(\lambda_2)K_*/(K_*+1)}\}\lambda_1 \geq \lambda_2 C_2 + \sigma\sqrt{2\log(p/\epsilon)}\max_{j\leq p}\|\boldsymbol{x}_j\|/n$, (8.3), (8.7), (8.8) and Condition (A) imply

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

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). □

Proof of Proposition 1. The $\tilde{\boldsymbol{\beta}}$ satisfies

$$-\frac{1}{n}\boldsymbol{x}_j'(\boldsymbol{y} - X\tilde{\boldsymbol{\beta}}) + \lambda_2(d_j\tilde{\beta}_j - \boldsymbol{a}_j'\tilde{\boldsymbol{\beta}}) = 0,\ 1 \leq j \leq q. \tag{8.9}$$

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

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

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

For part (ii), note that we have

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

Thus

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

Part (ii) follows. □.

Proof of Proposition 2. The $\tilde{\boldsymbol{\beta}}$ must satisfy

$$-\frac{1}{n}\boldsymbol{x}_j'(\boldsymbol{y} - X\tilde{\boldsymbol{\beta}}) + \lambda_2(\tilde{\beta}_j - v_g^{-1}\mathbf{1}_g'\tilde{\boldsymbol{\beta}}_g) = 0,\ j \in V_g,\ 1 \leq g \leq J. \tag{8.10}$$

Taking the difference between the jth and kth equations in (8.10) for $j, k \in V_g$, we get

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

Therefore,

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

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 $k \in V_h, g \neq h$, we have

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

Thus part (ii) follows. This completes the proof of Proposition 2. □

References

[1] Bondell, H. D. and Reich, B. J. (2008). Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR. *Biometrics*, **64**, 115-123.

[2] Breheny, P. and Huang, J. (2009). Coordinate descent algorithms for nonconvex penalized regression methods. *Technical Report #403*, Department of Biostatistics, University of Kentukey.

[3] Chen, S. S., Donoho, D. L. & Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, **20**, 3361.

[4] 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). *Proceedings of the National Academy of Sciences*, **103**, 6287-6292.

[5] Chung, F. R. K. (1997). *Spectral Graph Theory. CBMS Regional Conference Series in Mathematics*, No. 92. Amer. Math. Soc.

[6] Chung, F. R. K. and Lu, L. (2006). *Complex Graphs and Networks. CBMS Regional Conference Series in Mathematics*, No. 107. Amer. Math. Soc.

[7] Daye, J. Z. and Jeng, J. X. (2009). Shrinkage and model selection with correlated variables via weighted fusion. *Computational Statistics and Data Analysis*. **53**, 1284-1298.

[8] Fan, J. (1997). Comments on "Wavelets in statistics: a review" by A. Antoniadis. *J. Italian Statist. Assoc.* **6**, 131-138.

[9] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348-1360.

[10] Fan, J., Feng, Y. and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *Ann. Appl. Statist.*, **3**, 521-541.

[11] Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, **35**, 109-148.

[12] Friedman, J., Hastie, Hoefling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.*, **35**, 302-332.

[13] Friedman J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatist.*, **9**, 432-441.

[14] Fu, W. J. (1998). Penalized regressions: the bridge versus the LASSO. *J. Comp. Graph. Statist.* **7**, 397-416.

[15] Genkin, A. Lewis, D.D. and Madigan, D. (2004). Large-scale Bayesian logistic regression for text categorization. Technical Report, DIMACS, Rutgers University.

[16] 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*.

[17] Huang, J., Breheny, P., Ma, S. and Zhang, C.-H. (2010). The Mnet method for variable selection. *Technical report # 402*, Department of Statistics and Actuarial Science, University of Iowa.

[18] 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.

[19] Jia, J. & Yu, B. (2010). On model selection consistency of elastic net when $p \gg n$. *Statistica Sinica*, **20**, 595-611.

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

[21] Li, C. and Li, H. (2010). Variable seleciton and regression analysis for covariates with graphical structure. Accepted for publication by *Ann. Appl. Statist.*.

[22] Mazumder, R., Friedman, J. & Hastie, T. (2009). *SparseNet*: Coordinate descent with non-convex penalties. *Tech Report*. Department of Statistics, Stanford University.

[23] Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, **34**, 1436-1462.

[24] Pan, W., Xie, B. and Shen, X. (2009). Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*. In press.

[25] Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp1, 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. *Proceedings of the National Academy of Sciences*, **103**, 14429-14434.

[26] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. B*, **58**, 267-288.

[27] Tutz, G. and Ulbricht, J. (2007). Penalized regression with correlation-based penalty. *Statist. Comput.* **19**, 239-253.

[28] Wu, T. & Lange, K. (2007). Coordinate descent procedures for lasso penalized regression. *Ann. Appl. Statist.* **2**, 224-244.

[29] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, **68**, 49-67.

[30] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**, 19-35.

[31] Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statist. Appl. Genet. Mol. Bio.*, **4**, article 17.

[32] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894-942.

[33] Zhao, P. and Yu, B. (2004). Boosted Lasso. Technical report, Department of Statistics, UC Berkeley, 2004. URL *http://www.stat.berkeley.edu/users/binyu/ps/blasso.ps*.

[34] Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. *J. Machine Learning Res.*, **7**, 2541 - 2563.

[35] Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, **67**, 301-320.

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

Jian Huang,

Department of Statistics and Actuarial Science, 241 SH

University of Iowa

Iowa City, Iowa 52242.

E-mail: jian-huang@uiowa.edu

Shuangge Ma

Division of Biostatistics

School of Public Health

Yale University

New Haven, CT 06520 Email: shuangge.ma@yale.edu

Hongzhe Li

Department of Biostatistics and Epidemiology

University of Pennsylvania School of Medicine

Philadelphia, PA 19104 Email: hongzhe@upenn.edu

Cun-Hui Zhang

Department of Statistics and Biostatistics

Rutgers University

Piscataway, NJ 08854

Email: cunhui@stat.rutgers.edu