Journal of Machine Learning Research 14 (2013) 3385-3418 Submitted 2/12; Revised 8/13; Published 11/13

Sparse Matrix Inversion with Scaled Lasso

Tingni Sun TINGNI@WHARTON.UPENN.EDU
Statistics Department
The Wharton School
University of Pennsylvania
Philadelphia, Pennsylvania, 19104, USA

Cun-Hui Zhang CZHANG@STAT.RUTGERS.EDU
Department of Statistics and Biostatistics
Rutgers University
Piscataway, New Jersey 08854, USA

Editor: Bin Yu

Abstract

We propose a new method of learning a sparse nonnegative-definite target matrix. Our primary example of the target matrix is the inverse of a population covariance or correlation matrix. The algorithm first estimates each column of the target matrix by the scaled Lasso and then adjusts the matrix estimator to be symmetric. The penalty level of the scaled Lasso for each column is completely determined by data via convex minimization, without using cross-validation.

We prove that this scaled Lasso method guarantees the fastest proven rate of convergence in the spectrum norm under conditions of weaker form than those in the existing analyses of other ℓ_1 regularized algorithms, and has faster guaranteed rate of convergence when the ratio of the ℓ_1 and spectrum norms of the target inverse matrix diverges to infinity. A simulation study demonstrates the computational feasibility and superb performance of the proposed method.

Our analysis also provides new performance bounds for the Lasso and scaled Lasso to guarantee higher concentration of the error at a smaller threshold level than previous analyses, and to allow the use of the union bound in column-by-column applications of the scaled Lasso without an adjustment of the penalty level. In addition, the least squares estimation after the scaled Lasso selection is considered and proven to guarantee performance bounds similar to that of the scaled Lasso.

Keywords: precision matrix, concentration matrix, inverse matrix, graphical model, scaled Lasso, linear regression, spectrum norm

1. Introduction

We consider the estimation of the matrix inversion Θ^* satisfying $\overline{\Sigma}\Theta^* \approx I$ for a given data matrix $\overline{\Sigma}$. When $\overline{\Sigma}$ is a sample covariance matrix, our problem is the estimation of the inverse of the corresponding population covariance matrix. The inverse covariance matrix is also called precision matrix or concentration matrix. With the dramatic advances in technology, the number of variables p, or the size of the matrix Θ^*, is often of greater order than the sample size n in statistical and engineering applications. In such cases, the sample covariance matrix is always singular and a certain type of sparsity condition is typically imposed for proper estimation of the precision matrix and for theoretical investigation of the problem. In a simple version of our theory, this condition is expressed as the ℓ_0 sparsity, or equivalently the maximum degree, of the target inverse matrix Θ^*. A weaker condition of capped ℓ_1 sparsity is also studied to allow many small signals.

Several approaches have been proposed to the estimation of sparse inverse matrices in high-dimensional setting. The ℓ_1 penalization is one of the most popular methods. Lasso-type methods, or convex minimization algorithms with the ℓ_1 penalty on all entries of Θ^*, have been developed in Banerjee et al. (2008) and Friedman et al. (2008), and in Yuan and Lin (2007) with ℓ_1 penalization on the off-diagonal matrix only. This is refereed to as the graphical Lasso (GLasso) due to the connection of the precision matrix to Gaussian Markov graphical models. In this GLasso framework, Ravikumar et al. (2008) provides sufficient conditions for model selection consistency, while Rothman et al. (2008) provides the convergence rate $\{((p+s)/n)\log p\}^{1/2}$ in the Frobenius norm and $\{(s/n)\log p\}^{1/2}$ in the spectrum norm, where s is the number of nonzero off-diagonal entries in the precision matrix. Concave penalty has been studied to reduce the bias of the GLasso (Lam and Fan, 2009). Similar convergence rates have been studied under the Frobenius norm in a unified framework for penalized estimation in Negahban et al. (2012). Since the spectrum norm can be controlled via the Frobenius norm, this provides a sufficient condition $(s/n)\log p \to 0$ for the convergence to the unknown precision matrix under the spectrum norm. However, in the case of $p \geq n$, this condition does not hold for banded precision matrices, where s is of the order of the product of p and the width of the band.

A potentially faster rate $d\sqrt{(\log p)/n}$ can be achieved by ℓ_1 regularized estimation of individual columns of the precision matrix, where d, the matrix degree, is the largest number of nonzero entries in a column. This was done in Yuan (2010) by applying the Dantzig selector to the regression of each variable against others, followed by a symmetrization step via linear programming. When the ℓ_1 operator norm of the precision matrix is bounded, this method achieves the convergence rate $d\sqrt{(\log p)/n}$ in ℓ_q matrix operator norms. The CLIME estimator (Cai et al., 2011), which uses the Dantzig selector directly to estimate each column of the precision matrix, also achieves the $d\sqrt{(\log p)/n}$ rate under the boundedness assumption of the ℓ_1 operator norm. In Yang and Kolaczyk (2010), the Lasso is applied to estimate the columns of the target matrix under the assumption of equal diagonal, and the estimation error is studied in the Frobenius norm for $p = n^{\nu}$. This column-by-column idea reduces a graphical model to p regression models. It was first introduced by Meinshausen and Bühlmann (2006) for identifying nonzero variables in a graphical model, called neighborhood selection. In addition, Rocha et al. (2008) proposed a pseudo-likelihood method by merging all p linear regressions into a single least squares problem.

In this paper, we propose to apply the scaled Lasso (Sun and Zhang, 2012) column-by-column to estimate a precision matrix in the high dimensional setting. Based on the connection of precision matrix estimation to linear regression, we construct a column estimator with the scaled Lasso, a joint estimator for the regression coefficients and noise level. Since we only need a sample covariance matrix as input, this estimator could be extended to generate an approximate inverse of a nonnegative-definite data matrix in a more general setting. This scaled Lasso algorithm provides a fully specified map from the space of nonnegative-definite matrices to the space of symmetric matrices. For each column, the penalty level of the scaled Lasso is determined by data via convex minimization, without using cross-validation.

We study theoretical properties of the proposed estimator for a precision matrix under a normality assumption. More precisely, we assume that the data matrix is the sample covariance matrix $\overline{\Sigma} = X^T X/n$, where the rows of X are iid $N(0,\Sigma^*)$ vectors. Let $R^* = (\mathrm{diag}\Sigma^*)^{-1/2}\Sigma^*(\mathrm{diag}\Sigma^*)^{-1/2}$ be the population correlation matrix. Our target is to estimate the inverse matrices $\Theta^* = (\Sigma^*)^{-1}$ and

$\Omega^* = (R^*)^{-1}$. Define

$$d = \max_{1\le j\le p} \#\{k : \Theta^*_{jk} \ne 0\}. \tag{1}$$

A simple version of our main theoretical result can be stated as follows.

Theorem 1 *Let $\widehat{\Theta}$ and $\widehat{\Omega}$ be the scaled Lasso estimators defined in (4), (7) and (9) below with penalty level $\lambda_0 = A\sqrt{4(\log p)/n}$, $A > 1$, based on n iid observations from $N(0,\Sigma^*)$. Suppose the spectrum norm of $\Omega^* = (\text{diag}\Sigma^*)^{1/2}\Theta^*(\text{diag}\Sigma^*)^{1/2}$ is bounded and that $d^2(\log p)/n \to 0$. Then,*

$$\|\widehat{\Omega} - \Omega^*\|_2 = O_P(1)d\sqrt{(\log p)/n} = o(1),$$

where $\|\cdot\|_2$ is the spectrum norm (the ℓ_2 matrix operator norm). If in addition the diagonal elements of Θ^* is uniformly bounded, then

$$\|\widehat{\Theta} - \Theta^*\|_2 = O_P(1)d\sqrt{(\log p)/n} = o(1).$$

Theorem 1 provides a simple boundedness condition on the spectrum norm of Ω^* for the convergence of $\widehat{\Omega}$ in spectrum norm with sample size $n \gg d^2 \log p$. The additional condition on the diagonal of Θ^* is natural due to scale change. The boundedness condition on the spectrum norm of $(\text{diag}\Sigma^*)^{1/2}\Theta^*(\text{diag}\Sigma^*)^{1/2}$ and the diagonal of Θ^* is weaker than the boundedness of the ℓ_1 operator norm assumed in Yuan (2010) and Cai et al. (2011) since the boundedness of $\text{diag}\Sigma^*$ is also needed there. When the ratio of the ℓ_1 operator norm and spectrum norm of the precision matrix diverges to infinity, the proposed estimator has a faster proven convergence rate. This sharper result is a direct consequence of the faster convergence rate of the scaled Lasso estimator of the noise level in linear regression. To the best of our knowledge, it is unclear if the ℓ_1 regularization method of Yuan (2010) and Cai et al. (2011) also achieve the convergence rate under the weaker spectrum norm condition.

An important advantage of the scaled Lasso is that the penalty level is automatically set to achieve the optimal convergence rate in the regression model for the estimation of each column of the inverse matrix. This raises the possibility for the scaled Lasso to outperform methods using a single unscaled penalty level for the estimation of all columns such as the GLasso and CLIME. We provide an example in Section 7 to demonstrate the feasibility of such a scenario.

Another contribution of this paper is to study the scaled Lasso at a smaller penalty level than those based on ℓ_∞ bounds of the noise. The ℓ_∞-based analysis requires a penalty level λ_0 satisfying $P\{N(0,1/n) > \lambda_0/A\} = \varepsilon/p$ for a small ε and $A > 1$. For $A \approx 1$ and $\varepsilon = p^{o(1)}$, this penalty level is comparable to the universal penalty level $\sqrt{(2/n)\log p}$. However, $\varepsilon = o(1/p)$, or equivalently $\lambda_0 \approx \sqrt{(4/n)\log p}$, is required if the union bound is used to simultaneously control the error of p applications of the scaled Lasso in the estimation of individual columns of a precision matrix. This may create a significant gap between theory and implementation. We close this gap by providing a theory based on a sparse ℓ_2 measure of the noise, corresponding to a penalty level satisfying $P\{N(0,1/n) > \lambda_0/A\} = k/p$ with $A > 1$ and a potentially large k. This penalty level provides a faster convergence rate than the universal penalty level in linear regression when $\log(p/k) \approx \log(p/\|\beta\|_0) \ll \log p$. Moreover, the new analysis provides a higher concentration of the error so that the same penalty level $\lambda_0 \approx \sqrt{(2/n)\log(p/k)}$ can be used to simultaneously control the estimation error in p applications of the scaled Lasso for the estimation of a precision matrix.

The rest of the paper is organized as follows. In Section 2, we present the scaled Lasso method for the estimation of the inversion of a nonnegative definite matrix. In Section 3, we study the estimation error of the proposed method. In Section 4, we provide a theory for the Lasso and its scaled version with higher proven concentration at a smaller, practical penalty level. In Section 5, we study the least square estimation after the scaled Lasso selection. Simulation studies are presented in Section 6. In Section 7, we discuss the benefits of using the scaled penalty levels for the estimation of different columns of the precision matrix, compared with an optimal fixed penalty level for all columns. An appendix provides all the proofs.

We use the following notation throughout the paper. For real x, $x_+ = \max(x,0)$. For a vector $v = (v_1,\dots,v_p)$, $\|v\|_q = (\sum_j |v_j|^q)^{1/q}$ is the ℓ_q norm with the special $\|v\| = \|v\|_2$ and the usual extensions $\|v\|_\infty = \max_j |v_j|$ and $\|v\|_0 = \#\{j : v_j \neq 0\}$. For matrices M, $M_{i,*}$ is the i-th row and $M_{*,j}$ the j-th column, $M_{A,B}$ represents the submatrix of M with rows in A and columns in B, $\|M\|_q = \sup_{\|v\|_q=1} \|Mv\|_q$ is the ℓ_q matrix operator norm. In particular, $\|\cdot\|_2$ is the spectrum norm for symmetric matrices. Moreover, we may denote the set $\{j\}$ by j and denote the set $\{1,\dots,p\}\setminus\{j\}$ by $-j$ in the subscript.

2. Matrix Inversion via Scaled Lasso

Let $\overline{\Sigma}$ be a nonnegative-definite data matrix and Θ^* be a positive-definite target matrix with $\overline{\Sigma}\Theta^* \approx I$. In this section, we describe the relationship between positive-definite matrix inversion and linear regression and propose an estimator for Θ^* via scaled Lasso, a joint convex minimization for the estimation of regression coefficients and noise level.

We use the scaled Lasso to estimate Θ^* column by column. Define $\sigma_j > 0$ and $\beta \in \mathbb{R}^{p\times p}$ by

$$\sigma_j^2 = (\Theta^*_{jj})^{-1}, \quad \beta_{*,j} = -\Theta^*_{*,j}\sigma_j^2 = -\Theta^*_{*,j}(\Theta^*_{jj})^{-1}.$$

In the matrix form, we have the following relationship

$$\mathrm{diag}\Theta^* = \mathrm{diag}(\sigma_j^{-2}, j = 1,\dots,p), \quad \Theta^* = -\beta(\mathrm{diag}\Theta^*). \tag{2}$$

Let $\Sigma^* = (\Theta^*)^{-1}$. Since $(\partial/\partial b_{-j})b^T\Sigma^* b = 2\Sigma^*_{-j,*}b = 0$ at $b = \beta_{*,j}$, one may estimate the j-th column of β by minimizing the ℓ_1 penalized quadratic loss. In order to penalize the unknown coefficients in the same scale, we adjust the ℓ_1 penalty with diagonal standardization, leading to the following penalized quadratic loss:

$$b^T\overline{\Sigma}b/2 + \lambda\sum_{k=1}^p \overline{\Sigma}_{kk}^{1/2}|b_k|. \tag{3}$$

For $\overline{\Sigma} = X^TX/n$ and $b_j = -1$, $b^T\overline{\Sigma}b = \|x_j - \sum_{k\neq j} b_k x_k\|_2^2/n$, so that (3) is the penalized loss for the Lasso in linear regression of x_j against $\{x_k, k \neq j\}$. This is similar to the procedures in Yuan (2010) and Cai et al. (2011) that use the Dantzig selector to estimate $\Theta^*_{*,j}$ column-by-column. However, one still needs to choose a penalty level λ and to estimate σ_j in order to recover Θ^* via (2). A solution to resolve these two issues is the scaled Lasso (Sun and Zhang, 2012):

$$\{\widehat{\beta}_{*,j}, \widehat{\sigma}_j\} = \mathop{\arg\min}_{b,\sigma}\Big\{\frac{b^T\overline{\Sigma}b}{2\sigma} + \frac{\sigma}{2} + \lambda_0\sum_{k=1}^p \overline{\Sigma}_{kk}^{1/2}|b_k| : b_j = -1\Big\} \tag{4}$$

with $\lambda_0 \approx \sqrt{(2/n)\log p}$. The scaled Lasso (4) is a solution of joint convex minimization in $\{b,\sigma\}$ (Huber and Ronchetti, 2009; Antoniadis, 2010). Since $\beta^T\Sigma^*\beta = (\text{diag}\Theta^*)^{-1}\Theta^*(\text{diag}\Theta^*)^{-1}$,

$$\text{diag}\big(\beta^T\Sigma^*\beta\big) = (\text{diag}\Theta^*)^{-1} = \text{diag}(\sigma_j^2, j=1,\ldots,p).$$

Thus, (4) is expected to yield consistent estimates of $\sigma_j = (\Theta^*_{jj})^{-1/2}$.

An iterative algorithm has been provided in Sun and Zhang (2012) to compute the scaled Lasso estimator (4). We rewrite the algorithm in the form of matrices. For each $j \in \{1,\ldots,p\}$, the Lasso path is given by the estimates $\widehat{\beta}_{-j,j}(\lambda)$ satisfying the following Karush-Kuhn-Tucker conditions: for all $k \neq j$,

$$\begin{cases} \overline{\Sigma}_{kk}^{-1/2}\overline{\Sigma}_{k,*}\widehat{\beta}_{*,j}(\lambda) = -\lambda\text{sgn}(\widehat{\beta}_{k,j}(\lambda)), & \widehat{\beta}_{k,j} \neq 0, \\ \overline{\Sigma}_{kk}^{-1/2}\overline{\Sigma}_{k,*}\widehat{\beta}_{*,j}(\lambda) \in \lambda[-1,1], & \widehat{\beta}_{k,j} = 0, \end{cases} \tag{5}$$

where $\widehat{\beta}_{jj}(\lambda) = -1$. Based on the Lasso path $\widehat{\beta}_{*,j}(\lambda)$, the scaled Lasso estimator $\{\widehat{\beta}_{*,j}, \widehat{\sigma}_j\}$ is computed iteratively by

$$\widehat{\sigma}_j^2 \leftarrow \widehat{\beta}_{*,j}^T\overline{\Sigma}\widehat{\beta}_{*,j}, \quad \lambda \leftarrow \widehat{\sigma}_j\lambda_0, \quad \widehat{\beta}_{*,j} \leftarrow \widehat{\beta}_{*,j}(\lambda). \tag{6}$$

Here the penalty level of the Lasso is determined by the data without using cross-validation. We then simply take advantage of the relationship (2) and compute the coefficients and noise levels by the scaled Lasso for each column

$$\text{diag}\widetilde{\Theta} = \text{diag}(\widehat{\sigma}_j^{-2}, j=1,\ldots,p), \quad \widetilde{\Theta} = -\widehat{\beta}(\text{diag}\widetilde{\Theta}). \tag{7}$$

Now we have constructed an estimator for Θ^*. In our primary example of taking $\overline{\Sigma}$ as a sample covariance matrix, the target Θ^* is the inverse covariance matrix. One may also be interested in estimating the inverse correlation matrix

$$\Omega^* = (R^*)^{-1} = \big\{D^{-1/2}\Sigma^* D^{-1/2}\big\}^{-1} = D^{1/2}(\Sigma^*)^{-1}D^{1/2}, \tag{8}$$

where $D = \text{diag}(\Sigma^*)$ and $R^* = D^{-1/2}\Sigma^* D^{-1/2}$ is the population correlation matrix. The diagonal matrix D can be approximated by the diagonal of $\overline{\Sigma}$. Thus, the inverse correlation matrix is estimated by

$$\widetilde{\Omega} \;=\; \widehat{D}^{1/2}\widetilde{\Theta}\widehat{D}^{1/2} \text{ with } \widehat{D} = \text{diag}(\overline{\Sigma}_{jj}, j=1,\ldots,p).$$

The estimator $\widetilde{\Omega}$ here is a result of normalizing the precision matrix estimator by the population variances. Alternatively, we may estimate the inverse correlation matrix by using the population correlation matrix

$$\overline{R} = (\text{diag}\overline{\Sigma})^{-1/2}\overline{\Sigma}(\text{diag}\overline{\Sigma})^{-1/2} = \widehat{D}^{-1/2}\overline{\Sigma}\widehat{D}^{-1/2}$$

as data matrix. Let $\{\widehat{\alpha}_{*,j}, \widehat{\tau}_j\}$ be the solution of (4) with $\overline{R}$ in place of $\overline{\Sigma}$. We combine these column estimators as in (7) to have an alternative estimator for Ω^* as follows:

$$\text{diag}\big(\widetilde{\Omega}^{\text{Alt}}\big) = \text{diag}(\widehat{\tau}_j^{-2}, j=1,\ldots,p), \quad \widetilde{\Omega}^{\text{Alt}} = -\widehat{\alpha}\,\text{diag}\big(\widetilde{\Omega}^{\text{Alt}}\big).$$

Since $\overline{R}_{jj} = 1$ for all j, it follows from (4) that

$$\widehat{\alpha}_{*,j} = \widehat{D}^{1/2}\widehat{\beta}_{*,j}\widehat{D}_{jj}^{-1/2}, \quad \widehat{\tau}_j = \widehat{\sigma}_j\widehat{D}_{jj}^{-1/2}.$$

This implies

$$\widetilde{\Omega}^{\mathrm{Alt}} = -\widehat{D}^{1/2}\widehat{\beta}\mathrm{diag}(\widehat{D}_{jj}^{-1/2}\widehat{\sigma}_j^{-2}\widehat{D}_{jj}, j=1,\ldots,p) = \widehat{D}^{1/2}\widetilde{\Theta}\widehat{D}^{1/2} = \widetilde{\Omega}.$$

Thus, in this scaled Lasso approach, the estimator based on the normalized data matrix is exactly the same as the one based on the original data matrix followed by a normalization step. The scaled Lasso methodology is scale-free in the noise level, and as a result, the estimator for inverse correlation matrix is also scale free in diagonal normalization.

It is noticed that a good estimator for Θ^* or Ω^* should be a symmetric matrix. However, the estimators $\widetilde{\Theta}$ and $\widetilde{\Omega}$ do not have to be symmetric. We improve them by using a symmetrization step as in Yuan (2010),

$$\widehat{\Theta} = \mathop{\arg\min}_{M:M^T=M} \|M - \widetilde{\Theta}\|_1, \quad \widehat{\Omega} = \mathop{\arg\min}_{M:M^T=M} \|M - \widetilde{\Omega}\|_1, \tag{9}$$

which can be solved by linear programming. It is obvious that $\widehat{\Theta}$ and $\widehat{\Omega}$ are both symmetric, but not guaranteed to be positive-definite. It follows from Theorem 1 that $\widehat{\Theta}$ and $\widehat{\Omega}$ are positive-definite with large probability. Alternatively, semidefinite programming, which is somewhat more expensive computationally, can be used to produce a nonnegative-definite $\widehat{\Theta}$ and $\widehat{\Omega}$ in (9).

According to the definition, the estimators $\widehat{\Theta}$ and $\widehat{\Omega}$ have the same ℓ_1 error rate as $\widetilde{\Theta}$ and $\widetilde{\Omega}$ respectively. A nice property of symmetric matrices is that the spectrum norm is bounded by the ℓ_1 matrix norm. The ℓ_1 matrix norm can be expressed more explicitly as the maximum ℓ_1 norm of the columns, while the ℓ_∞ matrix norm is the maximum ℓ_1 norm of the rows. Hence, for any symmetric matrix, the ℓ_1 matrix norm is equivalent to the ℓ_∞ matrix norm, and the spectrum norm can be bounded by either of them. Since our estimators and target matrices are all symmetric, the error bound based on the spectrum norm could be studied by bounding the ℓ_1 error as typically done in the existing literature. We will study the estimation error of (9) in Section 3.

To sum up, we propose to estimate the matrix inversion by (4), (7) and (9). The iterative algorithm (6) computes (4) based on a Lasso path determined by (5). Then (7) translates the resulting estimators of (6) to column estimators and thus a preliminary matrix estimator is constructed. Finally, the symmetrization step (9) produces a symmetric estimate for our target matrix.

3. Theoretical Properties

From now on, we suppose that the data matrix is the sample covariance matrix $\overline{\Sigma} = X^TX/n$, where the rows of X are iid $N(0,\Sigma^*)$. Let $\Theta^* = (\Sigma^*)^{-1}$ be the precision matrix as the inverse of the population covariance matrix. Let D be the diagonal of Σ^*, $R^* = D^{-1/2}\Sigma^*D^{-1/2}$ the population correlation matrix, $\Omega^* = (R^*)^{-1}$ its inverse as in (8). In this section, we study $\widehat{\Omega}$ and $\widehat{\Theta}$ in (9), respectively for the estimation of Ω^* and Θ^*.

We consider a certain capped ℓ_1 sparsity for individual columns of the inverse matrix as follows. For a certain $\varepsilon_0 > 0$, a threshold level $\lambda_{*,0} > 0$ not depending on j and an index set $S_j \subset \{1,\ldots,p\} \setminus \{j\}$, the capped ℓ_1 sparsity condition measures the complexity of the j-th column of Ω^* by

$$|S_j| + (1-\varepsilon_0)^{-1} \sum_{k\neq j, k\notin S_j} \frac{|\Omega^*_{kj}|}{(\Omega^*_{jj})^{1/2}\lambda_{*,0}} \leq s_{*,j}. \tag{10}$$

The condition can be written as

$$\sum_{j\neq k}\min\left\{\frac{|\Omega^*_{kj}|}{(1-\varepsilon_0)(\Omega^*_{jj})^{1/2}\lambda_{*,0}},1\right\}\le s_{*,j}$$

if we do not care about the choice of S_j. In the ℓ_0 sparsity case of $S_j=\{k:k\neq j,\Omega^*_{kj}\neq 0\}$, we may set $s_{*,j}=|S_j|+1$ as the degree for the j-th node in the graph induced by matrix Ω^* (or Θ^*). In this case, $d=\max_j(1+|S_j|)$ is the maximum degree as in (1).

In addition to the sparsity condition on the inverse matrix, we also require a certain invertibility condition on R^*. Let $S_j\subseteq B_j\subseteq\{1,\ldots,p\}\setminus\{j\}$. A simple version of the required invertibility condition can be written as

$$\inf\left\{\frac{u^T(R^*_{-j,-j})u}{\|u_{B_j}\|_2^2}:u_{B_j}\neq 0\right\}\ge c_* \tag{11}$$

with a fixed constant $c_*>0$. This condition requires a certain partial invertibility of the population correlation matrix. It certainly holds if the smallest eigenvalue of $R^*_{-j,-j}$ is no smaller than c_* for all $j\le p$, or the spectrum norm of Ω^* is no greater than $1/c_*$ as assumed in Theorem 1. In the proof of Theorems 2 and 3, we only use a weaker version of condition (11) in the form of (35) with $\{\Sigma^*,\overline{\Sigma}\}$ replaced by $\{R^*_{-j,-j},\overline{R}_{-j,-j}\}$ there.

Theorem 2 *Suppose $\overline{\Sigma}$ is the sample covariance matrix of n iid $N(0,\Sigma^*)$ vectors. Let $\Theta^*=(\Sigma^*)^{-1}$ and Ω^* as in (8) be the inverses of the population covariance and correlation matrices. Let $\widehat{\Theta}$ and $\widehat{\Omega}$ be their scaled Lasso estimators defined in (4), (7) and (9) with a penalty level $\lambda_0=A\sqrt{4(\log p)/n}$, $A>1$. Suppose (10) and (11) hold with $\varepsilon_0=0$ and $\max_{j\le p}(1+s_{*,j})\lambda_0\le c_0$ for a certain constant $c_0>0$ depending on c_* only. Then, the spectrum norm of the errors are bounded by*

$$\|\widehat{\Theta}-\Theta^*\|_2\le\|\widehat{\Theta}-\Theta^*\|_1\le C\Big(\max_{j\le p}(\|D^{-1}_{-j}\|_\infty\Theta^*_{jj})^{1/2}s_{*,j}\lambda_0+\|\Theta^*\|_1\lambda_0\Big), \tag{12}$$

$$\|\widehat{\Omega}-\Omega^*\|_2\le\|\widehat{\Omega}-\Omega^*\|_1\le C\Big(\max_{j\le p}(\Omega^*_{jj})^{1/2}s_{*,j}\lambda_0+\|\Omega^*\|_1\lambda_0\Big), \tag{13}$$

with large probability, where C is a constant depending on $\{c_0,c_*,A\}$ only. Moreover, the term $\|\Theta^*\|_1\lambda_0$ in (12) can be replaced by

$$\max_{j\le p}\|\Theta^*_{*,j}\|_1 s_{*,j}\lambda_0^2+\tau_n(\Theta^*), \tag{14}$$

where $\tau_n(M)=\inf\{\tau:\sum_j\exp(-n\tau^2/\|M_{*,j}\|_1^2)\le 1/e\}$ for a matrix M.

Theorem 2 implies Theorem 1 due to $s_{*,j}\le d-1$, $1/D_{jj}\le\Theta^*_{jj}\le\|\Theta^*\|_2$, $\|\Theta^*\|_1\le d\max_j\Theta^*_{jj}$ and similar inequalities for Ω^*. We note that $B_j=S_j$ in (11) gives the largest c_* and thus the sharpest error bounds in Theorem 2. In Section 7, we give an example to demonstrate the advantage of this theorem.

In a 2011 arXiv version of this paper (http://arxiv.org/pdf/1202.2723v1.pdf), we are able to demonstrate good numerical performance of the scaled Lasso estimator with the universal penalty level $\lambda_0=\sqrt{2(\log p)/n}$, compared with some existing methods, but not the larger penalty level $\lambda_0>\sqrt{4(\log p)/n}$ in Theorems 1 and 2. Since a main advantage of our proposal is automatic selection of the penalty level without resorting to cross validation, a question arises as to whether a theory can be developed for a smaller penalty level to match the choice in a demonstration of good performance of the scaled Lasso in our simulation experiments.

We are able to provide an affirmative answer in this version of the paper by proving a higher concentration of the error of the scaled Lasso at a smaller penalty level as follows. Let $L_n(t)$ be the $N(0,1/n)$ quantile function satisfying

$$P\big\{N(0,1) > n^{1/2}L_n(t)\big\} = t.$$

Our earlier analysis is based on existing oracle inequalities of the Lasso which holds with probability $1-2\varepsilon$ when the inner product of design vectors and noise are bounded by their ε/p and $1-\varepsilon/p$ quantiles. Application of the union bound in p applications of the Lasso requires a threshold level $\lambda_{*,0} = L_n(\varepsilon/p^2)$ with a small $\varepsilon > 0$, which matches $\sqrt{4(\log p)/n}$ with $\varepsilon \asymp 1/\sqrt{\log p}$ in Theorems 1 and 2. Our new analysis of the scaled Lasso allows a threshold level

$$\lambda_{*,0} = L_{n-3/2}(k/p)$$

with $k \asymp s\log(p/s)$, where $s = 1+\max_j s_{*,j}$. More precisely, we require a penalty level $\lambda_0 \geq A\lambda_{*,0}$ with a constant A satisfying

$$A-1 > A_1 \geq \max_j \Big\{ \Big[\frac{e^{1/(4n-6)^2}4k}{m_j(L^4+2L^2)}\Big]^{1/2} + \frac{e^{1/(4n-6)^2}}{L\sqrt{2\pi}}\sqrt{\psi_j} + \frac{L_1(\varepsilon/p)}{L}\sqrt{\psi_j}\Big\}, \tag{15}$$

where $L = L_1(k/p)$, $s_{*,j} \leq m_j \leq \min(|B_j|, C_0 s_{*,j})$ with the $s_{*,j}$ and B_j in (10) and (11), and $\psi_j = \kappa_+(m_j; R^*_{-j,-j})/m_j + L_n(5\varepsilon/p^2)$ with

$$\kappa_+(m;\Sigma) = \max_{\|u\|_0=m, \|u\|_2=1} u^T\Sigma u. \tag{16}$$

Theorem 3 *Let $\{\overline{\Sigma}, \Sigma^*, \Theta^*, \Omega^*\}$ be matrices as in Theorem 2, and $\widehat{\Theta}$ and $\widehat{\Omega}$ be the scaled Lasso estimators with a penalty level $\lambda_0 \geq A\lambda_{*,0}$ where $\lambda_{*,0} = L_{n-3/2}(k/p)$. Suppose (10) and (11) hold with certain $\{S_j, s_{*,j}, \varepsilon_0, B_j, c_*\}$, (15) holds with constants $\{A, A_1, C_0\}$ and certain integers m_j, and $P\{(1-\varepsilon_0)^2 \leq \chi_n^2/n \leq (1+\varepsilon_0)^2\} \leq \varepsilon/p$. Then, there exist constants c_0 depending on c_* only and C depending on $\{A, A_1, C_0, c_*, c_0\}$ only such that when $\max_j s_{*,j}\lambda_0 \leq c_0$, the conclusions of Theorem 2 hold with at least probability $1-6\varepsilon-2k\sum_j(p-1-|B_j|)/p$.*

The condition $\max_j s_{*,j}\lambda_0 \leq c_0$ on (10), which controls the capped ℓ_1 sparsity of the inverse correlation matrix, weakens the ℓ_0 sparsity condition $d\sqrt{(\log p)/n} \to 0$.

The extra condition on the upper sparse eigenvalue $\kappa_+(m; R^*_{-j,-j})$ in (15) is mild, since it only requires a small $\kappa_+(m;R^*)/m$ that is actually decreasing in m.

The invertibility condition (11) is used to regularize the design matrix in linear regression procedures. As we mentioned earlier, condition (11) holds if the spectrum norm of Ω^* is bounded by $1/c_*$. Since $(R^*)^{-1} = \Omega^* = (\mathrm{diag}\Sigma^*)^{1/2}\Theta^*(\mathrm{diag}\Sigma^*)^{1/2}$, it suffices to have

$$\|(R^*)^{-1}\|_2 \leq \max \Sigma^*_{jj}\|\Theta^*\|_2 \leq 1/c_*.$$

To achieve the convergence rate $d\sqrt{(\log p)/n}$, both Yuan (2010) and Cai et al. (2011) require conditions $\|\Theta^*\|_1 = O(1)$ and $\max\Sigma^*_{jj} = O(1)$. In comparison, the spectrum norm condition is not only weaker than the ℓ_1 operator norm condition, but also more natural for the convergence in spectrum norm.

Our sharper theoretical results are consequences of using the scaled Lasso estimator (4) and its fast convergence rate in linear regression. In Sun and Zhang (2012), a convergence rate of order $s_*(\log p)/n$ was established for the scaled Lasso estimation of the noise level, compared with an oracle noise level as the moment estimator based on the noise vector. In the context of the column-by-column application of the scaled Lasso for precision matrix estimation, the results in Sun and Zhang (2012) can be written as

$$\Big|\frac{\sigma_j^*}{\widehat{\sigma}_j}-1\Big| \le C_1 s_{*,j}\lambda_0^2, \quad \sum_{k\neq j} \overline{\Sigma}_{kk}^{1/2}|\widehat{\beta}_{k,j}-\beta_{k,j}|\sqrt{\Theta_{jj}^*} \le C_2 s_{*,j}\lambda_0, \tag{17}$$

where $\sigma_j^* = \|X\beta_{*,j}\|_2/\sqrt{n}$. We note that $n(\sigma_j^*)^2\Theta_{jj}^*$ is a chi-square variable with n degrees of freedom when X has iid $N(0,\Sigma^*)$ rows. The oracle inequalities in (17) play a crucial role in our analysis of the proposed estimators for inverse matrices, as the following proposition attests.

Proposition 4 *Let Θ^* be a nonnegative definite target matrix, $\Sigma^* = (\Theta^*)^{-1}$, and $\beta = -\Theta^*(\mathrm{diag}\Theta^*)^{-1}$. Let $\widehat{\Theta}$ and $\widehat{\Omega}$ be defined as (7) and (9) based on certain $\widehat{\beta}$ and $\widehat{\sigma}_j$ satisfying (17). Suppose further that*

$$|\Theta_{jj}^*(\sigma_j^*)^2-1| \le C_0\lambda_0, \ \max_j|(\overline{\Sigma}_{jj}/\Sigma_{jj}^*)^{-1/2}-1| \le C_0\lambda_0, \tag{18}$$

and that $\max\{4C_0\lambda_0, 4\lambda_0, C_1 s_{*,j}\lambda_0\} \le 1$. Then, (12) and (13) hold with a constant C depending on $\{C_0, C_2\}$ only. Moreover, if $n\Theta_{jj}^*(\sigma_j^*)^2 \sim \chi_n^2$, then the term $\lambda_0\|\Theta^*\|_1$ in (12) can be replaced by (14) with large probability.

While the results in Sun and Zhang (2012) requires a penalty level $A\sqrt{(2/n)\log(p^2)}$ to allow simultaneous application of (17) for all $j \le p$ via the union bound in proving Theorem 2, Theorem 3 allows a smaller penalty level $\lambda_{*,0} = AL_{n-3/2}(k/p)$ with $A > 1$ and a potentially large $k \asymp s\log(p/s)$. This is based on new theoretical results for the Lasso and scaled Lasso developed in Section 4.

4. Linear Regression Revisited

This section provides certain new error bounds for the Lasso and scaled Lasso in the linear regression model. Compared with existing error bounds, the new results characterize the concentration of the estimation and prediction errors at fixed, smaller threshold levels. The new results also allow high correlation among certain nuisance design vectors.

Consider the linear regression model with standardized design and normal error:

$$y = X\beta + \varepsilon, \ \|x_j\|_2^2 = n, \ \varepsilon \sim N(0, \sigma^2 I_n).$$

Let $\lambda_{univ} = \sqrt{(2/n)\log p}$ be the universal penalty level (Donoho and Johnstone, 1994). For the estimation of β and variable selection, existing theoretical results with $p \gg n$ typically require a penalty level $\lambda = A\sigma\lambda_{univ}$, with $A > 1$, to guarantee rate optimality of regularized estimators. This includes the scaled Lasso with a jointly estimated σ. For the Dantzig selector (Candès and Tao, 2007), performance bounds have been established for $A = 1$.

It is well understood that $\sigma\lambda_{univ}$ in such theorems is a convenient probabilistic upper bound of $\|X^T\varepsilon/n\|_\infty$ for controlling the maximum gradient of the squared loss $\|y-Xb\|_2^2/(2n)$ at $b=\widehat{\beta}$. For $\lambda<\|X^T\varepsilon/n\|_\infty$, variable selection is known to be inconsistent for the Lasso and most other regularized estimates of β, and the analysis of such procedures become more complicated due to false selection. However, this does not preclude the possibility that such a smaller λ outperforms the theoretical $\lambda\geq\sigma\lambda_{univ}$ for the estimation of β or prediction.

In addition to theoretical studies, a large volume of numerical comparisons among regularized estimators exists in the literature. In such numerical studies, the choice of penalty level is typically delegated to computationally more expensive cross-validation methods. Since cross-validation aims to optimize prediction performance, it may lead to a smaller penalty level than $\lambda=\sigma\lambda_{univ}$. However, this gap between $\lambda\geq\sigma\lambda_{univ}$ in theoretical studies and the possible choice of $\lambda<\sigma\lambda_{univ}$ in numerical studies is largely ignored in the existing literature.

The purpose of this section is to provide rate optimal oracle inequalities for the Lasso and its scaled version, which hold with at least probability $1-\varepsilon/p$ for a reasonably small ε, at a fixed penalty level λ satisfying $P\{N(0,\sigma^2/n)>\lambda/A\}=k/p$, with a given $A>1$ and potentially large k, up to $k/(2\log(p/k))^2\asymp s_*$, where s_* is a complexity measure of β, for example, $s_*=\|\beta\|_0$.

When the (scaled) Lasso is simultaneously applied to p subproblems as in the case of matrix estimation, the new oracle inequalities allow the use of the union bound to uniformly control the estimation error in subproblems at the same penalty level.

Rate optimal oracle inequalities have been established for ℓ_1 and concave regularized estimators in Zhang (2010) and Ye and Zhang (2010) for penalty level $\lambda=A\sigma\sqrt{c^*(2/n)\log(p/(\varepsilon s_*))}$, where c^* is an upper sparse eigenvalue, $A>1$ and $1-\varepsilon$ is the guaranteed probability for the oracle inequality to hold. The new oracle inequalities remove the factors c^* and ε from the penalty level, as long as $1/\varepsilon$ is polynomial in p. The penalty level $A\sigma\sqrt{(2/n)\log(p/(\varepsilon s))}$ has been considered for models of size s under ℓ_0 regularization (Birge and Massart, 2001, 2007; Bunea et al., 2007; Abramovich and Grinshtein, 2010).

To bound the effect of the noise when $\lambda<\|X^T\varepsilon/n\|_\infty$, we use a certain sparse ℓ_2 norm to control the excess of $X^T\varepsilon/n$ over a threshold level λ_*. The sparse ℓ_q norm was used in the analysis of regularized estimators before (Candès and Tao, 2007; Zhang and Huang, 2008; Zhang, 2009; Cai et al., 2010; Ye and Zhang, 2010; Zhang, 2010), but it was done without a formal definition of the quantity to the best of our knowledge. To avoid repeating existing calculation, we define the norm and its dual here and summarize their properties in a proposition.

For $1\leq q\leq\infty$ and $t>0$, the sparse ℓ_q norm and its dual are defined as

$$\|v\|_{(q,t)}=\max_{|B|<t+1}\|v_B\|_q,\ \|v\|^*_{(q,t)}=\max_{\|u\|_{(q,t)}\leq 1}u^Tv.$$

The following proposition describes some of their basic properties.

Proposition 5 *Let $m\geq 1$ be an integer, $q'=q/(q-1)$ and $a_q=(1-1/q)/q^{1/(q-1)}$.*
(i) Properties of $\|\cdot\|_{(q,m)}$: $\|v\|_{(q,m)}\downarrow q$, $\|v\|_{(q,m)}/m^{1/q}\downarrow m$, $\|v\|_{(q,m)}/m^{1/q}\uparrow q$,

$$\|v\|_\infty=\|v\|_{(q,1)}\leq\|v\|_{(q,m)}\leq(\|v\|_q)\wedge(m^{1/q}\|v\|_\infty),$$

and $\|v\|_q^q\leq\|v\|^q_{(q,m)}+(a_q/m)^{q-1}\|v\|_1^q$.
(ii) Properties of $\|\cdot\|^*_{(q,m)}$: $m^{1/q}\|v\|^*_{(q,m)}\downarrow q$, and

$$\max\left(\|v\|_{q'},m^{-1/q}\|v\|_1\right)\leq\|v\|^*_{(q,m)}\leq\min\left(\|v\|_{(q',m/a_q)}+m^{-1/q}\|v\|_1,\|v\|_1\right).$$

(iii) Let $\overline{\Sigma} = X^T X/n$ *and* $\kappa_+(m;M)$ *be the sparse eigenvalue in (16). Then,*

$$\|\overline{\Sigma}v\|_{(2,m)} \le \min\Big\{\kappa_+^{1/2}(m;\overline{\Sigma})\|\overline{\Sigma}^{1/2}v\|_2, \kappa_+(m;\overline{\Sigma})\|v\|_2\Big\}.$$

4.1 Lasso with Smaller Penalty: Analytical Bounds

The Lasso path is defined as an $\mathbb{R}^p$-valued function of $\lambda > 0$ as

$$\widehat{\beta}(\lambda) = \arg\min_b \Big\{\|y - Xb\|_2^2/(2n) + \lambda\|b\|_1\Big\}.$$

For threshold levels $\lambda_* > 0$, we consider β satisfying the following complexity bound,

$$|S| + \sum_{j \notin S} |\beta_j|/\lambda_* \le s_* \tag{19}$$

with a certain $S \subset \{1, \ldots, p\}$. This includes the ℓ_0 sparsity condition $\|\beta\|_0 = s_*$ with $S = \text{supp}(\beta)$ and allows $\|\beta\|_0$ to far exceed s_* with many small $|\beta_j|$.

The sparse ℓ_2 norm of a soft-thresholded vector v, at threshold level λ_* in (19), is

$$\zeta_{(2,m)}(v, \lambda_*) = \|(|v| - \lambda_*)_+\|_{(2,m)} = \max_{|J| \le m} \Big\{\sum_{j \in J} (|v_j| - \lambda_*)_+^2\Big\}^{1/2}. \tag{20}$$

Let $B \subseteq \{1, \ldots, p\}$ and

$$z = (z_1, \ldots, z_p)^T = X^T\varepsilon/n.$$

We bound the effect of the excess of the noise over λ_* under the condition

$$\|z_{B^c}\|_\infty \le \lambda_*,\ \zeta_{(2,m)}(z_B, \lambda_*) \le A_1 m^{1/2}\lambda_*, \tag{21}$$

for some $A_1 \ge 0$. We prove that when $\lambda \ge A\lambda_*$ with $A > 1 + A_1$ and (21) holds, a scaled version of $\widehat{\beta}(\lambda) - \beta$ belongs to a set $\mathscr{U}(\overline{\Sigma}, S, B; A, A_1, m, s_* - |S|)$, where

$$\begin{aligned} &\mathscr{U}(\Sigma, S, B; A, A_1, m, m_1) \\ = \ &\Big\{u : u^T\Sigma u + (A-1)\|u_{S^c}\|_1 \le (A+1)\|u_S\|_1 + A_1 m^{1/2}\|u_B\|_{(2,m)}^* + 2Am_1\Big\}. \end{aligned} \tag{22}$$

This leads to the definition of

$$M^*_{pred} = \sup\Big\{\frac{u^T\overline{\Sigma}u/A^2}{m_1 + |S|} : u \in \mathscr{U}(\overline{\Sigma}, S, B; A, A_1, m, m_1)\Big\} \tag{23}$$

as a constant factor for the prediction error of the Lasso and

$$M^*_q = \sup\Big\{\frac{\|u\|_q/A}{(m_1 + |S|)^{1/q}} : u \in \mathscr{U}(\overline{\Sigma}, S, B; A, A_1, m, m_1)\Big\} \tag{24}$$

for the ℓ_q estimation error of the Lasso.

The following theorem provides analytic error bounds for the Lasso prediction and estimation under the sparse ℓ_2 norm condition (21) on the noise. This is different from existing analyses of the Lasso based on the ℓ_∞ noise bound $\|X^T\varepsilon/n\|_\infty \le \lambda_*$. In the case of Gaussian error, (21) allows a fixed threshold level $\lambda_* = \sigma\sqrt{(2/n)\log(p/m)}$ to uniformly control the error of p applications of the Lasso for the estimation of a precision matrix. When $m \asymp s_*$ and $\sigma\sqrt{(2/n)\log(p/m)} \ll \sigma\sqrt{(2/n)\log p}$, using such smaller λ_* is necessary for achieving error bounds with the sharper rate corresponding to $\sigma\sqrt{(2/n)\log(p/m)}$.

Theorem 6 *Suppose (19) holds with certain $\{S, s_*, \lambda_*\}$. Let $A > 1$, $\widehat{\beta} = \widehat{\beta}(\lambda)$ be the Lasso estimator with penalty level $\lambda \geq A\lambda_*$, $h = \widehat{\beta} - \beta$, and $m_1 = s_* - |S|$. If (21) holds with $A_1 \geq 0$, a positive integer m and $B \subseteq \{1, \ldots, p\}$, then*

$$\|Xh\|_2^2/n \leq M^*_{pred} s_* \lambda^2, \quad \|h\|_q \leq M^*_q s_*^{1/q} \lambda. \tag{25}$$

Remark 7 *Theorem 6 covers $\|X^T\varepsilon/n\|_\infty \leq \lambda_*$ as a special case with $A_1 = 0$. In this case, the set (22) does not depend on $\{m, B\}$. For $A_1 = 0$ and $|S| = s_*$ ($m_1 = 0$), (22) contains all vectors satisfying a basic inequality $u^T\Sigma u + (A-1)\|u_{S^c}\|_1 \leq (A+1)\|u_S\|_1$ (Bickel et al., 2009; van de Geer and Bühlmann, 2009; Ye and Zhang, 2010) and Theorem 6 still holds when (22) is replaced by the smaller*

$$\mathscr{U}_-(\Sigma, S, A) = \Big\{u : \|\overline{\Sigma}_{S,*}u\|_\infty \leq A+1, u_j\overline{\Sigma}_{j,*}u \leq -|u_j|(A-1)\ \forall j \notin S\Big\}.$$

Thus, in what follows, we always treat $\mathscr{U}(\Sigma, S, B; A, 0, m, 0)$ as $\mathscr{U}_-(\Sigma, S, A)$ when $A_1 = 0$ and $|S| = s_*$. This yields smaller constants $\{M^*_{pred}, M^*_q\}$ in (23) and (24).

The purpose of including a choice B in (21) is to achieve bounded $\{M^*_{pred}, M^*_1\}$ in the presence of some highly correlated design vectors outside $S \cup B$ when $\overline{\Sigma}_{S\cup B,(S\cup B)^c}$ is small. Since $\|u_B\|^*_{(2,m)}$ is increasing in B, a larger B leads to a larger set (22) and larger $\{M^*_{pred}, M^*_q\}$. However, (21) with smaller B typically requires larger λ_*. Fortunately, the difference in the required λ_* in (21) is of smaller order than λ_* between the largest $B = \{1, \ldots, p\}$ and smaller B with $|B^c| \leq p/m$. We discuss the relationship between $\{M^*_{pred}, M^*_q\}$ and existing conditions on the design in the next section, along with some simple upper bounds for $\{M^*_{pred}, M^*_1, M^*_2\}$.

4.2 Scaled Lasso with Smaller Penalty: Analytical Bounds

The scaled Lasso estimator is defined as

$$\{\widehat{\beta}, \widehat{\sigma}\} = \arg\min_{b,\sigma} \Big\{\|y - Xb\|_2^2/(2n\sigma) + \lambda_0\|b\|_1 + \sigma/2\Big\}, \tag{26}$$

where $\lambda_0 > 0$ is a scale-free penalty level. In this section, we describe the implication of Theorem 6 on the scaled Lasso.

A scaled version of (19) is

$$|S| + \sum_{j \notin S} |\beta_j|/(\sigma^*\lambda_{*,0}) = s_{*,0} \leq s_*, \tag{27}$$

where $\sigma^* = \|\varepsilon\|_2/\sqrt{n}$ is an oracle estimate of the noise level and $\lambda_{*,0} > 0$ is a scaled threshold level. This holds automatically under (19) when $S \supseteq \mathrm{supp}(\beta)$. When $\beta_{S^c} \neq 0$, (27) can be viewed as an event of large probability. When

$$|S| + (1-\varepsilon_0)^{-1} \sum_{j \notin S} \frac{|\beta_j|}{\sigma\lambda_{*,0}} \leq s_* \tag{28}$$

and $\varepsilon \sim N(0, \sigma I_n)$, $P\{s_{*,0} \leq s_*\} \geq P\{\chi_n^2/n \geq (1-\varepsilon_0)^2\} \to 1$ for fixed $\varepsilon_0 > 0$. Let

$$M^*_\sigma = \sup_{u \in \mathscr{U}} \left\{\frac{u^T\overline{\Sigma}u}{s_*A^2} + \frac{2\|u\|_1}{s_*A^2} + \frac{2A_1 m^{1/2}\|u_B\|^*_{(2,m)}}{s_*A^2}\right\} \tag{29}$$

with $\mathscr{U} = \mathscr{U}(\overline{\Sigma}, S, B; A, A_1, m, m_1)$ in (22), as in (23) and (24). Set

$$\eta_* = M_\sigma^* A^2 \lambda_{*,0}^2 s_*, \ \lambda_0 \ge A\lambda_{*,0}/\sqrt{(1-\eta_*)_+}, \ \eta_0 = M_\sigma^* \lambda_0^2 s_*.$$

Theorem 8 *Suppose $\eta_0 < 1$. Let $\{\widehat{\beta}, \widehat{\sigma}\}$ be the scaled Lasso estimator in (26), $\phi_1 = 1/\sqrt{1+\eta_0}$, $\phi_2 = 1/\sqrt{1-\eta_0}$, and $\sigma^* = \|\varepsilon\|_2/\sqrt{n}$. Suppose (21) holds with $\{z_B, \lambda_*\}$ replaced by $\{z_B/\sigma^*, \lambda_{*,0}\}$.*
(i) Let $h^* = (\widehat{\beta} - \beta)/\sigma^*$. Suppose (27) holds. Then,

$$\phi_1 < \widehat{\sigma}/\sigma^* < \phi_2, \ \|Xh^*\|_2^2/n < M_{pred}^* s_* (\phi_2 \lambda_0)^2, \ \|h^*\|_q < M_q^* s_* \phi_2 \lambda_0. \tag{30}$$

(ii) Let $h = \widehat{\beta} - \beta$. Suppose (28) holds and $1 - \varepsilon_0 \le \sigma^*/\sigma \le 1 + \varepsilon_0$. Then,

$$\begin{aligned} &(1-\varepsilon_0)\phi_1 < \widehat{\sigma}/\sigma < \phi_2(1+\varepsilon_0), \\ &\|Xh\|_2^2/n < (1+\varepsilon_0)^2 M_{pred}^* s_* (\sigma\phi_2\lambda_0)^2, \\ &\|h\|_q < (1+\varepsilon_0) M_q^* s_* \sigma \phi_2 \lambda_0. \end{aligned} \tag{31}$$

Compared with Theorem 6, Theorem 8 requires nearly identical conditions on the design X, the noise and penalty level under proper scale. It essentially allows the substitution of $\{y, X, \beta\}$ by $\{y/\sigma^*, X, \beta/\sigma^*\}$ when η_0 is small.

Theorems 6 and 8 require an upper bound (21) for the sparse ℓ_2 norm of the excess noise as well as upper bounds for the constant factors $\{M_{pred}^*, M_q^*, M_\sigma^*\}$ in (23), (24) and (29). Probabilistic upper bounds for the noise and consequences of their combination with Theorems 6 and 8 are discussed in Section 4.3. We use the rest of this subsection to discuss $\{M_{pred}^*, M_q^*, M_\sigma^*\}$.

Existing analyses of the Lasso and Dantzig selector can be to used find upper bounds for $\{M_{pred}^*, M_q^*, M_\sigma^*\}$ via the sparse eigenvalues (Candès and Tao, 2005, 2007; Zhang and Huang, 2008; Zhang, 2009; Cai et al., 2010; Zhang, 2010; Ye and Zhang, 2010). In the simpler case $A_1 = m_1 = 0$, shaper bounds can be obtained using the compatibility factor (van de Geer, 2007; van de Geer and Bühlmann, 2009), the restricted eigenvalue (Bickel et al., 2009; Koltchinskii, 2009), or the cone invertibility factors (Ye and Zhang, 2010; Zhang and Zhang, 2012). Detailed discussions can be found in van de Geer and Bühlmann (2009), Ye and Zhang (2010) and Zhang and Zhang (2012) among others. The main difference here is the possibility of excluding some highly correlated vectors from B in the case of $A_1 > 0$. The following lemma provide some simple bounds used in our analysis of the scaled Lasso estimation of the precision matrix.

Lemma 9 *Let $\{M_{pred}^*, M_q^*, M_\sigma^*\}$ be as in (23), (24) and (29) with the vector class $\mathscr{U}(\overline{\Sigma}, S, B; A, A_1, m, m_1)$ in (22). Suppose that for a nonnegative-definite matrix Σ, $\max_j \|\overline{\Sigma}_{j,*} - \Sigma_{j,*}\|_\infty \le \lambda^*$ and $c_*\|u_{S\cup B}\|_2^2 \le u^T \Sigma u$ for $u \in \mathscr{U}(\overline{\Sigma}, S, B; A, A_1, m, m_1)$. Suppose further that $\lambda^*\{(s_* \vee m)/c_*\}(2A + A_1)^2 \le (A - A_1 - 1)_+^2/2$. Then,*

$$M_{pred}^* + M_1^*\Big(1 - \frac{A_1+1}{A}\Big) \le \max\Big\{\frac{4 \vee (4m/s_*)}{c_*(2 + A_1/A)^{-2}}, \frac{c_*(1 - |S|/s_*)}{A^2}\Big\} \tag{32}$$

and

$$M_\sigma^* \le \Big(1 + \frac{2A_1}{c_* A}\Big) M_{pred}^* + 2(1+A_1)\frac{M_1^*}{A} + \frac{A_1 m}{A s_*} + \frac{2A_1}{A^3}\Big(1 - \frac{|S|}{s_*}\Big). \tag{33}$$

Moreover, if in addition $B = \{1, \ldots, p\}$ then

$$M_2^* \le (2/c_*) M_{pred}^* + 2(1 - |S|/s_*)/(A^2). \tag{34}$$

The main condition of Lemma 9,

$$c_* \le \inf\left\{ \frac{u^T \Sigma u}{\|u_{S\cup B}\|_2^2} : u \in \mathscr{U}(\overline{\Sigma}, S, B; A, A_1, m, m_1) \right\}, \tag{35}$$

can be viewed as a restricted eigenvalue condition (Bickel et al., 2009) on a population version of the Gram matrix. However, one may also pick the sample version $\Sigma = \overline{\Sigma}$ with $\lambda^* = 0$. Let $\{A, A_1\}$ be fixed constants satisfying $A_1 < A - 1$. Lemma 9 asserts that the factors $\{M^*_{pred}, M^*_1, M^*_\sigma\}$ can be all treated as constants when $1/c_*$ and m/s_* are bounded and $\lambda^*(s_* \vee m)/c_*$ is smaller than a certain constant. Moreover, M^*_2 can be also treated as a constant when (35) holds for $B = \{1, \ldots, p\}$.

4.3 Probabilistic Error Bounds

Theorems 6 and 8 provides analytical error bounds based on the size of the excess noise over a given threshold. Here we provide probabilistic upper bounds for the excess noise and describe their implications in combination with Theorems 6 and 8. We use the following notation:

$$L_n(t) = n^{-1/2}\Phi^{-1}(1-t), \tag{36}$$

where $\Phi^{-1}(t)$ is the standard normal quantile function.

Proposition 10 *Let $\zeta_{(2,m)}(v, \lambda_*)$ be as in (20) and $\kappa_+(m) = \kappa_+(m; \overline{\Sigma})$ as in (16) with $\overline{\Sigma} = X^T X/n$. Suppose $\varepsilon \sim N(0, \sigma^2 I_n)$ and $\|x_j\|_2^2 = n$. Let $k > 0$.*
(i) Let $z = X^T\varepsilon/n$ and $\lambda_* = \sigma L_n(k/p)$. Then, $P\{\zeta_{(2,p)}(z, \lambda_*) > 0\} \le 2k$, and

$$E\zeta^2_{(2,p)}(z, \lambda_*) \le 4k\lambda_*^2/\{L_1^4(k/p) + 2L_1^2(k/p)\},$$
$$P\Big\{\zeta_{(2,m)}(z, \lambda_*) > E\zeta_{(2,p)}(z, \lambda_*) + \sigma L_n(\varepsilon)\sqrt{\kappa_+(m)}\Big\} \le \varepsilon. \tag{37}$$

(ii) Let $\sigma^* = \|\varepsilon\|_2/\sqrt{n}$, $z^* = z/\sigma^*$, $\lambda_{*,0} = L_{n-3/2}(k/p)$ and $\varepsilon_n = e^{1/(4n-6)^2} - 1$. Then, $P\{\zeta_{(2,p)}(z^*, \lambda_{*,0}) > 0\} \le (1+\varepsilon_n)k$, $E\zeta^2_{(2,p)}(z^*, \lambda_{*,0}) \le (1+\varepsilon_n)4k\lambda^2_{*,0}/\{L_1^4(k/p) + 2L_1^2(k/p)\}$, and

$$P\Big\{\zeta_{(2,m)}(z^*, \lambda_{*,0}) > \mu_{(2,m)} + L_{n-3/2}(\varepsilon)\sqrt{\kappa_+(m)}\Big\} \le (1+\varepsilon_n)\varepsilon, \tag{38}$$

where $\mu_{(2,m)}$ is the median of $\zeta_{(2,m)}(z^*, \lambda_{*,0})$. Moreover,

$$\mu_{(2,m)} \le E\zeta_{(2,p)}(z^*, \lambda_{*,0}) + (1+\varepsilon_n)\{\lambda_{*,0}/L_1(k/p)\}\sqrt{\kappa_+(m)/(2\pi)}. \tag{39}$$

We describe consequences of combining Proposition 10 with Theorems 6 and 8 in three theorems, respectively using the probability of no excess noise over the threshold, the Markov inequality with the second moment, and the concentration bound on the excess noise.

Theorem 11 *Let $0 < \varepsilon < p$. Suppose $\varepsilon \sim N(0, \sigma^2 I_n)$.*
(i) Let the notation be as in Theorem 6 and (36) with $A_1 = 0$ and $\lambda_* = \sigma L_n(\varepsilon/p^2)$. If (19) holds, then (25) holds with at least probability $1 - 2\varepsilon/p$.
(ii) Let the notation be as in Theorem 8 and (36) with $A_1 = 0$ and $\lambda_{*,0} = L_{n-3/2}(\varepsilon/p^2)$. If (28) holds with $P\{(1-\varepsilon_0)^2 \le \chi_n^2/n \le (1+\varepsilon_0)^2\} \le \varepsilon/p$, then (30) and (31) hold with at least probability $1 - 3\varepsilon/p$.

For a single application of the Lasso or scaled Lasso, $\epsilon/p = o(1)$ guarantees $\|z\|_\infty \le \lambda_*$ in Theorem 11 (i) and $\|z^*\|_\infty \le \lambda_{*,0}$ in Theorem 11 (ii) with high probability. The threshold levels are $\lambda_*/\sigma \approx \lambda_{*,0} \approx \lambda_{univ} = \sqrt{(2/n)\log p}$, as typically considered in the literature. In numerical experiments, this often produces nearly optimal results although the threshold level may still be somewhat higher than optimal for the prediction and estimation of β. However, if we use the union bound to guarantee the simultaneous validity of the oracle inequalities in p applications of the scaled Lasso in the estimation of individual columns of a precision matrix, Theorem 11 requires $\epsilon = o(1)$, or equivalently a significantly higher threshold level $\lambda_{*,0} \approx \sqrt{(4/n)\log p}$. This higher $\lambda_{*,0}$, which does not change the theoretical results by much, may produce clearly suboptimal results in numerical experiments.

Theorem 12 *Let $k > 0$. Suppose $\epsilon \sim N(0, \sigma^2 I_n)$.*
(i) Let the notation be as in Theorem 6 and Proposition 10, $\lambda_* = \sigma L_n(k/p)$, and $A - 1 > A_1 \ge \sqrt{4k/(\epsilon m(L_1^4(k/p) + 2L_1^2(k/p)))}$. If (19) holds, then (25) holds with at least probability $1 - \epsilon - 2|B^c|k/p$.
(ii) Let the notation be as in Theorem 8 and Proposition 10, $\lambda_{*,0} = L_{n-3/2}(k/p)$, $\epsilon_n = e^{1/(4n-6)^2} - 1$, and $A - 1 > A_1 \ge \sqrt{(1+\epsilon_n)4k/(\epsilon m(L_1^4(k/p) + 2L_1^2(k/p)))}$. If (28) holds with $P\{(1-\epsilon_0)^2 \le \chi_n^2/n \le (1+\epsilon_0)^2\} \le \epsilon$, then (30) and (31) hold with at least probability $1 - 2\epsilon - 2|B^c|k/p$.

Theorem 12 uses the upper bounds for $E\zeta^2_{(2,p)}(z, \lambda_*)$ and $E\zeta^2_{(2,p)}(z^*, \lambda_{*,0})$ to verify (21). Since $L_n(k/p) \approx \sqrt{(2/n)\log(p/k)}$, it allows smaller threshold levels λ_* and $\lambda_{*,0}$ as long as $k/(\epsilon m(L_1^4(k/p) + 2L_1^2(k/p)))$ is small. However, it does not allow $\epsilon \le 1/p$ for using the union bound in p applications of the Lasso in precision matrix estimation.

Theorem 13 *Let $k > 0$. Suppose $\epsilon \sim N(0, \sigma^2 I_n)$.*
(i) Let the notation be as in Theorem 6 and Proposition 10, $\lambda_* = \sigma L_n(k/p)$, and

$$A - 1 > A_1 \ge \Big(\frac{4k/m}{L_1^4(k/p) + 2L_1^2(k/p)}\Big)^{1/2} + \frac{L_1(\epsilon/p)}{L_1(k/p)}\Big(\frac{\kappa_+(m)}{m}\Big)^{1/2}.$$

If (19) holds, then (25) holds with at least probability $1 - \epsilon/p - 2|B^c|k/p$.
(ii) Let the notation be as in Theorem 8 and Proposition 10, $\lambda_{*,0} = L_{n-3/2}(k/p)$, $\epsilon_n = e^{1/(4n-6)^2} - 1$, and

$$A - 1 > A_1 \ge \Big(\frac{(1+\epsilon_n)4k/m}{L_1^4(k/p) + 2L_1^2(k/p)}\Big)^{1/2} + \Big(\frac{L_1(\epsilon/p)}{L_1(k/p)} + \frac{1+\epsilon_n}{L_1(k/p)\sqrt{2\pi}}\Big)\Big(\frac{\kappa_+(m)}{m}\Big)^{1/2}.$$

If (28) holds with $P\{(1-\epsilon_0)^2 \le \chi_n^2/n \le (1+\epsilon_0)^2\} \le \epsilon/p$, then (30) and (31) hold with at least probability $1 - 2\epsilon/p - 2|B^c|k/p$.

Theorem 13 uses concentration inequalities (37) and (38) to verify (21). Let $B = \{1, \dots, p\}$ and $L_n(t)$ be as in (36). By guaranteeing the validity of the oracle inequalities with $1 - \epsilon/p$ probability, with a reasonably small ϵ, Theorem 13 justifies the use of a fixed smaller threshold level $\lambda_{*,0} =$

$L_{n-3/2}(k/p) \approx \sqrt{(2/n)\log(p/k)}$ in p applications of the scaled Lasso to estimate columns of a precision matrix.

Since $L_1(\varepsilon/p) \approx L_1(k/p)$ typically holds, Theorem 13 only requires $(k/m)/(L_1^4(k/p)+2L_1^2(k/p))$ and $\kappa_+(m)/m$ be smaller than a fixed small constant. This condition relies on the upper sparse eigenvalue only in a mild way since $\kappa_+(m)/m$ is decreasing in m and $\kappa_+(m)/m \le 1/m+(1-1/m)\max_{j\neq k}|x_j^T x_k/n|$ (Zhang and Huang, 2008).

For $k \asymp m$ and $\log(p/k) \asymp \log(p/(s_*\vee 1))$, Theorem 13 provides prediction and ℓ_q error bounds of the orders $\sigma^2(s_*\vee 1)\lambda_{*,0}^2 \approx \sigma^2((s_*\vee 1)/n)2\log(p/(s_*\vee 1))$ and $\sigma(s_*\vee 1)^{1/q}\lambda_{*,0}$ respectively. For $\log(p/n) \ll \log n$, this could be of smaller order than the error bounds with $\lambda_{*,0} \approx \lambda_{univ} = \sqrt{(2/n)\log p}$.

Theorem 13 suggests the use of a penalty level satisfying $\lambda/\sigma = \lambda_0 = AL_n(k/p) \approx A\sqrt{(2/n)\log(p/k)}$ with $1 < A \le \sqrt{2}$ and a real solution of $k = L_1^4(k/p)+2L_1^2(k/p)$. This is conservative since the constraint on A in the theorem is valid with a moderate $m = O(s_*+1)$. For p applications of the scaled Lasso in the estimation of precision matrix, this also provides a more practical penalty level compared with $A'L_n(\varepsilon/p^2) \approx A'\sqrt{(4/n)\log(p/\varepsilon^{1/2})}$, $A' > 1$ and $\varepsilon \ll 1$, based on existing results and Theorem 11. In our simulation study, we use $\lambda_0 = \sqrt{2}L_n(k/p)$ with $k = L_1^4(k/p)+2L_1^2(k/p)$.

4.4 A Lower Performance Bound

It is well understood that in the class of β satisfying the sparsity condition (19), $s_*\sigma^2 L_n^2(s_*/p)$ and $s_*^{1/q}\sigma L_n(s_*/p)$ are respectively lower bounds for the rates of minimax prediction and ℓ_q estimation error (Ye and Zhang, 2010; Raskutti et al., 2011). This can be achieved by the Lasso with $\lambda_* = \sigma L_n(m/p)$, $m \asymp s_*$, or scaled Lasso with $\lambda_{*,0} = L_{n-3/2}(m/p)$. The following proposition asserts that for each fixed β, the minimax error rate cannot be achieved by regularizing the gradient with a threshold level of smaller order.

Proposition 14 *Let $y = X\beta+\varepsilon$, $\widetilde{\beta}(\lambda)$ satisfy $\|X^T(y-X\widetilde{\beta}(\lambda))/n\|_\infty \le \lambda$, and $\widetilde{h}(\lambda) = \widetilde{\beta}(\lambda)-\beta$. Let $\overline{\Sigma} = X^TX/n$ and $\kappa_+(m;\cdot)$ be as in (16).*
(i) If $\|X^T\varepsilon/n\|_{(2,k)} \ge k^{1/2}\lambda_* > 0$, then for all $A > 1$

$$\inf_{\lambda\le\lambda_*/A}\min\Big\{\frac{\|X\widetilde{h}(\lambda)\|^2/n}{\kappa_+^{-1}(k;\overline{\Sigma})},\frac{\|\widetilde{h}(\lambda)\|_2^2}{\kappa_+^{-2}(k;\overline{\Sigma})}\Big\} \ge (1-1/A)^2 k\lambda_*^2. \tag{40}$$

(ii) Let $\sigma^* = \|\varepsilon\|_2/\sqrt{n}$ and $N_k = \#\{j : |x_j^T\varepsilon|/(n\sigma^*) \ge \widetilde{L}_n(k/p)\}$ with $\widetilde{L}_n(t) = L_n(t)-n^{-1/2}$. Suppose X has iid $N(0,\Sigma)$ rows, $\mathrm{diag}(\Sigma) = I_p$, and $2k-4\|\Sigma\|_2 \ge \big(\sqrt{k-1}+\sqrt{2\|\Sigma\|_2\log(1/\varepsilon)}\big)^2$. Then, $P\{N_k \ge k\} \ge 1-\varepsilon$ and

$$P\Big\{\|X^T\varepsilon/n\|_{(q,k)} \ge \sigma^* k^{1/q}\sigma\widetilde{L}_n(k/p)\Big\} \ge 1-\varepsilon. \tag{41}$$

Consequently, there exist numerical constants c_1 and c_2 such that

$$P\Big\{\inf_{\lambda\le c_1\sigma L_n(k/p)}\min\Big(\frac{\|X\widetilde{h}(\lambda)\|^2/n}{\kappa_+^{-1}(k;\overline{\Sigma})},\frac{\|\widetilde{h}(\lambda)\|_2^2}{\kappa_+^{-2}(k;\overline{\Sigma})}\Big) \ge c_2\sigma^2 kL_n^2(k/p)\Big\} \ge 1-\varepsilon-e^{-n/9}.$$

It follows from Proposition 14 (ii) that the prediction and ℓ_2 estimation error is of no smaller order than $k\sigma^2 L_n^2(k/p)$ for all $\lambda \le c_1\sigma L_n(k/p)$. This rate is suboptimal when $k\log(p/k) \gg s_*\log(p/s_*)$.

5. Estimation after Model Selection

We have presented theoretical properties of the scaled Lasso for linear regression and precision matrix estimation. After model selection, the least squares estimator is often used to remove bias of regularized estimators. The usefulness of this technique after the scaled Lasso was demonstrated in Sun and Zhang (2012), along with its theoretical justification. In this section, we extend the theory to smaller threshold level and to the estimation of precision matrix.

In linear regression, the least squares estimator $\overline{\beta}$ and the corresponding estimate of σ in the model selected by a regularized estimator $\widehat{\beta}$ are given by

$$\overline{\beta} = \arg\min_{b} \Big\{ \|y - Xb\|_2^2 : \mathrm{supp}(b) \subseteq \widehat{S} \Big\}, \quad \overline{\sigma} = \big\|y - X\overline{\beta}\big\|_2 / \sqrt{n}, \tag{42}$$

where $\widehat{S} = \mathrm{supp}(\widehat{\beta})$. To study the performance of (42), we define sparse eigenvalues relative to a support set S as follows:

$$\begin{aligned} \kappa_-^*(m^*, S; \Sigma) &= \min_{J \supseteq S, |J \setminus S| \le m^*} \min_{\|u_J\|_2 = 1} u_J^T \Sigma_{J,J} u_J, \\ \kappa_+^*(m^*, S; \Sigma) &= \min_{J \cap S = \emptyset, |J| \le m^*} \max_{\|u_J\|_2 = 1} u_J^T \Sigma_{J,J} u_J. \end{aligned}$$

It is proved in Sun and Zhang (2012) that $\{\overline{\beta}, \overline{\sigma}\}$ satisfies prediction and estimation error bounds of the same order as those for the scaled Lasso (26) under some extra conditions on $\kappa_\pm^*(m^*, S; \overline{\Sigma})$. The extra condition on $\kappa_+^*(m^*, S; \overline{\Sigma})$ is used to derive an upper bound for the false positive $|\widehat{S} \setminus S|$, and then the extra condition on $\kappa_-^*(m^*, S; \overline{\Sigma})$ is used to invert $X_{S \cup \widehat{S}}$. The following theorem extends the result to the smaller threshold level $\lambda_{*,0} = L_{n-3/2}(k/p)$ in Theorem 13 (ii). Let

$$M_{lse}^* = \frac{\Big[\big\{ |S| + \big(\sqrt{m^*} + \sqrt{2m^* \log(ep/m^*)} \big)^2 \big\}^{1/2} + L_1(\varepsilon/p) \Big]^2}{s_* \log(p/s_*)}.$$

Theorem 15 *Let $(\widehat{\beta}, \widehat{\sigma})$ be the scaled lasso estimator in (26) and $(\overline{\beta}, \overline{\sigma})$ be the least squares estimator (42) in the selected model $\widehat{S} = \mathrm{supp}(\widehat{\beta})$. Let the notation be as in Theorem 13 (ii) and $m^* > m$ be an integer satisfying $s_* M_{pred}^* / \{(1-\xi_1)(1-1/A)\}^2 \le m^* / \kappa_+^*(m^*, S)$. Suppose $\beta_{S^c} = 0$ and (21) holds with $\{z_B, \lambda_*\}$ replaced by $\{z_B^*, \lambda_{*,0}\}$. Then,*

$$|\widehat{S} \setminus S| \le m^* \tag{43}$$

with at least probability $1 - 2\varepsilon/p - 2|B^c|k/p$. Moreover,

$$\begin{aligned} & (\sigma^*)^2 - \sigma^2 M_{lse}^* (s_*/n) \log(p/s_*) \\ \le\ & \overline{\sigma}^2 \\ \le\ & \widehat{\sigma}^2, \\ & \kappa_-^*(m^* - 1, S) \|\overline{h}\|_2^2 \\ \le\ & \|X\overline{h}\|_2^2 / n \\ \le\ & (1 + \varepsilon_0)^2 M_{pred}^* s_* (\sigma \phi_2 \lambda_0)^2 + \sigma^2 M_{lse}^* (s_*/n) \log(p/s_*), \end{aligned} \tag{44}$$

with at least probability $1 - 3\varepsilon/p - 2|B^c|k/p$, where $\overline{h} = \overline{\beta} - \beta$.

Theorem 15 asserts that when $k \vee m^* \asymp s_*$, the least squares estimator $\{\overline{\beta}, \overline{\sigma}\}$ after the scaled Lasso selection enjoys estimation and prediction properties comparable to that of the scaled Lasso:

$$\lambda_0^{-2}\Big\{\big|\overline{\sigma}/\sigma^* - 1\big| + \big\|\overline{\beta} - \beta\big\|_2^2 + \big\|X\overline{\beta} - X\beta\big\|_2^2/n\Big\} + \|\overline{\beta}\|_0 = O_P(1)s_*.$$

Now we apply this method for precision matrix estimation. Let $\widehat{\beta}$ be as in (4) and define $\overline{\beta}$ and $\overline{\sigma}$ as follows:

$$\begin{aligned} \overline{\beta}_{*,j} &= \arg\min_b \Big\{\|Xb\|_2^2 : b_j = -1, \mathrm{supp}(b) \subseteq \mathrm{supp}(\widehat{\beta}_{*,j})\Big\}, \\ \overline{\sigma}_j &= \big\|X\overline{\beta}_{*,j}\big\|_2/\sqrt{n}. \end{aligned} \tag{45}$$

We define $\widetilde{\Theta}^{\mathrm{LSE}}$ and $\widehat{\Theta}^{\mathrm{LSE}}$ as in (7) and (9) with $\overline{\beta}$ and $\overline{\sigma}$ in place of $\widehat{\beta}$ and $\widehat{\sigma}$.

Under an additional condition on the upper sparse eigenvalue, Theorem 15 is parallel to Theorem 8 (ii), and the theoretical results in Sun and Zhang (2012) are parallel to Theorem 6 (ii). These results can be used to verify the condition (17), so that Proposition 4 also applies to (45) with the extra upper sparse eigenvalue condition on the population correlation matrix R^*. We formally state this result as a corollary.

Corollary 16 *Under the additional condition $\|R^*\|_2 = O(1)$ on the population correlation matrix R^*, Theorems 2 and 3 are applicable to the estimator $\widehat{\Theta}^{LSE}$ and the corresponding estimator for $\Omega^* = (R^*)^{-1}$ with possibly different numerical constants.*

6. Numerical Study

In this section, we present some numerical comparison between the proposed and existing methods. In addition to the proposed estimator (7) and (9) based on the scaled Lasso (4) and the least squares estimation after the scale Lasso (45), the graphical Lasso and CLIME are considered. The following three models are considered. Models 1 and 2 have been considered in Cai et al. (2011), while Model 2 in Rothman et al. (2008).

- Model 1: $\Theta_{ij} = 0.6^{|i-j|}$.
- Model 2: Let $\Theta = B + \delta I$, where each off-diagonal entry in B is generated independently and equals to 0.5 with probability 0.1 or 0 with probability 0.9. The constant δ is chosen such that the condition number of Θ^* is p. Finally, we rescale the matrix Θ^* to the unit in diagonal.
- Model 3: The diagonal of the target matrix has unequal values. $\Theta = D^{1/2}\Omega D^{1/2}$, where $\Omega_{ij} = 0.6^{|i-j|}$ and D is a diagonal matrix with diagonal elements $d_{ii} = (4i + p - 5)/\{5(p-1)\}$.

Among the three models, Model 2 is the densest. For $p = 1000$, the capped ℓ_1 sparsity s_* is 8.84, 24.63, and 8.80 for three models respectively.

In each model, we generate a training sample of size 100 from a multivariate normal distribution with mean zero and covariance matrix $\Sigma = \Theta^{-1}$ and an independent sample of size 100 from the same distribution for validating the tuning parameter λ for the graphical Lasso and CLIME. The GLasso and CLIME estimators are computed based on training data with various λ's and we choose λ by minimizing likelihood loss $\{\mathrm{trace}(\overline{\Sigma}\widehat{\Theta}) - \log\det(\widehat{\Theta})\}$ on the validation sample. The scaled

Lasso estimators are computed based on the training sample alone with penalty level $\lambda_0 = AL_n(k/p)$, where $A = \sqrt{2}$ and k is the solution of $k = L_1^4(k/p) + 2L_1^2(k/p)$. The symmetrization step in Cai et al. (2011) is applied. We consider six different dimensions $p = 30, 60, 90, 150, 300, 1000$ and replicate 100 times in each setting. The CLIME estimators for $p = 300$ and $p = 1000$ are not computed due to computational costs.

Table 1 presents the mean and standard deviation of estimation errors based on 100 replications. The estimation error is measured by three matrix norms: the spectrum norm, the matrix ℓ_1 norm and the Frobenius norm. The scaled Lasso estimator, labeled as SLasso, outperforms the graphical Lasso (GLasso) in all cases except for the smaller $p \in \{30, 60, 90\}$ in the Frobenius loss in the denser Model 2. It also outperforms the CLIME in most cases, except for smaller p in sparser models ($p = 30$ in Model 1 and $p \in \{30, 60\}$ in Model 3). The least squares estimator after the scaled Lasso selection outperforms all estimators by large margin in the spectrum and Frobenius losses in Models 1 and 3, but in general underperforms in the ℓ_1 operator norm and in Model 2. It seems that post processing by the least squares method is a somewhat aggressive procedure for bias correction. It performs well in sparse models, where variable selection is easier, but may not perform very well in denser models.

Both the scaled Lasso and the CLIME are resulting from sparse linear regression solutions. A main advantage of the scaled Lasso over the CLIME is adaptive choice of the penalty level for the estimation of each column of the precision matrix. The CLIME uses cross-validation to choose a common penalty level for all p columns. When p is large, it is computationally difficult. In fact, this prevented us from completing the simulation experiment for the CLIME for the larger $p \in \{300, 1000\}$.

7. Discussion

Since the scaled Lasso choose penalty levels adaptively in the estimation of each column of the precision matrix, it is expected to outperform methods using a fixed penalty level for all columns in the presence of heterogeneity of the diagonal of the precision matrix. Let $\widetilde{\Theta}(\lambda)$ be an estimator with columns

$$\widetilde{\Theta}_{*j}(\lambda) = \mathop{\arg\min}_{v \in \mathbb{R}^p} \Big\{ \|v\|_1 : \|\overline{\Sigma} v - e_j\|_\infty \le \lambda \Big\},\ j = 1, \ldots, p. \tag{46}$$

The CLIME is a symmetrization of this estimator $\widetilde{\Theta}(\lambda)$ with fixed penalty level for all columns. In the following example, the scaled Lasso estimator has a faster convergence rate than (46). The example also demonstrates the possibility of achieving the rate $d\lambda_0$ in Theorem 2 with unbounded $\|\Theta^*\|_2 \ge d^2$, when Theorem 1 is not applicable.

Example 1 *Let $p > n^2 + 3 + m$ with $(m, m^4(\log p)/n) \to (\infty, 0)$ and $4m^2 \le \log p$. Let $L_n(t) \approx \sqrt{(2/n)\log(1/t)}$ be as in (36). Let $\{J_1, J_2, J_3\}$ be a partition of $\{1, \ldots, p\}$ with $J_1 = \{1, 2\}$ and $J_2 = \{3, \ldots, 3+m\}$. Let $\rho_1 = \sqrt{1 - 1/m^2}$, $v = (v_1, \ldots, v_m)^T \in \mathbb{R}^m$ with $v_j^2 = 1/m$, $\rho_2 = c_0 m^{3/2} L_n(m/p) = o(1)$, and*

$$\Sigma^* = \begin{pmatrix} \Sigma^*_{J_1,J_1} & 0 & 0 \\ 0 & \Sigma^*_{J_2,J_2} & 0 \\ 0 & 0 & I_{p-m-3} \end{pmatrix}, \ \Sigma^*_{J_1,J_1} = \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix}, \ \Sigma^*_{J_2,J_2} = \begin{pmatrix} 1 & \rho_2 v^T \\ \rho_2 v & I_m \end{pmatrix}.$$

	Model 1											
	Spectrum norm				Matrix ℓ_1 norm				Frobenius norm			
p	SLasso	SLasso/LSE	GLasso	CLIME	SLasso	SLasso/LSE	GLasso	CLIME	SLasso	SLasso/LSE	GLasso	CLIME
30	2.46(0.07)	1.77(0.15)	2.49(0.14)	2.29(0.21)	2.95(0.10)	2.73(0.21)	3.09(0.11)	2.92(0.17)	4.20(0.11)	3.37(0.14)	4.24(0.26)	3.80(0.36)
60	2.68(0.05)	2.04(0.11)	2.94(0.05)	2.68(0.10)	3.12(0.08)	3.17(0.25)	3.55(0.07)	3.27(0.09)	6.41(0.09)	5.35(0.15)	7.15(0.15)	6.32(0.28)
90	2.75(0.04)	2.09(0.08)	3.07(0.03)	2.87(0.09)	3.21(0.07)	3.49(0.31)	3.72(0.06)	3.42(0.07)	8.09(0.10)	6.87(0.15)	9.25(0.12)	8.42(0.31)
150	2.84(0.03)	2.18(0.06)	3.19(0.02)	3.05(0.04)	3.29(0.07)	3.81(0.31)	3.88(0.06)	3.55(0.06)	10.79(0.11)	9.32(0.14)	12.55(0.09)	11.68(0.20)
300	2.93(0.02)	2.25(0.05)	3.29(0.01)	NA	3.39(0.05)	4.36(0.38)	4.06(0.05)	NA	15.83(0.09)	13.89(0.16)	18.44(0.09)	NA
1000	3.08(0.02)	2.38(0.08)	3.39(0.00)	NA	3.51(0.03)	5.13(0.37)	4.44(0.07)	NA	30.55(0.09)	26.68(0.19)	35.11(0.06)	NA

	Model 2											
	Spectrum norm				Matrix ℓ_1 norm				Frobenius norm			
p	SLasso	SLasso/LSE	GLasso	CLIME	SLasso	SLasso/LSE	GLasso	CLIME	SLasso	SLasso/LSE	GLasso	CLIME
30	0.72(0.08)	1.08(0.17)	0.82(0.07)	0.81(0.09)	1.27(0.15)	1.86(0.33)	1.49(0.15)	1.45(0.18)	1.86(0.10)	2.40(0.24)	1.84(0.09)	1.87(0.11)
60	1.06(0.05)	1.41(0.19)	1.15(0.06)	1.19(0.08)	1.93(0.15)	2.68(0.35)	2.21(0.12)	2.20(0.23)	3.27(0.08)	4.19(0.26)	3.18(0.13)	3.42(0.09)
90	1.48(0.04)	1.73(0.23)	1.54(0.05)	1.61(0.04)	2.58(0.15)	3.81(0.46)	2.89(0.16)	2.90(0.17)	4.42(0.07)	6.18(0.29)	4.40(0.11)	4.65(0.08)
150	1.96(0.03)	2.04(0.28)	2.02(0.05)	2.06(0.03)	3.25(0.17)	5.21(0.63)	3.60(0.15)	3.65(0.19)	5.95(0.06)	9.17(0.33)	6.19(0.16)	6.33(0.08)
300	2.88(0.02)	2.34(0.19)	2.89(0.02)	NA	4.45(0.13)	6.75(0.52)	4.92(0.17)	NA	9.26(0.05)	13.99(0.37)	9.79(0.05)	NA
1000	5.46(0.01)	4.88(0.03)	5.52(0.01)	NA	7.09(0.09)	10.26(0.53)	7.98(0.15)	NA	18.85(0.06)	26.08(0.30)	20.81(0.02)	NA

	Model 3											
	Spectrum norm				Matrix ℓ_1 norm				Frobenius norm			
p	SLasso	SLasso/LSE	GLasso	CLIME	SLasso	SLasso/LSE	GLasso	CLIME	SLasso	SLasso/LSE	GLasso	CLIME
30	1.84(0.09)	1.28(0.15)	2.08(0.10)	1.63(0.19)	2.30(0.12)	2.00(0.19)	2.59(0.10)	2.17(0.20)	2.69(0.09)	2.15(0.13)	2.91(0.16)	2.37(0.25)
60	2.18(0.07)	1.58(0.13)	2.63(0.04)	2.10(0.10)	2.63(0.10)	2.46(0.18)	3.10(0.05)	2.65(0.14)	4.10(0.08)	3.41(0.10)	4.84(0.08)	3.98(0.13)
90	2.34(0.06)	1.71(0.11)	2.84(0.03)	2.38(0.18)	2.77(0.10)	2.73(0.20)	3.30(0.06)	2.91(0.12)	5.19(0.08)	4.40(0.10)	6.25(0.08)	5.37(0.37)
150	2.51(0.05)	1.84(0.09)	3.06(0.02)	2.76(0.05)	2.93(0.09)	3.04(0.28)	3.45(0.04)	3.18(0.09)	6.93(0.08)	5.96(0.10)	8.43(0.07)	7.75(0.08)
300	2.70(0.05)	1.99(0.08)	3.26(0.01)	NA	3.10(0.07)	3.39(0.27)	3.58(0.03)	NA	10.18(0.08)	8.89(0.10)	12.41(0.04)	NA
1000	2.94(0.03)	2.16(0.07)	3.47(0.01)	NA	3.32(0.06)	4.07(0.32)	3.73(0.03)	NA	19.63(0.07)	17.04(0.15)	23.55(0.02)	NA

Table 1: Estimation errors under various matrix norms of scaled Lasso, GLasso and CLIME for three models.

The eigenvalues of $\Sigma^*_{J_1,J_1}$ *are* $1 \pm \rho_1$*, those of* $\Sigma^*_{J_2,J_2}$ *are* $1 \pm \rho_2, 1, \ldots, 1$*, and*

$$(\Sigma^*_{J_1,J_1})^{-1} = m^2 \begin{pmatrix} 1 & -\rho_1 \\ -\rho_1 & 1 \end{pmatrix}, \ (\Sigma^*_{J_2,J_2})^{-1} = \frac{1}{1-\rho_2^2} \begin{pmatrix} 1 & -\rho_2 v^T \\ -\rho_2 v & (1-\rho_2^2) I_m + \rho_2^2 v v^T \end{pmatrix}.$$

We note that $\mathrm{diag}(\Sigma^*) = I_p$*,* $d = m+1$ *is the maximum degree,* $\|\Theta^*\|_2 = 1/(1-\rho_1) \approx 2d^2$*, and* $\|\Theta^*\|_1 \approx 2d^2$*. The following statements are proved in the Appendix.*
(i) Let $\widehat{\Theta}$ *be the scaled Lasso estimator of* $\Theta^* = (\Sigma^*)^{-1}$ *with penalty level* $\lambda_0 = A\sqrt{(4/n)\log p}$*,* $A > 1$*, as in Theorem 2. Then, there exists a constant* M_1^* *such that*

$$P\Big\{ \|\widehat{\Theta} - \Theta^*\|_2 \le \|\widehat{\Theta} - \Theta^*\|_1 \le M_1^* m L_n(m/p) \Big\} \to 1.$$

(ii) If $\rho_2 = c_0 m^{3/2} L_n(m/p)$ *with a sufficiently small constant* $c_0 > 0$*, then*

$$P\Big\{ \inf_{\lambda > 0} \|\widetilde{\Theta}(\lambda) - \Theta^*\|_2 \ge c_0 m^{3/2} L_n(m/p)/\sqrt{1+1/m} \Big\} \to 1.$$

Thus, the order of the ℓ_1 *and spectrum norms of the error of (46) for the best data dependent penalty level* λ *is larger than that of the scaled Lasso by a factor* $\sqrt{m}$*.*

Acknowledgments

This research is partially supported by National Science Foundation Grants DMS-11-06753 and DMS-12-09014.

Appendix A.

We provide all proofs in this appendix. We first prove the results in Section 4 since they are used to prove the results in Section 3.

A.1 Proof of Proposition 5

Lemma 20 in Ye and Zhang (2010) gives $\|v\|_q^q \le \|v\|_{(q,m)}^q + (a_q/m)^{q-1}\|v\|_1^q$. The rest of part (i) follows directly from definition. Lemma 20 in Ye and Zhang (2010) also gives $\|v\|^*_{(q,m)} \le \|v\|_{(q',m/a_q)} + m^{-1/q}\|v\|_1$. The rest of part (ii) is dual to the corresponding parts of part (i). Since $\|\overline{\Sigma} v\|_{(2,m)} = \max_{\|u\|_0 = m, \|u\|_2 = 1} u^T \overline{\Sigma} v$ and $\|u^T \overline{\Sigma}^q\|_2 \le \kappa_+^q(m; \overline{\Sigma})$ for $q \in \{1/2, 1\}$, part (iii) follows. □

A.2 Proof of Theorem 6

By the Karush-Kuhn-Tucker conditions,

$$(X^T X/n) h = z - \lambda g, \ \mathrm{sgn}(\widehat{\beta}_j) g_j \in \{0, 1\}, \ \|g\|_\infty \le 1. \tag{47}$$

Since $\zeta_{(2,m)}(z_B, \lambda_*)$ is the $\|\cdot\|_{(2,m)}$ norm of $(|z_B| - \lambda_*)_+$, (21) implies

$$|h^T z| \quad \le \quad \lambda_* \|h\|_1 + \sum_{j \in B} |h_j| (|z_j| - \lambda_*)_+$$

$$\leq \quad \lambda_* \|h\|_1 + A_1 \lambda_* m^{1/2} \|h_B\|^*_{(2,m)}. \tag{48}$$

Since $-h_j \text{sgn}(\widehat{\beta}_j) \leq |\beta_j| - |\widehat{\beta}_j| \leq \min(|h_j|, -|h_j| + 2|\beta_j|) \ \forall \ j \in S^c$, (19) and (47) yield

$$\begin{aligned} -\lambda h^T g &\leq \lambda \|h_S\|_1 - \lambda \|h_{S^c}\|_1 + 2\lambda \|\beta_{S^c}\|_1 \\ &\leq \lambda \|h_S\|_1 - \lambda \|h_{S^c}\|_1 + 2\lambda\lambda_*(s_* - |S|). \end{aligned}$$

By applying the above bounds to the inner product of h and (47), we find

$$\begin{aligned} \|Xh\|_2^2/n &\leq A_1 \lambda_* m^{1/2} \|h_B\|^*_{(2,m)} + (\lambda_* - \lambda)\|h_{S^c}\|_1 \\ &\quad + (\lambda + \lambda_*)\|h_S\|_1 + 2\lambda\lambda_*(s_* - |S|). \end{aligned}$$

Let $u = (A/\lambda)h$. It follows that when $A\lambda_* \leq \lambda$,

$$\frac{\|Xu\|_2^2}{n} \leq A_1 m^{1/2} \|u_B\|^*_{(2,m)} - (A-1)\|u_{S^c}\|_1 + (A+1)\|u_S\|_1 + 2A(s_* - |S|).$$

Since $X^T X/n = \overline{\Sigma}$, $u \in \mathscr{U}(\overline{\Sigma}, S, B; A, A_1, m, m_1)$ with $m_1 = s_* - |S|$. Since $h = \lambda u/A$ and $s_* = m_1 + |S|$, the conclusion follows from (23) and (24). □

A.3 Proof of Theorem 8

It follows from the scale equivariance of (26) that

$$\{\widehat{\beta}/\sigma^*, \widehat{\sigma}/\sigma^*\} = \{\widehat{b}, \widehat{\phi}\} = \underset{b,\phi}{\arg\min} \Big\{ \|y^* - Xb\|_2^2/(2n\phi) + \lambda_0 \|b\|_1 + \phi/2 \Big\}, \tag{49}$$

where $y^* = y/\sigma^* = Xb^* + \varepsilon^*$ with $b^* = \beta/\sigma^*$ and $\varepsilon^* = \varepsilon/\sigma^*$. Our objective is to bound $\|X(\widehat{b} - b^*)\|_2^2/n$ and $\|\widehat{b} - b^*\|_q$ from the above and $\widehat{\sigma}/\sigma^*$ from both sides. To this end, we apply Theorem 6 to the Lasso estimator

$$\widehat{b}(\lambda) = \underset{b}{\arg\min} \Big\{ \|y^* - Xb\|_2^2/(2n) + \lambda \|b\|_1 \Big\}.$$

Let $z^* = z/\sigma^*$ and $h^*(\lambda) = \widehat{b}(\lambda) - b^*$. Since $\|y^* - Xb^*\|_2^2/n = \|\varepsilon^*\|_2^2/n = 1$,

$$\begin{aligned} 1 - \|y^* - X\widehat{b}(\lambda)\|_2^2/n &= h^*(\lambda)^T X^T (y^* - X\widehat{b}(\lambda))/n + h^*(\lambda)^T z^* \\ &= 2h^*(\lambda)^T z^* - \|Xh^*(\lambda)\|_2^2/n. \end{aligned}$$

Consider $\lambda \geq A\lambda_{*,0}$. Since (21) holds with $\{z, \lambda_*\}$ replaced by $\{z^*, \lambda_{*,0}\}$, we find as in the proof of Theorem 6 that

$$u(\lambda) \quad = \quad h^*(\lambda) A/\lambda \in \mathscr{U}(\overline{\Sigma}, S, B; A, A_1, m, m_1).$$

In particular, (48) gives

$$\begin{aligned} |h^*(\lambda)^T z^*| &\leq \lambda_{*,0} \|h^*(\lambda)\|_1 + A_1 \lambda_{*,0} m^{1/2} \|h^*_B(\lambda)\|^*_{(2,m)} \\ &\leq (\lambda^2/A^2) \Big\{ \|u(\lambda)\|_1 + A_1 m^{1/2} \|u_B(\lambda)\|^*_{(2,m)} \Big\}. \end{aligned}$$

Thus, the definition of M^*_σ in (29) gives

$$|2h^*(\lambda)^T z^* - \|Xh^*(\lambda)\|_2^2/n| < M^*_\sigma s_* \lambda^2.$$

We summarize the calculation in this paragraph with the following statement:

$$\lambda \geq A\lambda_{*,0} \ \Rightarrow\ \left|1 - \|y - X\widehat{b}(\lambda)\|_2^2/n\right| < M^*_\sigma s_* \lambda^2. \tag{50}$$

As in Sun and Zhang (2012), the convexity of the joint loss function in (49) implies

$$(\phi - \widehat{\phi})(\phi^2 - \|y - X\widehat{b}(\phi\lambda_0)\|_2^2/n) \geq 0,$$

so that $\widehat{\phi}$ can be bounded by testing the sign of $\phi^2 - \|y - X\widehat{b}(\phi\lambda_0)\|_2^2/n$. For $(\phi,\lambda) = (\phi_1, \phi_1\lambda_0)$, we have

$$\lambda^2 = \frac{\lambda_0^2}{1+\lambda_0^2 M^*_\sigma s_*} \geq \frac{A^2\lambda_{*,0}^2}{1-\eta_* + A^2\lambda_{*,0}^2 M^*_\sigma s_*} = A^2\lambda_{*,0}^2,$$

which implies $\|y - X\widehat{b}(\phi_1\lambda_0)\|_2^2/n > 1 - \phi_1^2\eta_0 = \phi_1^2$ by (50) and the definition of ϕ_1. This yields $\widehat{\phi} > \phi_1$. Similarly, $\widehat{\phi} < \phi_2$. The error bounds for the prediction and the estimation $\widehat{\beta}$ follow from Theorem 6 due to $A\lambda_{*,0} \leq \phi_1\lambda_0 < \widehat{\phi}\lambda_0 < \phi_2\lambda_0$. □

A.4 Proof of Lemma 9

By Proposition 5, $m^{1/2}\|u_B\|^*_{(2,m)} \leq \|u_B\|_1 + m^{1/2}\|u_B\|_{(2,4m)}$, so that for $u \in \mathscr{U}(\overline{\Sigma}, S, B; A, A_1, m, m_1)$,

$$u^T\overline{\Sigma}u + (A - A_1 - 1)\|u\|_1 \leq 2A\|u_S\|_1 + A_1 m^{1/2}\|u_B\|_2 + 2Am_1.$$

Let $\xi = A/(A - A_1 - 1)$ and $\xi_1 = A_1/(A - A_1 - 1)$. It follows that

$$\begin{aligned}
& (\xi/A)u^T\overline{\Sigma}u + \|u\|_1 \\
\leq\ & 2\xi\|u_S\|_1 + \xi_1 m^{1/2}\|u_B\|_2 + 2\xi m_1 \\
\leq\ & (2\xi|S| + \xi_1 m + 2\xi m_1)^{1/2}\{(2\xi+\xi_1)\|u_{S\cup B}\|_2^2 + 2\xi m_1\}^{1/2} \\
\leq\ & \{(2\xi s_* + \xi_1 m)/c_*\}^{1/2}\{(2\xi+\xi_1)u^T\Sigma u + 2\xi c_* m_1\}^{1/2} \\
\leq\ & \{(s_* \vee m)/c_*\}^{1/2}(2\xi+\xi_1)(u^T\Sigma u + c_* m_1)^{1/2}
\end{aligned} \tag{51}$$

due to $s_* = |S| + m_1$ and $c_*\|u_{S\cup B}\|_2^2 \leq u^T\Sigma u$. In terms of $\{\xi, \xi_1\}$, the condition of the Lemma can be stated as $\lambda^*\{(s_* \vee m)/c_*\}(2\xi+\xi_1)^2 \leq 1/2$. Thus,

$$u^T\Sigma u - u^T\overline{\Sigma}u \leq \lambda^*\|u\|_1^2 \leq u^T\Sigma u/2 + c_* m_1/2. \tag{52}$$

Inserting this inequality back into (51), we find that

$$(\xi/A)u^T\overline{\Sigma}u + \|u\|_1 \leq \{(s_* \vee m)/c_*\}^{1/2}(2\xi+\xi_1)(2u^T\overline{\Sigma}u + 2c_* m_1)^{1/2}.$$

If $(\xi/A)u^T\overline{\Sigma}u + \|u\|_1 \geq (\xi/A)(2u^T\overline{\Sigma}u + 2c_* m_1)/4$, we have

$$(\xi/A)u^T\overline{\Sigma}u + \|u\|_1 \leq \{(s_* \vee m)/c_*\}(2\xi+\xi_1)^2(4A/\xi).$$

Otherwise, we have $(\xi/A)u^T\overline{\Sigma}u+2\|u\|_1 \le (\xi/A)c_*m_1$. Consequently,

$$(\xi/A)u^T\overline{\Sigma}u+\|u\|_1 \le \max\Big\{\{(s_*\vee m)/c_*\}(2\xi+\xi_1)^2(4A/\xi), (\xi/A)c_*(s_*-|S|)\Big\}.$$

This and the definition of $\{M^*_{pred}, M^*_1\}$ yield (32) via

$$\xi M^*_{pred}+M^*_1 \le \max\Big\{\big(1\vee(m/s_*)\big)(2\xi+\xi_1)^2\frac{4}{\xi c_*}, \xi c_*(1-|S|/s_*)/A^2\Big\}.$$

Moreover, (52) gives $c_*\|u_{S\cup B}\|_2^2 \le u^T\Sigma u \le 2u^T\overline{\Sigma}u+2c_*m_1$, so that M^*_σ can be bounded via

$$\begin{aligned} & u^T\overline{\Sigma}u/(s_*A^2)+2\big(\|u\|_1+A_1m^{1/2}\|u_B\|^*_{(2,m)}\big)/(s_*A^2) \\ \le\ & M^*_{pred}+2(1+A_1)\|u\|_1/(s_*A^2)+(A_1/A)\Big\{m/s_*+\|u_B\|_2^2/(s_*A^2)\Big\} \\ \le\ & M^*_{pred}+2(1+A_1)M^*_1/A+(A_1/A)\Big(m/s_*+(2/c_*)M^*_{pred}+2(1-|S|/s_*)/A^2\Big). \end{aligned}$$

This gives (33). If in addition $B=\{1,\ldots,p\}$, then it yields (34)

$$M^*_2 = \sup_{u\in\mathscr{U}}\|u\|_2^2/(s_*A^2) \le (2/c_*)M^*_{pred}+2(1-|S|/s_*)/A^2.$$

This completes the proof. □

The tail probability bound for $\zeta^*(z^*,\lambda_*,m)/\sigma^*$ in part (ii) of Proposition 10 uses the following version of the Lévy concentration inequality in the sphere.

Lemma 17 *Let* $\widetilde{\varepsilon}_m=\sqrt{2/(m-1/2)}\Gamma(m/2+1/2)/\Gamma(m/2)-1$, $U=(U_1,\ldots,U_{m+1})^T$ *be a uniform random vector in* $S^m=\{u\in\mathbb{R}^{m+1}:\|u\|_2=1\}$, $f(u)$ *a unit Lipschitz function in* S^m, *and* m_f *the median of* $f(U)$. *Then,*

$$P\{U_1>x\}\le(1+\widetilde{\varepsilon}_m)P\Big\{N(0,1/(m-1/2))>\sqrt{-\log(1-x^2)}\Big\}, \tag{53}$$

$1<1+\widetilde{\varepsilon}_m<\exp\big(1/(4m-2)^2\big)$, *and*

$$\begin{aligned} P\{f(U)>m_f+x\} &\le P\Big\{U_1>x\sqrt{1-(x/2)^2}\Big\} \\ &\le (1+\widetilde{\varepsilon}_m)P\{N(0,1/(m-1/2))>x\}. \end{aligned} \tag{54}$$

PROOF. Since U_1^2 follows the beta$(1/2,m/2)$ distribution,

$$P\big\{U_1>x\big\}=\frac{\Gamma(m/2+1/2)/2}{\Gamma(m/2)\Gamma(1/2)}\int_{x^2}^1 t^{-1/2}(1-t)^{m/2-1}dt.$$

Let $y=\sqrt{-(m-1/2)\log(1-t)}$. We observe that $-t^{-1}\log(1-t)\le(1-t)^{-1/2}$ by inspecting the infinite series expansions of the two functions. This gives

$$\frac{e^{-y^2/2}dy}{t^{-1/2}(1-t)^{m/2-1}dt}=\frac{t^{1/2}e^{-y^2/2}(m-1/2)^{1/2}}{2(-\log(1-t))^{1/2}(1-t)^{m/2}}\ge 2^{-1}\sqrt{m-1/2}.$$

Since $y = \sqrt{-(m-1/2)\log(1-x^2)}$ when $t = x^2$, it follows that

$$P\{U_1 > x\} \le (1+\widetilde{\varepsilon}_m)\int_{\sqrt{-(m-1/2)\log(1-x^2)}}^{\infty}(2\pi)^{-1/2}e^{-y^2/2}dt.$$

Let $\widetilde{A} = \{u \in S^m : f(u) \le m_f\}$, $H = \{u : u_1 \le 0\}$, and $A_x = \cup_{v\in A}\{u \in S^m : \|u-v\|_2 \le x\}$ for all $A \subset S^m$. Since $u \in \widetilde{A}_x$ implies $f(u) \le m_f + x$ and $P\{U \in \widetilde{A}\} \ge P\{U \in H\}$, the Lévy concentration inequality gives

$$P\{f(U) > m_f + x\} \le P\{U \notin H_x\} = P\Big\{U_1 > x\sqrt{1-(x/2)^2}\Big\}.$$

The second inequality of (54) then follows from $(d/dx)\{-\log\{1-(x^2-x^4/4)\}-x^2\} \ge 0$ for $x^2 \le 2$ and $\|U_1\|_\infty \le 1$.

It remains to bound $1+\widetilde{\varepsilon}_m$. Let $x = m+1/2$. Since

$$\frac{1+\widetilde{\varepsilon}_m}{1+\widetilde{\varepsilon}_{m+2}} = \frac{(m/2)\sqrt{m+3/2}}{(m/2+1/2)\sqrt{m-1/2}} = \frac{(x-1/2)\sqrt{x+1}}{(x+1/2)\sqrt{x-1}},$$

the infinite series expansion of its logarithm is bounded by

$$\log\Big(\frac{1+\widetilde{\varepsilon}_m}{1+\widetilde{\varepsilon}_{m+2}}\Big) = \frac{1}{2}\log\Big(\frac{1+1/x}{1-1/x}\Big) + \log\Big(\frac{1-1/(2x)}{1+1/(2x)}\Big) \le \frac{x^{-3}}{4} + \frac{x^{-5}}{5} + \cdots$$

Since $\{(x-1)^{-2}-(x+1)^{-2}\}/2 = 2x^{-3}+4x^{-5}+\cdots$ by Newton's binomial formula,

$$\log\Big(\frac{1+\widetilde{\varepsilon}_m}{1+\widetilde{\varepsilon}_{m+2}}\Big) \le \{(x-1)^{-2}-(x+1)^{-2}\}/16.$$

This gives $\log(1+\widetilde{\varepsilon}_m) \le 1/\{16(x-1)^2\}$. □

A.5 Proof of Proposition 10

(i) Let $L = L_1(k/p)$. Since $P\{N(0,\sigma^2/n) > \lambda_*\} = k/p$, $\lambda_* = \sigma L/\sqrt{n}$. Since $z_j = x_j^T\varepsilon/n \sim N(0,\sigma^2/n)$, $P\{\zeta_{(2,p)}(z,\lambda_*) > 0\} \le 2k$ and

$$\begin{aligned} E\zeta^2_{(2,p)}(z,\lambda_*) &= p(\sigma^2/n)E(|N(0,1)|-L)_+^2 \\ &= 2p(\sigma^2/n)\int_L^\infty (x-L)^2\varphi(x)dx. \end{aligned}$$

Let $J_k(t) = \int_0^\infty x^k e^{-x-x^2/(2t^2)}dx$. By definition

$$\frac{t^2\int_t^\infty (x-t)^2\varphi(x)dx}{\Phi(-t)} = \frac{t^2\int_0^\infty x^2e^{-tx-x^2/2}dx}{\int_0^\infty e^{-tx-x^2/2}dx} = \frac{\int_0^\infty u^2e^{-u-u^2/(2t^2)}du}{\int_0^\infty e^{-u-u^2/(2t^2)}du} = \frac{J_2(t)}{J_0(t)}.$$

Since $J_{k+1}+J_{k+2}/t^2 = -\int_0^\infty x^{k+1}de^{-x-x^2/(2t^2)} = (k+1)J_k(t)$, we find

$$\frac{J_2(t)}{J_0(t)} = \frac{J_2(t)}{\{J_2(t)+J_3(t)/t^2\}/2+J_2(t)/t^2} \le \frac{1}{1/2+1/t^2}.$$

Thus, $E\zeta^2_{(2,p)}(z,\lambda_*) = 2p(\sigma^2/n)(k/p)L^{-2}J_2(L)/J_0(L) \le 2k\lambda_*^2L^{-4}/(1/2+1/L^2)$.

Since $z_j = x_j^T\varepsilon/n$, $\big(\sum_{j\in B}(|z_j|-\lambda_*)_+^2\big)^{1/2}$ is a function of ε with the Lipschitz norm $\|X_B/n\|_2$. Thus, $\zeta_{(2,m)}(z,\lambda_*)$ is a function of ε with the Lipschitz norm $\max_{|B|=m}\|X_B/n\|_2 = \sqrt{\kappa_+(m)/n}$. In addition, since $\zeta_{(2,m)}(z,\lambda_*)$ is an increasing convex function of $(|z_j|-\lambda_*)_+$ and $(|z_j|-\lambda_*)_+$ are convex in ε, $\zeta_{(2,m)}(z,\lambda_*)$ is a convex function of ε. The mean of $\zeta_{(2,m)}(z,\lambda_*)$ is no smaller than its median. This gives (37) by the Gaussian concentration inequality (Borell, 1975).

(ii) The scaled version of the proof uses Lemma 17 with $m=n-1$ there. Let $U=\varepsilon/\|\varepsilon\|_2$, $z_j^* = x_j^T\varepsilon/(n\sigma^*) = (x_j/\sqrt{n})^TU$ and $z^* = X^T\varepsilon/(n\sigma^*)$. Since $z_j^*\sim U_1$, (53) yields the bound $P\{\zeta_{(2,p)}(z^*,\lambda_{*,0})>0\}\le(1+\varepsilon_n)2k$ and

$$E\zeta^2_{(2,p)}(z^*,\lambda_{*,0}) = pE(|U_1|-\lambda_*)_+^2 \le (1+\varepsilon_n)pE(|N(0,1)|-L)_+^2/(n-3/2).$$

The bound for $E\zeta^2_{(2,p)}(z^*,\lambda_{*,0})$ is then derived as in (i). Lemma 17 also gives

$$\begin{aligned} & P\big\{\pm(\zeta_{(2,m)}(z^*,\lambda_{*,0})-\mu_{(2,m)})>x\sqrt{\kappa_+(m)}\big\} \\ \le\ & (1+\varepsilon_n)P\{|N(0,1/(n-3/2))|>x\} \end{aligned}$$

and

$$\begin{aligned} |E\zeta_{(2,m)}(z^*,\lambda_{*,0})-\mu_{(2,m)}| &\le (1+\varepsilon_n)\sqrt{\kappa_+(m)/(n-3/2)}E\big(N(0,1)\big)_+ \\ &= (1+\varepsilon_n)\sqrt{\kappa_+(m)/\{2\pi(n-3/2)\}}. \end{aligned}$$

The above two inequalities yield (38) and (39). □

A.6 Proof of Theorems 11, 12 and 13

The conclusions follow from Theorems 6 and 8 once (21) is proved to hold with the given probability. In Theorem 11, the tail probability bounds for $\zeta_{(0,p)}$ in Proposition 10 yield (21) with $A_1=0$. In Theorem 12, the moment bounds for $\zeta_{(0,p)}$ in Proposition 10 controls the excess noise in (21). In Theorem 13 (i), we need $A_1\lambda_*m^{1/2}\ge E\zeta_{(0,p)}(z,\lambda_*)+\sigma L_n(\varepsilon/p)\sqrt{\kappa_+(m)}$ by (37), so that the given lower bound of A_1 suffices due to $L_n(\varepsilon)/L_n(k/p) = L_1(\varepsilon/p)/L_1(k/p)$. The proof of Theorem 13 (ii) is nearly identical, with (38) and (39) in place of (38). We omit the details. □

A.7 Proof of Proposition 14

(i) By the ℓ_∞ constraint,

$$\|X^T(\varepsilon - X\widetilde{h}(\lambda))/n\|_{(2,k)} = \|X^T(y-X\widetilde{\beta}(\lambda))/n\|_{(2,k)} \le \lambda\sqrt{k}.$$

Thus, when $\lambda\sqrt{k}\le\|X^T\varepsilon/n\|_{(2,k)}/A$,

$$\|X^T\varepsilon/n\|_{(2,k)}(1-1/A) \le \|X^T\varepsilon/n\|_{(2,k)} - \lambda\sqrt{k} \le \|X^TX\widetilde{h}(\lambda)/n\|_{(2,k)}.$$

Thus, Proposition 5 (iii) gives (40).

(ii) Let $f(x) = (x-\widetilde{L}_1(k/p))_+\wedge 1$ and $z^* = X^T\varepsilon/\|\varepsilon\|_2$. Since $z^*\sim N(0,\Sigma)$ and $\|f(z^*)\|_2$ has unit Lipschitz norm, the Gaussian concentration theorem gives

$$P\Big\{Ef(z^*)-f(z^*)\ge\sqrt{2\|\Sigma\|_2\log(1/\varepsilon)}\Big\}\le\varepsilon.$$

This implies $\mathrm{Var}(f(z^*)) \le 4\|\Sigma\|_2$. Since $Ef^2(z^*) \ge pP\{|N(0,1)| \ge L_1(k/p)\} = 2k$,

$$Ef(z^*) - \sqrt{2\|\Sigma\|_2 \log(1/\varepsilon)} \ge \sqrt{2k - 4\|\Sigma\|_2} - \sqrt{2\|\Sigma\|_2 \log(1/\varepsilon)} \ge \sqrt{k-1}.$$

This gives $P\{N_k \ge k\} \ge P\{f(z^*) > \sqrt{k-1}\} \ge 1-\varepsilon$ due to $N_k \ge f^2(z^*)$. Thus, (41) follows from $\|X^T\varepsilon/n\|_{(q,k)} \ge \sigma^* k^{1/q}\widetilde{L}_n(k/p)$ when $N_k \ge k$. The final conclusion follows from part (i) and large deviation for $(\sigma^*/\sigma)^2 \sim \chi^2_n/n$. □

Lemma 18 *Let $\chi^2_{m,j}$ be χ^2 distributed variables with m degrees of freedom. Then,*

$$E \max_{1\le j\le t} \chi^2_{m,j} \le \left(\sqrt{m} + \sqrt{2\log t}\right)^2, \quad t \ge 1.$$

PROOF. Let $f(t) = 2\log t - \int_0^\infty \min\left(1, tP\{N(0,1) > x\}\right) dx^2$. We first proof $f(t) \ge 0$ for $t \ge 2$. Let $L_1(x) = -\Phi^{-1}(x)$. We have $f(2) \ge 2\log 2 - 1 > 0$ and

$$f'(t) = 2/t - 2\int_{L_1(1/t)}^\infty P\big\{N(0,1) > x\big\} x dx \ge 2/t - 2\int_{L_1(1/t)}^\infty \varphi(x) dx = 0.$$

The conclusion follows from $P\{\chi_{m,j} > \sqrt{m} + x\} \le P\{N(0,1) > x\}$ for $x > 0$. □

A.8 Proof of Theorem 15

Let $h = \widehat{\beta} - \beta$ and $\widehat{\lambda} = \widehat{\sigma}\lambda_0$. Consider $J \subseteq \widehat{S} \setminus S$ with $m \le |J| \le m^*$. For any $j \in \widehat{S}$, it follows from the KKT conditions that $|x_j^T X h/n| = |x_j^T(y - X\widehat{\beta} - \varepsilon)| \ge \widehat{\lambda} - |z_j|$. By the definition of $\kappa^*_+(m^*, S)$ and (25),

$$\begin{aligned} \sum_{j\in J} (\widehat{\lambda} - |z_j|)_+^2 &\le \sum_{j\in J} |x_j^T X h/n|^2 \\ &= (X_J^T X h/n)^T (X_J^T X h/n) \\ &\le \kappa^*_+(m^*, S)\|Xh\|_2^2/n \\ &\le \kappa^*_+(m^*, S) M^*_{pred} s_* \widehat{\lambda}^2. \end{aligned} \tag{55}$$

Since $\zeta_{(2,k)}(z^*_B, \lambda_{*,0})/k^{1/2} \downarrow k$ by Proposition 5 (i), the $\{z^*, \lambda_{0,*}\}$ version of (21) gives $\zeta_{(2,|J|)}(z^*, \lambda_{*,0})/|J|^{1/2} \le \zeta_{(2,m)}(z^*_B, \lambda_{*,0})/m^{1/2} \le \xi_1(A-1)\lambda_{*,0}$. Thus, with $z^*_j = z_j/\sigma^*$,

$$\begin{aligned} \sum_{j\in J} (\widehat{\lambda} - |z_j|)_+^2 &\ge \sum_{j\in J} \Big\{\widehat{\lambda} - \sigma^*\lambda_{*,0} - \sigma^*(|z^*_j| - \lambda_{*,0})_+\Big\}_+^2 \\ &\ge \Big\{|J|^{1/2}(\widehat{\lambda} - \sigma^*\lambda_{*,0}) - \sigma^*\zeta_{(2,|J|)}(z^*_B, \lambda_{*,0})\Big\}_+^2 \\ &\ge |J|\Big\{\widehat{\lambda} - \sigma^*\lambda_{*,0} - \sigma^*\xi_1(A-1)\lambda_{*,0}\Big\}_+^2 \end{aligned}$$

Since $\lambda^2_{*,0}/(\lambda_0\phi_1)^2 = (\lambda_{*,0}/\lambda_0)^2(1+\eta_0) \le (1-\eta_*)/A^2 + \eta_*/A^2 = 1/A^2$, we have $\lambda_{*,0}\sigma^* < \lambda_{*,0}\widehat{\sigma}/\phi_1 \le \widehat{\lambda}/A$. The above inequalities and (55) yield

$$|J| \le \frac{\kappa^*_+(m^*, S) M^*_{pred} s_* \widehat{\lambda}^2}{\big\{\widehat{\lambda} - \sigma^*\lambda_{*,0} - \sigma^*\xi_1(A-1)\lambda_{*,0}\big\}_+^2} < \frac{\kappa^*_+(m^*, S) M^*_{pred} s_*}{\big\{1 - 1/A - \xi_1(1-1/A)\big\}_+^2} \le m^*.$$

Since $\widehat{S}\setminus S$ does not have a subset of size m^*, we have $|\widehat{S}\setminus S| < m^*$ as stated in (43). Let P_B be the projection to the linear span of $\{x_j, j\in B\}$. We have

$$\widehat{\sigma}^2 \ge \|P_{\widehat{S}}^{\perp}y\|_2^2/n = \overline{\sigma}^2 \ge \|P_{S\cup\widehat{S}}^{\perp}y\|_2^2/n = (\sigma^*)^2 - \|P_{S\cup\widehat{S}}\varepsilon\|_2^2/n,$$
$$\|X\overline{h}\|_2^2/n = \|P_{\widehat{S}}y - X\beta\|_2^2/n = \|P_{\widehat{S}}\varepsilon\|_2^2/n + \|P_{\widehat{S}}^{\perp}X\beta\|_2^2/n. \tag{56}$$

Let $N = \binom{p}{m^*}$. We have $\log N \le m^*\log(ep/m^*)$ by Stirling. By Lemma 18,

$$E\|P_{S\cup\widehat{S}}\varepsilon\|_2^2/\sigma^2 \le E\max_{|B|=m^*}\|P_{S\cup B}\varepsilon\|_2^2/\sigma^2 \le |S| + (\sqrt{m^*}+\sqrt{2\log N})^2.$$

Since $\max_{|B|=m}\|P_{S\cup B}\varepsilon\|_2$ is a unit Lipschitz function,

$$\begin{aligned}\|P_{S\cup\widehat{S}}\varepsilon\|_2/\sigma &\le \Big\{|S| + \big(\sqrt{m^*}+\sqrt{2m^*\log(ep/m^*)}\big)^2\Big\}^{1/2} + L_1(\varepsilon/p)\\ &\le \sqrt{M^*_{lse}s_*\log(p/s_*)}\end{aligned}$$

with probability ε/p. In addition, Theorem 8 gives $\|P_{\widehat{S}}^{\perp}X\beta\|_2^2 \le \|X\widehat{\beta}-X\beta\|_2^2 \le (1+\varepsilon_0)^2M^*_{pred}s_*(\sigma\phi_2\lambda_0)^2$. Inserting these bounds into (56) yields (44). □

Lemma 19 *Suppose that the rows of $X\in\mathbb{R}^{n\times p}$ are iid $N(0,\Sigma)$ random vectors.*
(i) Let $Y = \mathrm{trace}(AX'X/n)$ and $\sigma^2 = \mathrm{trace}\{(A+A')\Sigma(A+A')\Sigma\}/2$ with a deterministic matrix A. Then, $EY = \mu = \mathrm{trace}(A\Sigma)$, $\mathrm{Var}(Y) = \sigma^2/n$ and

$$E\exp\Big\{t(Y-\mu)\Big\} \le \exp\Big\{-\frac{t\sigma}{\sqrt{2}} - \frac{n}{2}\log(1-\sqrt{2}t\sigma/n)\Big\}.$$

Consequently, for $0 < x \le 1$,

$$P\Big\{(Y-\mu)/\sigma > x\Big\} \le \exp\Big\{-\frac{n}{2}\Big(\sqrt{2}x - \log(1+\sqrt{2}x)\Big)\Big\} \le e^{-nx^2/4}.$$

(ii) Let R^* and $\overline{R}$ be the population and sample correlation matrices of X. Then,

$$P\Big\{|\overline{R}_{jk}-R^*_{jk}| > x\sqrt{1-(R^*_{jk})^2}\Big\} \le 2P\Big\{|t_n| > n^{1/2}x\Big\}$$

where t_n has the t-distribution with n degrees of freedom. In particular, for $n\ge 4$,

$$P\Big\{|\overline{R}_{jk}-R^*_{jk}| > \sqrt{2}x\Big\} \le 2e^{1/(4n-2)^2}P\big\{|N(0,1/n)| > x\big\},\ 0\le x\le 1.$$

PROOF. (i) This part can be proved by computing the moment generating function with $t\sigma/n = x/(1+\sqrt{2}x)$. We omit details. For $0 < x < 1$,

$$f(x) = \frac{\sqrt{2}x - \log(1+\sqrt{2}x)}{x^2} = \int_0^{\sqrt{2}x}\frac{udu}{x^2(1+u)} = \int_0^{\sqrt{2}}\frac{udu}{1+xu} \ge f(1) > 1/2.$$

(ii) Conditionally on $\overline{\Sigma}_{kk}$, $\overline{\Sigma}_{jk}/\overline{\Sigma}_{kk} \sim N(\Sigma_{jk}/\Sigma_{kk}, (1-(R^*_{jk})^2)\Sigma_{jj}/(n\overline{\Sigma}_{kk}))$. Thus,

$$z_{jk} = \Big(\frac{n\overline{\Sigma}_{kk}}{(1-(R^*_{jk})^2)\Sigma_{jj}}\Big)^{1/2}\Big(\frac{\overline{\Sigma}_{jk}}{\overline{\Sigma}_{kk}} - \frac{\Sigma_{jk}}{\Sigma_{kk}}\Big)$$

$$= \Big(\frac{n}{1-(R^*_{jk})^2}\Big)^{1/2}\Big(\overline{R}_{jk}\frac{\overline{\Sigma}_{jj}^{1/2}}{\Sigma_{jj}^{1/2}} - R_{jk}\frac{\overline{\Sigma}_{kk}^{1/2}}{\Sigma_{kk}^{1/2}}\Big)$$

is a $N(0,1)$ variable independent of $\overline{\Sigma}_{kk}$. Consequently,

$$\frac{n^{1/2}|\overline{R}_{jk}-R_{jk}|}{(1-(R^*_{jk})^2)^{1/2}} = \frac{|z_{jk}+z_{kj}|}{\overline{\Sigma}_{jj}^{1/2}/\Sigma_{jj}^{1/2}+\overline{\Sigma}_{kk}^{1/2}/\Sigma_{kk}^{1/2}} \le |t_{jk}|\vee|t_{kj}|$$

with $t_{jk} = z_{jk}\Sigma_{kk}^{1/2}/\overline{\Sigma}_{kk}^{1/2} \sim t_n$. Let U_1 be a uniformly distributed variable in the unit sphere of $\mathbb{R}^{n+1}$. Since $t_n^2/n \sim U_1^2/(1-U_1^2)$, Lemma 17 provides

$$P\big\{t_n^2/n > e^{x^2}-1\big\} = P\big\{U_1^2 > 1-e^{-x^2}\big\} \le 2e^{1/(4n-2)^2}P\Big\{N(0,1/(n-1/2)) > x\Big\}.$$

The conclusion follows from $e^{x^2 n/(n-1/2)}-1 \le 2x^2$ for $0<x\le 1$. □

A.9 Proof of Proposition 4

Since $\widetilde{\Theta}_{jj} = 1/\widehat{\sigma}_j^2$ and $\max\{C_0\lambda_0, C_1 s_{*,j}\lambda_0^2\} \le 1/4$, (17) and the condition on σ_j^* implies

$$|\widetilde{\Theta}_{jj}/\Theta^*_{jj}| \le (5/4)^3 \le 2,\ |\widetilde{\Theta}_{jj}/\Theta^*_{jj}-1| \le \{(5/4)^2C_0+5/4+1\}\lambda_0.$$

It follows from (7), (17) and the condition on $\overline{D} = \mathrm{diag}(\overline{\Sigma}_{jj}, j\le p)$ that

$$\begin{aligned}\big\|\widetilde{\Theta}_{*,j}-\Theta^*_{*,j}\big\|_1 &= \big\|-\widehat{\beta}_{*,j}\widetilde{\Theta}_{jj}-\Theta^*_{*,j}\big\|_1\\ &\le \big\|(\widehat{\beta}_{-j,j}-\beta_{-j,j})\widetilde{\Theta}_{jj}\big\|_1 + \big\|\Theta^*_{*,j}(\widetilde{\Theta}_{jj}/\Theta^*_{jj}-1)\big\|_1\\ &\le \big\|\widehat{D}_{-j}^{-1/2}\big\|_\infty\big\|\widehat{D}_{-j}^{1/2}(\widehat{\beta}_{-j,j}-\beta_{-j,j})\big\|_1\widetilde{\Theta}_{jj} + \big\|\Theta^*_{*,j}\big\|_1\big|\widetilde{\Theta}_{jj}/\Theta^*_{jj}-1\big|\\ &\le (5/2)\Theta^*_{jj}\big\|D_{-j}^{-1/2}\big\|_\infty(\Theta^*_{jj})^{-1/2}C_2 s_{*,j}\lambda_0 + \big\|\Theta_{*,j}\big\|_1\{(3/2)C_0+5/2\}\lambda_0\\ &\le C\Big\{(\big\|D_{-j}^{-1}\big\|_\infty\Theta^*_{jj})^{1/2}s_{*,j}\lambda_0 + \big\|\Theta_{*,j}\big\|_1\lambda_0\Big\}\end{aligned}$$

with $C = \max(5C_2/2, 3C_0/2+5/2)$. This gives (12) due to $\|\widehat{\Theta}-\Theta^*\|_1 \le 2\|\widetilde{\Theta}-\Theta^*\|_1$ by (9). Similarly,

$$\begin{aligned}\big\|\widetilde{\Omega}_{*,j}-\Omega^*_{*,j}\big\|_1 &= \big\|-\widehat{D}_{-j}^{1/2}\widehat{\beta}_{*,j}\widetilde{\Theta}_{jj}\widehat{D}_{jj}^{1/2}-\Omega^*_{*,j}\big\|_1\\ &\le \big\|\widehat{D}_{-j}^{1/2}(\widehat{\beta}_{-j,j}-\beta_{-j,j})\big\|_1\widetilde{\Theta}_{jj}\widehat{D}_{jj}^{1/2}\\ &\quad+\big\|\widehat{D}_{-j}^{1/2}D_{-j}^{-1/2}\Omega^*_{*,j}(\widetilde{\Theta}_{jj}/\Theta^*_{jj})(\widehat{D}_{jj}/D_j)^{1/2}-\Omega^*_{*,j}\big\|_1\\ &\le C\big\{(\Theta^*_{jj})^{-1/2}s_{*,j}\lambda_0\Theta^*_{jj}D_{jj}^{1/2}+\big\|\Omega_{*,j}\big\|_1\lambda_0\big\}\end{aligned}$$

This gives (13) due to $D_{jj}\Theta^*_{jj} = \Omega^*_{jj}$. We omit an explicit calculation of C.

Let $\chi^2_{n,j} = n\Theta^*_{jj}(\sigma^*_j)^2$. When $\chi^2_{n,j}\sim\chi^2_n$, we have

$$|\widetilde{\Theta}_{jj}/\Theta^*_{jj}-1| \le \{(5/4)^2+5/4\}C_1 s_{*,j}\lambda_0^2 + (4/3)|\chi^2_{n,j}/n-1|$$

It follows from Lemma 19 that $P\{|\chi^2_{n,j}/n-1| > \sqrt{2}x\} \le 2e^{-nx^2/4}$ for $x \le 1$. Let $a_j = \|\Theta_{*,j}\|_1$, $t = \max\{M\max_j a_j/\sqrt{n}, \tau_n(\Theta^*)\}$ and $B_0 = \{j : a_j \le \sqrt{8}t\}$. By definition $t \le M\tau_n(\Theta^*)$ and $nt^2/a_j^2 \ge M^2$. It follows that

$$
\begin{aligned}
P\Big\{\max_j |\chi^2_{n,j}/n-1|a_j > 4t\Big\} &\le |B_0|e^{-n/4} + \sum_{j\notin B_0} \exp(-2nt^2/a_j^2) \\
&\le pe^{-n/4} + e^{-M^2}\sum_{j\notin B_0} \exp\Big(-n\tau^2(\Theta^*)/a_j^2\Big). \\
&\le pe^{-n/4} + e^{-M^2}.
\end{aligned}
$$

Thus, $\max_j |\widetilde{\Theta}_{jj}/\Theta^*_{jj}-1|a_j = O_P\big(\tau_n(\Theta^*) + \max_j s_{*,j}a_j\lambda_0^2\big)$. □

A.10 Proof of Theorem 2

We need to verify conditions (17) and (18) in order to apply Proposition 4. Since $\Theta^*_{jj}(\sigma^*_j)^2 \sim \overline{\Sigma}_{jj}/\Sigma^*_{jj} \sim \chi^2_n/n$, (18) follows from Lemma 19 (i) with $\lambda_0 \asymp \sqrt{(\log p)/n}$. Moreover, the condition $P\{(1-\varepsilon_0)^2 \le \chi^2_n/n \le (1+\varepsilon_0)^2\} \le \varepsilon/p$ holds with small ε_0 and ε since $\sqrt{(\log p)/n} = \lambda_0/(2A)$ is assumed to be sufficiently small. We take $\varepsilon_0 = 0$ in (10) since its value does not change the order of $s_{*,j}$.

If we treat $\overline{\Sigma}^{1/2}_{kk}\beta_k$ as the regression coefficient in (4) for the standardized design vector $\overline{\Sigma}^{-1/2}_{kk}x_k$, $k \ne j$, Theorem 11 (ii) asserts that the conclusions of Theorem 8 hold with probability $1-3\varepsilon/p$ for each j, with $\lambda_0 = A\sqrt{4(\log p)/n}$, $A_1 = 0$ and $\varepsilon \asymp 1/\sqrt{\log p}$. By the union bound, the conclusions of Theorem 8 holds simultaneously for all j with probability $1-3\varepsilon$. Moreover, (17) is included in the conclusions of Theorem 8 when M^*_σ and M^*_1 are uniformly bounded in the p regression problems with large probability. Thus, it suffices to verify the uniform boundedness of these quantities.

We use Lemma 9 to verify the uniform boundedness of M^*_σ and M^*_1 with $A_1 = 0$, $B_j = S_j$, $m_j = 0$ and $\{\overline{\Sigma}, \Sigma^*\}$ replaced by $\{\overline{R}_{-j,-j}, R^*_{-j,-j}\}$. Note that the Gram matrix for the regression problem in (4) is $\overline{R}_{-j,-j}$, which is random and dependent on j, so that M^*_σ and M^*_1 are random and dependent on j with the random design. It follows from Lemma 19 (ii) that

$$
\max_{k\ne j} \|\overline{R}_{k,-j} - R^*_{k,-j}\|_\infty \le \max_{j,k} |\overline{R}_{k,j} - R_{k,j}| \le L_n(5\varepsilon/p^2)
$$

with probability $1-\varepsilon$. We may take $L_n(5\varepsilon/p^2) = 2\sqrt{(\log p)/n}$ with $\varepsilon \asymp 1/\sqrt{\log p}$. This yields the first condition of Lemma 9 with $\lambda^* = 2\sqrt{(\log p)/n} \asymp \lambda_0$. The second condition $c_*\|u_S\|_2^2 \le u^T R^*_{-j,-j}u$ follows from (11). The third condition translates to $\max_{j\le p}\lambda_0 s_{*,j} \le c_0$, which is imposed in Theorem 2. Thus, all conditions of Lemma 9 hold simultaneously for all j with large probability. The proof is complete since the conclusions of Lemma 9 with $m = m_j = 0$ guarantee the uniform boundedness of M^*_σ and M^*_1. □

A.11 Proof of Theorem 3

The proof is parallel to that of Theorem 2. Since the smaller $\lambda_{*,0} = L_{n-3/2}(k/p)$ is used, we need to apply Theorem 13 (ii) with $A_1 > 0$, $m = m_j > 0$ and typically much larger B_j than S_j. Since the condition $m_j \le C_0 s_{*,j}$ is impose in (15), the conclusions of Lemma 9 still guarantee the uniform boundedness of M^*_σ and M^*_1. The verification of the conditions of Lemma 9 is identical to the case of larger $\lambda_{*,0}$ in Theorem 2. The only difference is the need to verify that condition (15) uniformly guarantees the condition on A_1 in Theorem 13 (ii), where κ_+/m has the interpretation of $\kappa_+(m_j;\overline{R}_{-j,-j})/m_j$, which depends on j and random $\overline{R}$. Anyway, it suffices to verify $\kappa_+(m_j;\overline{R}_{-j,-j})/m_j \le \psi_j$ simultaneously for all j with large probability.

We verify $\kappa_+(m_j;\overline{R}_{-j,-j})/m_j \le \psi_j$ with the same argument as in Lemma 9. For any vector u with $\|u\|_0 = m_j$ and $\|u\|_2 = 1$, it holds with probability $1-\varepsilon$ that

$$\left|u^T(\overline{R}_{-j,-j} - R_{-j,-j})u\right| \le \max_{j,k} |\overline{R}_{k,j} - R_{k,j}| \sum_{j,k} |u_j u_k| \le L_n(5\varepsilon/p^2) m_j.$$

Thus, it follows from the definition of $\kappa_+(m;\Sigma)$ in (16) that $\kappa_+(m_j;\overline{R}_{-j,-j})/m_j \le \kappa_+(m_j;R_{-j,-j})/m_j + L_n(5\varepsilon/p^2) = \psi_j$ for all j. This completes the proof. □

A.12 Proof of Example 1

(i) Let $s_{*,j} = d_j = \#\{k : \Theta^*_{jk} \ne 0\} \le m+1$. We have $\max_j(1+s_{*,j})\lambda_0 \le (m+2)\lambda_0 \to 0$. Let $B_j = \{k \ne j : \Theta^*_{kj} \ne 0\}$. Since $B_j = J_1 \setminus \{j\}$ for $j \in J_1$, (11) holds with

$$\inf\left\{u^T(R^*_{-j,-j})u/\|u_{B_j}\|_2^2 : u_{B_j} \ne 0\right\} \ge 1-\rho_2 \to 1.$$

Thus, Theorem 2 is directly applicable to this example.

Next, we calculate the error bound in (12) and (14). Since $d_j(\Theta^*_{jj})^{1/2} = 2/(1-\rho_1^2)^{1/2} = 2m$ for $j \in J_1$ and $d_j(\Theta^*_{jj})^{1/2} \le (m+1)/(1-\rho_2^2)^{1/2} \le 2m$ for $j \in J_2$,

$$(\|D^{-1}_{-j}\|_\infty \Theta^*_{jj})^{1/2} s_{*,j}\lambda_0 = (\Theta^*_{jj})^{1/2} s_{*,j}\lambda_0 \le 2m\lambda_0.$$

In addition, $\|\Theta_{*,j}\|_1 \le 2m^2$ for $j \in J_1$ and $\|\Theta_{*,j}\|_1 \le (1+\rho_2\|v\|_1)/(1-\rho_2^2) \le 3/2 + o(1)$ for $j \in J_2$, so that for $t = \sqrt{(2/n)\log p}$,

$$\sum_j \exp(-nt^2/\|\Theta_{*,j}\|_1^2) \le 2\exp\Big(-\frac{2\log p}{4m^2}\Big) + p\exp\Big(-\frac{2\log p}{3/2+o(1)}\Big) \to 0.$$

It follows that the quantities in (14) are bounded by

$$\max_{j\le p} s_{*,j}\|\Theta^*_{*,j}\|_1\lambda_0^2 \le 2(m\lambda_0)^2,\ \tau_n(\Theta^*) \le \sqrt{(2/n)\log p} \le \lambda_0/(A\sqrt{2}).$$

Since $m\lambda_0 \to 0$, the error for the scaled Lasso is of the order $m\lambda_0$ by Theorem 2. The conclusion follows since $L_n(m/p) = (1+o(1))\sqrt{(2/n)\log p}$ when $4m^2 \le \log p$.

(ii)Let $\widetilde{\lambda} = \max_j \|\overline{\Sigma}_{*,j} - \Sigma^*_{*,j}\|_\infty$ and $\widetilde{\lambda}^* = \rho_2/\sqrt{m} + \widetilde{\lambda}$. Since $\mathrm{diag}(\Sigma^*) = I_n$,

$$\widetilde{\lambda} \lesssim L_n(1/p) \ll \rho_2/\sqrt{m}, \quad \widetilde{\lambda}^* = (1+o(1))\rho_2/\sqrt{m} = (1+o(1))c_0 m L_n(m/p).$$

For $\lambda \ge \widetilde{\lambda}^*$, e_3 is feasible for (46) with $j = 3 \in J_2$, so that $\|\widetilde{\Theta}_{*j}(\lambda)\|_1 \le 1$. Since $\|\Theta^*_{J_2,3}\|_1 \ge 1 + m^{1/2}\rho_2$,

$$m^{1/2}\rho_2 \le \inf_{\lambda\ge\widetilde{\lambda}^*} \|\widetilde{\Theta}_{J_2,3}(\lambda) - \Theta^*_{J_2,3}(\lambda)\|_1 \le (m+1)^{1/2} \inf_{\lambda\ge\widetilde{\lambda}^*} \|\widetilde{\Theta}_{J_2,3}(\lambda) - \Theta^*_{J_2,3}(\lambda)\|_2.$$

It follows that for $\lambda \ge \widetilde{\lambda}^*$, $\widetilde{\Theta}(\lambda)$ is suboptimal in the sense of

$$\inf_{\lambda \ge \widetilde{\lambda}^*} \|\widetilde{\Theta}(\lambda) - \Theta^*(\lambda)\|_2 \ge \sqrt{m/(1+m)}\rho_2 = c_0 m^{3/2} L_n(m/p)/\sqrt{1+1/m}.$$

Consider $\lambda \le \widetilde{\lambda}^*$. Let $\beta_{-j,j} = -\Theta^*_{-j,j}/\Theta^*_{jj}$, $\widetilde{\beta}_{-j,j}(\lambda) = -\widetilde{\Theta}_{-j,j}(\lambda)/\widetilde{\Theta}_{jj}(\lambda)$, $\sigma_j = (\Theta^*_{j,j})^{-1/2}$, and $\widetilde{h}_j(\lambda) = \widetilde{\beta}_{-j,j}(\lambda) - \beta_{-j,j}$. By (46), $\|X^T_{-j}(x_j - X_{-j}\widetilde{\beta}(\lambda))/n\|_\infty \le \lambda/\widetilde{\Theta}_{jj}(\lambda)$. Since $m(\log p)/n \to 0$ and $\|\Sigma^*\|_2 \le 2$, $P\{\kappa_+(m;\overline{\Sigma}) \le 3\} \to 1$. Thus, by Proposition 14, there exist positive constants $\{c_1, c_2\}$ such that

$$\min_j P\Big\{ \inf_{\lambda/\widetilde{\Theta}_{jj}(\lambda) \le c_1 \sigma_j L_n(m/p)} \|\widetilde{h}_j(\lambda)\|_2 \ge c_2 \sigma_j \sqrt{m} L_n(m/p) \Big\} \to 1.$$

For $\widetilde{\Theta}_{jj}(\lambda) \ge \Theta^*_{jj}/2$,

$$\begin{aligned} \|\widetilde{h}_j(\lambda)\|_2 &= \|\widetilde{\Theta}_{-j,j}(\widetilde{\lambda})/\widetilde{\Theta}_{jj}(\lambda) - \Theta^*_{-j,j}/\Theta^*_{jj}\|_2 \\ &\le \|\widetilde{\Theta}_{-j,j}(\widetilde{\lambda}) - \Theta^*_{-j,j}\|_2/\widetilde{\Theta}_{jj}(\lambda) + \|\beta_{-j,j}\|_2 |\widetilde{\Theta}_{jj}(\lambda) - \Theta^*_{jj}|/\widetilde{\Theta}_{jj}(\lambda) \\ &\le \|\widetilde{\Theta}_{*,j}(\widetilde{\lambda}) - \Theta^*_{*,j}\|_2 (1 + \|\beta_{-j,j}\|_2)/(\Theta^*_{jj}/2). \end{aligned}$$

For $j = 1$, $\Theta^*_{jj} = m^2$ and $\|\beta_{-j,j}\|_2 = \rho_1$, so that $\|\widetilde{\Theta}_{*,j}(\widetilde{\lambda}) - \Theta^*_{*,j}\|_2 \ge m^2 \|\widetilde{h}_j(\lambda)\|_2/4$ when $\widetilde{\Theta}_{jj}(\lambda) \ge \Theta^*_{jj}/2$. Since $\|\widetilde{\Theta}_{*,j}(\widetilde{\lambda}) - \Theta^*_{*,j}\|_2 \ge m^2/2$ when $\widetilde{\Theta}_{jj}(\lambda) \le \Theta^*_{jj}/2$,

$$\inf_{\lambda \le \widetilde{\lambda}^*} \|\widetilde{\Theta}(\lambda) - \Theta^*\|_2 \ge \min\Big(m^2 \|\widetilde{h}_1(\lambda)\|_2/4, m^2/2 \Big).$$

Pick $0 < c_0 < \min(c_1/2, c_2/4)$. Since $\sigma_1 = (\Theta^*_{11})^{-1/2} = 1/m$,

$$\begin{aligned} & P\Big\{ \inf_{\lambda \le \widetilde{\lambda}^*} \|\widetilde{\Theta}(\lambda) - \Theta^*\|_2 \le \min\big(m^2/2, (c_2/4) m^{3/2} L_n(m/p)\big) \Big\} \\ \le\ & P\Big\{ \widetilde{\lambda}^* > (m^2/2)(c_1/m) L_n(m/p) \Big\} + o(1) = o(1). \end{aligned}$$

Since $L_n(m/p) \to 0$ implies $\min\big(m^2/2, (c_2/4) m^{3/2} L_n(m/p)\big) \ge c_0 m^{3/2} L_n(m/p)$, the conclusion follows. □

References

F. Abramovich and V. Grinshtein. Map model selection in gaussian regression. *Electr. J. Statist.*, 4: 932–949, 2010.

A. Antoniadis. Comments on: ℓ_1-penalization for mixture regression models. *Test*, 19(2):257–258, 2010.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

L. Birge and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3:203–268, 2001.

L. Birge and P. Massart. Minimal penalties for gaussian model selection. *Probability Theory Related Fields*, 138:33–73, 2007.

C. Borell. The brunn-minkowski inequality in gaussian space. *Invent. Math.*, 30:207–216, 1975.

F. Bunea, A. Tsybakov, and M.H. Wegkamp. Aggregation for gaussian regression. *Ann. Statist.*, 35: 1674–1697, 2007.

T. Cai, L. Wang, and G. Xu. Shifting inequality and recovery of sparse signals. *IEEE Transactions on Signal Processing*, 58:1300–1308, 2010.

T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106:594–607, 2011.

E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. on Information Theory*, 51:4203–4215, 2005.

E. J. Candès and T. Tao. The dantzig selector: statistical estimation when p is much larger than n (with discussion). *Annals of Statistics*, 35:2313–2404, 2007.

D. L. Donoho and I. Johnstone. Minimax risk over ℓ_p–balls for ℓ_q–error. *Probability Theory and Related Fields*, 99:277–303, 1994.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9:432–441, 2008.

P. J. Huber and E. M. Ronchetti. *Robust Statistics*, pages 172–175. Wiley, second edition, 2009.

V. Koltchinskii. The dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15:799–828, 2009.

C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrices estimation. *Annals of Statistics*, 37:4254–4278, 2009.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27: 538–557, 2012.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high–dimensional linear regression over ℓ_q–balls. *IEEE Trans. Info. Theory*, 57:6976–6994, 2011.

P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. Model selection in gaussian graphical models: High-dimensional consistency of ℓ_1-regularized MLE. *In Advances in Neural Information Processing Systems (NIPS)*, 21, 2008.

G. Rocha, P. Zhao, and B. Yu. A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice). Technical report, University of California, Berkeley, 2008.

A.J. Rothman, P.J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99:879–898, 2012.

S. van de Geer. The deterministic Lasso. Technical Report 140, ETH Zurich, Switzerland, 2007.

S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

S. Yang and E. D. Kolaczyk. Target detection via network filtering. *IEEE Transactions on Information Theory*, 56(5):2502–2515, 2010.

F. Ye and C.-H. Zhang. Rate minimaxity of theLasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research*, 11:3481–3502, 2010.

M. Yuan. Sparse inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286, 2010.

M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38:894–942, 2010.

C.-H. Zhang and J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

C.-H. Zhang and T. Zhang. A general theory of concave regularization for high dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012.

T. Zhang. Some sharp performance bounds for least squares regression with L_1 regularization. *Ann. Statist.*, 37(5A):2109–2144, 2009.