Supplementary Material for "Fast and Adaptive Sparse Precision Matrix Estimation in High Dimensions"

Weidong Liu and Xi Luo*

September 21, 2014

This supplementary material provides the derivation of the method, additional numerical results, and proof of the main results in the main text.

1 Methodology

For simplicity, we start with a population covariance matrix $\mathbf{\Sigma}$, and we define a covariance loss function for every column $i = 1, 2, \ldots, p$,

$$f_i(\mathbf{\Sigma}, \boldsymbol{B}) = \frac{1}{2}\boldsymbol{\beta}_i^T \mathbf{\Sigma} \boldsymbol{\beta}_i - \boldsymbol{e}_i^T \boldsymbol{\beta}_i, \tag{1}$$

where $\boldsymbol{B} = \big(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p\big)$, and each $\boldsymbol{\beta}_i$ is a column vector. Each function f_i in (1) is strictly convex in $\boldsymbol{\beta}_i$ as $\mathbf{\Sigma}$ is strictly positive-definite; more importantly, the minimal values of each f_i are achieved at $\boldsymbol{\beta}_i$'s that satisfy the following equality, for each i,

$$\mathbf{\Sigma}\boldsymbol{\beta}_i - \boldsymbol{e}_i = \mathbf{0}. \tag{2}$$

The quadratic function (1) is of the same form as the iterative conjugate gradient method that solves a linear system like (2). It is also straightforward to see that each column of the precision matrix $\mathbf{\Omega}$ satisfies these equalities, and thus minimizing all the loss functions in (1). In fact, $\mathbf{\Omega}$ is the unique solution of (2) if $\mathbf{\Sigma}$ is full rank, which can be seen by the inversion formula $\boldsymbol{\omega}_i = \mathbf{\Sigma}^{-1}\boldsymbol{e}_i = \mathbf{\Omega}\boldsymbol{e}_i$.

Certainly, the functions in (1) and the inversion formula cannot be directly applied to data, because the population covariance $\mathbf{\Sigma}$ is usually unknown. Thus, we replace with the sample covariance matrix $\hat{\mathbf{\Sigma}}$, to produce the sample based loss function of (1):

$$f_i(\hat{\mathbf{\Sigma}}, \boldsymbol{B}) = \frac{1}{2}\boldsymbol{\beta}_i^T \hat{\mathbf{\Sigma}} \boldsymbol{\beta}_i - \boldsymbol{e}_i^T \boldsymbol{\beta}_i.$$

One intuition is to minimize the above function, for every i, to produce an estimator for $\mathbf{\Omega}$. However, this is not possible because there may be multiple solutions if $\hat{\mathbf{\Sigma}}$ is not full rank. Moreover, it does not utilize the sparsity assumption of $\mathbf{\Omega}$. We will address these two issues momentarily.

Motivated by recent developments on using the ℓ_1 norm in sparse precision matrix estimation (Friedman, Hastie, and Tibshirani, 2008; Cai, Liu and Luo, 2011), we add the ℓ_1 penalty to the column loss function

$$\frac{1}{2}\boldsymbol{\beta}_i^T \hat{\mathbf{\Sigma}} \boldsymbol{\beta}_i - \boldsymbol{e}_i^T \boldsymbol{\beta}_i + \lambda_{ni} |\boldsymbol{\beta}_i|_1 \tag{3}$$

for $i = 1, 2, \ldots, p$, where the penalization parameter $\lambda_{ni} > 0$ and is allowed to be different from column to column, in order to adapt to the magnitude and sparsity of each column. Due to the shrinkage effect of the ℓ_1 penalty, some coordinates of $\boldsymbol{\beta}_i$ may be shrunk to zero exactly. The loss function (3) is connected to the CLIME estimator (Cai, Liu and Luo, 2011). By taking the subgradient of (3), the minimal values satisfy the following constraint for $i = 1, 2, \ldots, p$,

$$\left| \hat{\mathbf{\Sigma}} \boldsymbol{\beta} - \boldsymbol{e}_i \right|_\infty \leq \lambda_{ni} \tag{4}$$

which is exactly the CLIME constraint.

2 Numerical examples

2.1 Simulations

Table 1 lists the frequencies of correct zero/nonzero identification by SCIO, SCIOcv, CLIME, and glasso, as discussed in the main text.

Table 1: Comparison of average support recovery (SD) of SCIO, SCIOcv, CLIME, and glasso over 100 simulation runs.

TN%

	Decay				Sparse				Block			
p	SCIO	SCIOcv	CLIME	glasso	SCIO	SCIOcv	CLIME	glasso	SCIO	SCIOcv	CLIME	glasso
50	98.57(0.72)	97.22(0.84)	88.84(2.40)	76.18(3.06)	85.16(1.62)	97.73(0.54)	64.60(7.20)	83.16(2.45)	80.60(1.93)	95.67(0.98)	86.69(2.16)	87.40(5.12)
100	99.71(0.13)	98.97(0.21)	96.55(0.68)	86.03(1.60)	91.40(0.44)	98.73(0.20)	79.51(3.89)	86.69(1.16)	97.34(2.44)	98.69(0.32)	95.97(0.65)	96.72(1.27)
200	99.98(0.02)	99.61(0.06)	99.40(0.08)	94.97(2.20)	96.11(0.29)	99.42(0.07)	93.70(1.49)	90.55(0.66)	99.97(0.12)	99.71(0.07)	99.67(0.86)	99.03(0.36)
400	100.00(0.00)	99.84(0.02)	100.0(0.00)	98.90(0.16)	98.66(0.06)	99.72(0.03)	98.62(0.27)	95.60(0.44)	100.00(0.01)	99.94(0.01)	99.98(0.01)	99.68(0.11)
800	100.00(0.00)	99.94(0.01)	100.0(0.00)	95.29(0.06)	100.00(0.00)	99.86(0.01)	100.0(0.00)	94.20(0.38)	100.00(0.00)	99.98(0.00)	100.0(0.00)	95.85(0.06)
1600	100.00(0.00)	99.98(0.00)	100.0(0.00)	97.41(0.11)	100.00(0.00)	99.95(0.00)	100.0(0.00)	96.72(0.26)	100.00(0.00)	99.99(0.00)	100.0(0.00)	97.20(0.02)

TP%

	Decay				Sparse				Block			
p	SCIO	SCIOcv	CLIME	glasso	SCIO	SCIOcv	CLIME	glasso	SCIO	SCIOcv	CLIME	glasso
50	24.19(2.24)	21.60(1.65)	37.21(2.91)	35.92(2.32)	98.71(1.22)	93.27(2.75)	99.88(0.39)	96.00(2.28)	95.18(2.83)	58.26(5.12)	98.50(1.17)	62.45(6.20)
100	12.67(0.52)	13.77(0.76)	21.54(1.37)	26.44(1.37)	77.73(2.12)	75.73(2.50)	97.05(1.06)	83.55(2.66)	31.09(10.94)	41.94(3.33)	85.56(3.07)	48.98(3.48)
200	10.14(0.26)	9.92(0.38)	12.76(0.32)	16.15(3.46)	41.20(1.68)	29.78(1.33)	62.99(3.58)	62.98(1.73)	20.02(0.11)	30.11(1.70)	39.17(2.27)	38.81(3.11)
400	7.14(0.78)	7.84(0.18)	3.46(0.00)	8.81(0.37)	10.68(0.39)	12.03(0.44)	24.17(2.17)	33.83(1.41)	20.00(0.01)	24.63(0.75)	23.70(0.72)	32.15(2.02)
800	3.40(0.00)	6.81(0.08)	3.40(0.00)	15.43(0.24)	2.44(0.00)	5.02(0.16)	2.41(0.01)	25.50(0.79)	20.00(0.00)	22.47(0.36)	19.99(0.02)	52.19(0.56)
1600	3.37(0.00)	6.32(0.05)	3.36(0.00)	12.13(0.25)	1.24(0.00)	1.87(0.05)	1.23(0.80)	12.82(0.52)	20.00(0.00)	21.42(0.22)	19.98(0.02)	47.98(0.31)

2.2 A genetic dataset on HIV-1 associated neurocognitive disorders

Borjabad et al (2011) analyzed gene expression arrays on post-mortem brain tissues. They showed that patients with HAND on antiretroviral therapy have many fewer and milder gene expression changes than untreated patients, and these genes are postulated to regulate several important genetic pathways. Their dataset is publicly available from Gene Expression Ominibus (GEO) under the serial number GSE28160. We here apply our method to study how their genetic interactions/pathways are altered between treated and untreated patients, and compare with other methods using classification, due to lack of the golden truth.

This dataset contains gene expression profiles of post-mortem brain tissues under two biological replications. The first replication contains 6 control (healthy) samples, 7 treated HAND samples, and 8 untreated HAND samples; the second replication contains 3 controls, 5 treated, and 6 untreated. The data are preprocessed by GEO and then log-transformed using Bioconductor in R. We will use the first replications as a training set, and test the performance of classifying 3 classes on the second replications. The class label is denoted by q, where $q = 1, 2, 3$ for control, treated and untreated respectively. The model building procedure is similar to Cai, Liu and Luo (2011). On the training data, we first compare pair-wise mean differences between 3 classes for each gene using Wilcoxon's tests, and select the top 100 genes with the most significant p-values in testing any pair of classes. Based on these 100 genes and the training data, we estimate the inverse covariance matrix $\hat{\boldsymbol{\Omega}}_q$ for each class q using SCIO, CLIME, and glasso. To classify a new observation X from the testing dataset, we employ a classification score for each pair of class (q, q'), which is defined as the log-likelihood difference (ignoring constant factors)

$$s_{q,q'}(\boldsymbol{X}) = -\left(\boldsymbol{X} - \overline{\boldsymbol{X}}_q\right)^T \hat{\boldsymbol{\Omega}}_q \left(\boldsymbol{X} - \overline{\boldsymbol{X}}_q\right) + \left(\boldsymbol{X} - \overline{\boldsymbol{X}}_{q'}\right)^T \hat{\boldsymbol{\Omega}}_{q'} \left(\boldsymbol{X} - \overline{\boldsymbol{X}}_{q'}\right) + \log\det\left(\hat{\boldsymbol{\Omega}}_q\right) - \log\det\left(\hat{\boldsymbol{\Omega}}_{q'}\right)$$

where $\overline{\boldsymbol{X}}_j$ is the mean vector for class j using the training data, $j = q, q'$ and $q \neq q'$. This score is essentially the logarithm of the likelihood ratios under two estimated multivariate normals. Because each class has almost the same number of observations in the training, we will assign the label q if $s_{q,q'} > 0$ and q' otherwise.

Figure 1 plots the support maps with a representing case of 10% connected edges using both SCIO, CLIME, and glasso. Each label has different connection patterns as shown by all these methods, and all methods share similar patterns by visual inspection. However, it should be noted that glasso tends to have stripes in the support, which is also observed in simulations.

2.3 An fMRI dataset on attention deficit hyperactivity disorders

The ADHD-200 project (http://fcon_1000.projects.nitrc.org/indi/adhd200/) released a resting-state fMRI dataset of healthy controls and ADHD children. We apply our method using the data in one of the participating center, Kennedy Krieger Institute. There are 61 typically-developing controls (HC), and 22 ADHD cases. The fMRI data were preprocessed by from neurobureau (http://www.nitrc.org/plugins/mwiki/index.php/neurobureau:AthenaPipeline), and the preprocessing steps are described on the same website. After preprocessing, we have 148 time points from each of 116 brain regions, for each subject. We will use the data of each subject to estimate the precision matrix. We choose the precision matrix instead of the covariance because it is more relevant to direct connections rather than indirect ones.

3 Proof of the main results

To prove the main results, we need the following lemmas. The first one comes from (28) and (33) in Cai, Liu and Luo (2011).

Lemma 1 *Let* $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ *and the sample covariance* $\hat{\mathbf{\Sigma}} = (\hat{\sigma}_{ij})_{p\times p}$. *We have for some* $C > 0$,

$$P\Big(\max_{1\le i,j\le p}\{|\hat{\sigma}_{ij}-\sigma_{ij}|/(\sigma_{ii}^{1/2}\sigma_{jj}^{1/2})\} \ge C\sqrt{\frac{\log p}{n}}\Big) = O(p^{-1})$$

under (C2), and

$$P\Big(\max_{1\le i,j\le p}\{|\hat{\sigma}_{ij}-\sigma_{ij}|/(\sigma_{ii}^{1/2}\sigma_{jj}^{1/2})\} \ge C\sqrt{\frac{\log p}{n}}\Big) = O(p^{-1}+n^{-\delta/8})$$

under (C2).*

Figure 1: Comparison of support recovered by SCIO, CLIME, and glasso for the HIV dataset, when 10% of the edges are connected. Nonzeros are in black.

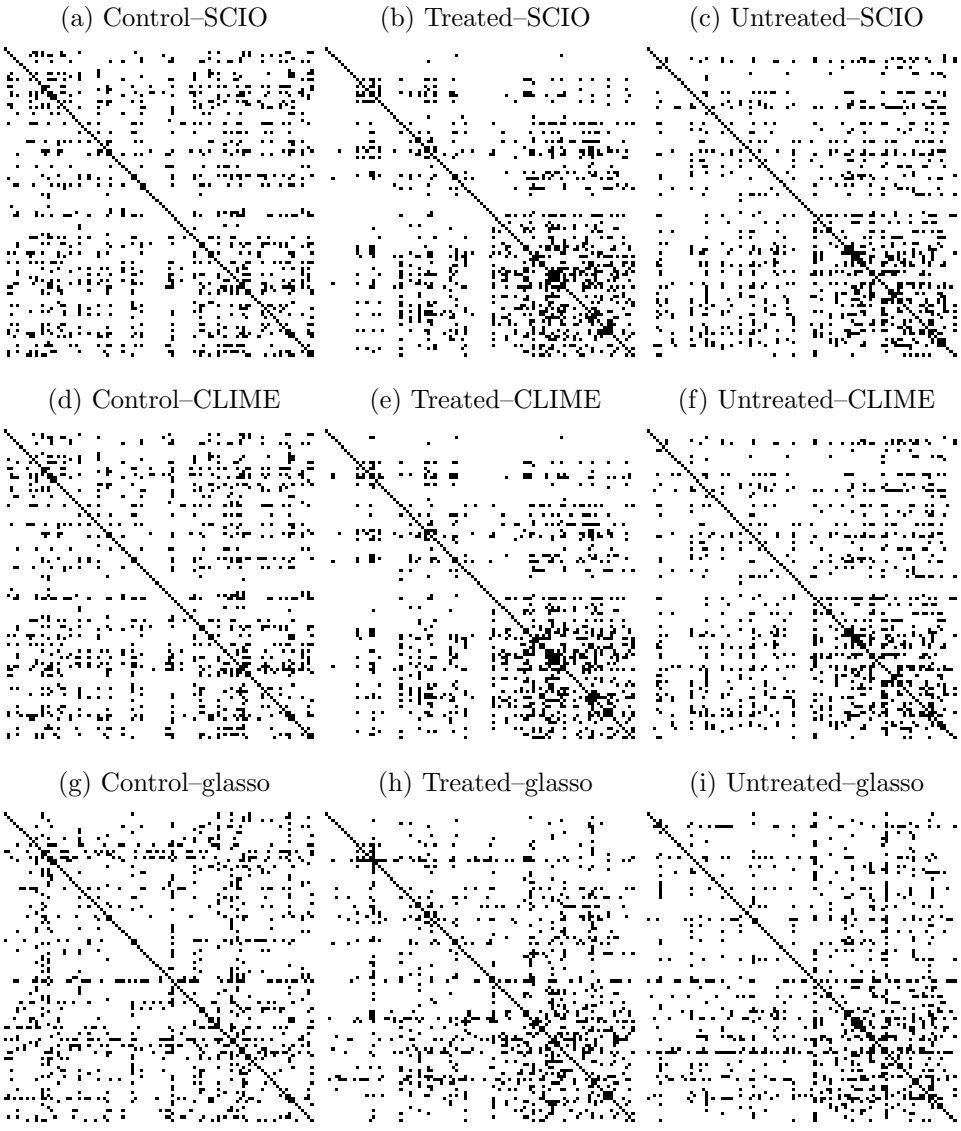
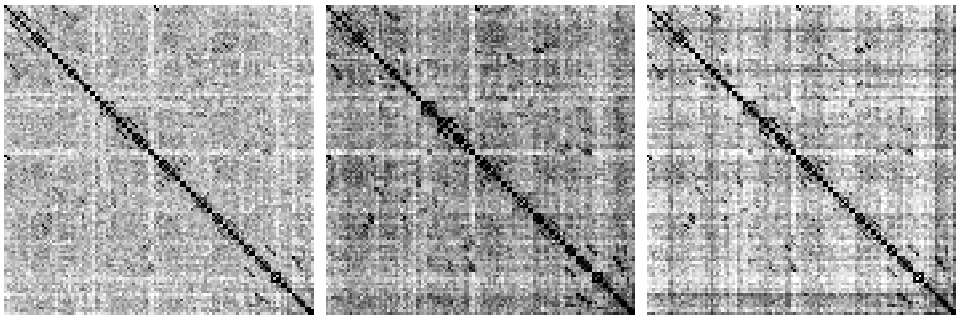
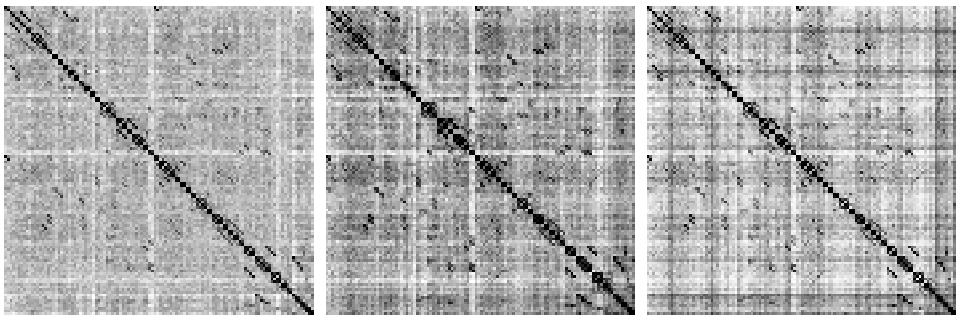


Figure 2: Comparison of support recovered by SCIO, CLIME, and glasso for the ADHD dataset, when 30% of the edges are connected. Black is nonzero over 100% of subjects, and white is 0%.

(a) ADHD–SCIO (b) ADHD–CLIME (c) ADHD–glasso



(d) Control–SCIO (e) Control–CLIME (f) Control–glasso



Let $\mathbf{\Omega} = (\omega_{ij}) = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_p)$, $\mathcal{S}_i$ be the support of $\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_{\mathcal{S}_i} = (\omega_{ji}; j \in \mathcal{S}_i)^T$. We will also need the following lemma from Cai, Liu and Zhou (2011).

Lemma 2 *Assume* $c_0^{-1} \leq \Lambda_{\min}(\mathbf{\Omega}) \leq \Lambda_{\max}(\mathbf{\Omega}) \leq c_0$. *We have for some* $C > 0$,

$$P\Big(\max_{1\leq i\leq p} |\hat{\mathbf{\Sigma}}_{\mathcal{S}_i\times\mathcal{S}_i}\boldsymbol{\omega}_{\mathcal{S}_i} - \boldsymbol{e}_{\mathcal{S}_i}|_\infty \geq C\sqrt{\frac{\log p}{n}}\Big) = O(p^{-1})$$

if (C2) holds;

$$P\Big(\max_{1\leq i\leq p} |\hat{\mathbf{\Sigma}}_{\mathcal{S}_i\times\mathcal{S}_i}\boldsymbol{\omega}_{\mathcal{S}_i} - \boldsymbol{e}_{\mathcal{S}_i}|_\infty \geq C\sqrt{\frac{\log p}{n}}\Big) = O(p^{-1} + n^{-\delta/8})$$

if (C2) holds.*

Proof of Theorem 1. For the solution $\hat{\boldsymbol{\beta}}_i$, it satisfies that

$$\hat{\mathbf{\Sigma}}\hat{\boldsymbol{\beta}}_i - \boldsymbol{e}_i = -\lambda_{ni}\hat{\boldsymbol{Z}}_i,$$

where $\hat{\boldsymbol{Z}}_i =: (\hat{Z}_{1i}, \ldots, \hat{Z}_{pi})^T$ is the subdifferential $\partial|\hat{\boldsymbol{\beta}}_i|_1$ satisfying

$$\hat{Z}_{ji} = \begin{cases} 1, & \hat{\beta}_{ji} > 0; \\ -1, & \hat{\beta}_{ji} < 0; \\ \in [-1, 1], & \hat{\beta}_{ji} = 0. \end{cases}$$

Define $\hat{\boldsymbol{\beta}}_i^o$ be the solution of the following optimization problem:

$$\hat{\boldsymbol{\beta}}_i^o = \mathop{\arg\min}_{\mathbf{supp}(\boldsymbol{\beta})\subseteq\mathcal{S}_i} \Big\{ \frac{1}{2}\boldsymbol{\beta}^T\hat{\mathbf{\Sigma}}\boldsymbol{\beta} - \boldsymbol{e}_i^T\boldsymbol{\beta} + \lambda_{ni}|\boldsymbol{\beta}|_1 \Big\},$$

where $\mathbf{supp}(\boldsymbol{\beta})$ denotes the support of $\boldsymbol{\beta}$. We will show that $\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}}_i^o$ with high probability.

Let $\hat{\boldsymbol{Z}}_{\mathcal{S}_i}^o$ is the subdifferential $\partial|\hat{\boldsymbol{\beta}}_i^o|_1$ on $\mathcal{S}_i$. We define the vector $\tilde{\boldsymbol{Z}}_i = (\tilde{Z}_{1i}, \ldots, \tilde{Z}_{pi})^T$ by letting $\tilde{Z}_{ji} = \hat{Z}_{ji}^o$ for $j \in \mathcal{S}_i$ and

$$\tilde{Z}_{ji} = -\lambda_{ni}^{-1}(\hat{\mathbf{\Sigma}}\hat{\boldsymbol{\beta}}_i^o)_j \quad \text{for } j \in \mathcal{S}_i^c.$$

By Lemma 3 proved momentarily, for $j \in \mathcal{S}_i^c$ and some $r < 1$,

$$|\tilde{Z}_{ji}| \leq r < 1 \tag{5}$$

with probability greater than $1 - O(p^{-1})$ under (C2) (or $1 - O(p^{-1} + n^{-\delta/8})$ under (C2*)). By this primal-dual witness construction and (9), the theorem is proved. ∎

The following lemma is employed when proving Theorem 1.

Lemma 3 *With probability greater than* $1 - O(p^{-1})$ *under (C2) (or* $1 - O(p^{-1} + n^{-\delta/8})$ *under (C2*)), we have*

$$|\tilde{Z}_{ji}| < 1 - \alpha/2$$

uniformly for $j \in \mathcal{S}_i^c$.

Proof. By the definition of $\tilde{\boldsymbol{Z}}_i$, we have

$$\hat{\boldsymbol{\Sigma}}_{\mathcal{S}_i\mathcal{S}_i}\hat{\boldsymbol{\beta}}^o_{\mathcal{S}_i} - \boldsymbol{e}_{\mathcal{S}_i} = -\lambda_{ni}\tilde{\boldsymbol{Z}}_{\mathcal{S}_i} \tag{6}$$

and

$$\hat{\boldsymbol{\Sigma}}_{\mathcal{S}_i^c\mathcal{S}_i}\hat{\boldsymbol{\beta}}^o_{\mathcal{S}_i} \;\; = \;\; -\lambda_{ni}\tilde{\boldsymbol{Z}}_{\mathcal{S}_i^c}. \tag{7}$$

Write (6) as

$$\boldsymbol{\Sigma}_{\mathcal{S}_i\mathcal{S}_i}(\hat{\boldsymbol{\beta}}^o_{\mathcal{S}_i} - \boldsymbol{\omega}_{\mathcal{S}_i}) + (\hat{\boldsymbol{\Sigma}}_{\mathcal{S}_i\mathcal{S}_i} - \boldsymbol{\Sigma}_{\mathcal{S}_i\mathcal{S}_i})(\hat{\boldsymbol{\beta}}^o_{\mathcal{S}_i} - \boldsymbol{\omega}_{\mathcal{S}_i}) + \hat{\boldsymbol{\Sigma}}_{\mathcal{S}_i\mathcal{S}_i}\boldsymbol{\omega}_{\mathcal{S}_i} - \boldsymbol{e}_{\mathcal{S}_i} = -\lambda_{ni}\tilde{\boldsymbol{Z}}_{\mathcal{S}_i}.$$

This implies that

$$\hat{\boldsymbol{\beta}}^o_{\mathcal{S}_i} - \boldsymbol{\omega}_{\mathcal{S}_i} = \boldsymbol{\Sigma}^{-1}_{\mathcal{S}_i\mathcal{S}_i}\Big(- \lambda_n\tilde{\boldsymbol{Z}}_{\mathcal{S}_i} - (\hat{\boldsymbol{\Sigma}}_{\mathcal{S}_i\mathcal{S}_i} - \boldsymbol{\Sigma}_{\mathcal{S}_i\mathcal{S}_i})(\hat{\boldsymbol{\beta}}^o_{\mathcal{S}_i} - \boldsymbol{\omega}_{\mathcal{S}_i}) - \hat{\boldsymbol{\Sigma}}_{\mathcal{S}_i\mathcal{S}_i}\boldsymbol{\omega}_{\mathcal{S}_i} + \boldsymbol{e}_{\mathcal{S}_i}\Big). \tag{8}$$

By (3) of the main text, Lemma 1 and Lemma 2, we have with probability greater than $1 - O(p^{-1})$ (or $1 - O(p^{-1} + n^{-\delta/8})$),

$$|\hat{\boldsymbol{\beta}}^o_{\mathcal{S}_i} - \boldsymbol{\omega}_{\mathcal{S}_i}|_2 \le C\sqrt{s_p \log p/n} + o(1)|\hat{\boldsymbol{\beta}}^o_{\mathcal{S}_i} - \boldsymbol{\omega}_{\mathcal{S}_i}|_2.$$

This implies that

$$|\hat{\boldsymbol{\beta}}^o_{\mathcal{S}_i} - \boldsymbol{\omega}_{\mathcal{S}_i}|_2 \le C\sqrt{s_p \log p/n}. \tag{9}$$

By (7) and the above equation, we have

$$\begin{aligned}
-\tilde{\boldsymbol{Z}}_{\mathcal{S}_i^c} &= \frac{1}{\lambda_n}\hat{\boldsymbol{\Sigma}}_{\mathcal{S}_i^c\mathcal{S}_i}(\hat{\boldsymbol{\beta}}^o_{\mathcal{S}_i} - \boldsymbol{\omega}_{\mathcal{S}_i}) + \frac{1}{\lambda_n}(\hat{\boldsymbol{\Sigma}}_{\mathcal{S}_i^c\mathcal{S}_i} - \boldsymbol{\Sigma}_{\mathcal{S}_i^c\mathcal{S}_i})\boldsymbol{\omega}_{\mathcal{S}_i} \\
&= \frac{1}{\lambda_n}(\hat{\boldsymbol{\Sigma}}_{\mathcal{S}_i^c\mathcal{S}_i} - \boldsymbol{\Sigma}_{\mathcal{S}_i^c\mathcal{S}_i})(\hat{\boldsymbol{\beta}}^o_{\mathcal{S}_i} - \boldsymbol{\omega}_{\mathcal{S}_i}) - \boldsymbol{\Sigma}_{\mathcal{S}_i^c\mathcal{S}_i}\boldsymbol{\Sigma}^{-1}_{\mathcal{S}_i\mathcal{S}_i}\tilde{\boldsymbol{Z}}_{\mathcal{S}_i} \\
&\quad -\frac{1}{\lambda_n}\boldsymbol{\Sigma}_{\mathcal{S}_i^c\mathcal{S}_i}\boldsymbol{\Sigma}^{-1}_{\mathcal{S}_i\mathcal{S}_i}(\hat{\boldsymbol{\Sigma}}_{\mathcal{S}_i\mathcal{S}_i} - \boldsymbol{\Sigma}_{\mathcal{S}_i\mathcal{S}_i})(\hat{\boldsymbol{\beta}}^o_{\mathcal{S}_i} - \boldsymbol{\omega}_{\mathcal{S}_i}) \\
&\quad -\frac{1}{\lambda_n}\boldsymbol{\Sigma}_{\mathcal{S}_i^c\mathcal{S}_i}\boldsymbol{\Sigma}^{-1}_{\mathcal{S}_i\mathcal{S}_i}(\hat{\boldsymbol{\Sigma}}_{\mathcal{S}_i\mathcal{S}_i}\boldsymbol{\omega}_{\mathcal{S}_i} - \boldsymbol{e}_{\mathcal{S}_i}) \\
&\quad +\frac{1}{\lambda_n}(\hat{\boldsymbol{\Sigma}}_{\mathcal{S}_i^c\mathcal{S}_i} - \boldsymbol{\Sigma}_{\mathcal{S}_i^c\mathcal{S}_i})\boldsymbol{\omega}_{\mathcal{S}_i}.
\end{aligned}$$

Since $\|\boldsymbol{\Sigma}_{\mathcal{S}_i^c\mathcal{S}_i}\boldsymbol{\Sigma}^{-1}_{\mathcal{S}_i\mathcal{S}_i}\|_\infty \le 1-\alpha$ and $|\tilde{\boldsymbol{Z}}_{\mathcal{S}_i}|_\infty \le 1$, we have $|\boldsymbol{\Sigma}_{\mathcal{S}_i^c\mathcal{S}_i}\boldsymbol{\Sigma}^{-1}_{\mathcal{S}_i\mathcal{S}_i}\tilde{\boldsymbol{Z}}_{\mathcal{S}_i}|_\infty \le 1-\alpha$. By (9) and Lemma 1, we obtain that with probability greater than $1-O(p^{-1})$ (or $1-O(p^{-1}+n^{-\delta/8})$)

$$|(\hat{\boldsymbol{\Sigma}}_{\mathcal{S}_i^c\mathcal{S}_i} - \boldsymbol{\Sigma}_{\mathcal{S}_i^c\mathcal{S}_i})(\hat{\boldsymbol{\beta}}^o_{\mathcal{S}_i} - \boldsymbol{\omega}_{\mathcal{S}_i})|_\infty \le Cs_p\log p/n. \tag{10}$$

This, together with Lemma 2, implies (5). ∎

Proof of Theorems 2 and 3. By the proof of Theorem 1, we have $\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}}^o_i$. Reorganize terms to yield that

$$\hat{\boldsymbol{\beta}}_i - \boldsymbol{\omega}_i = \boldsymbol{\Sigma}^{-1}\Big(-\lambda_n\hat{\boldsymbol{Z}}_i - (\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\omega}_i) - \hat{\boldsymbol{\Sigma}}\boldsymbol{\omega}_i + \boldsymbol{e}_i\Big). \tag{11}$$

By (9) and Lemma 1, we obtain that with probability greater than $1-O(p^{-1})$ (or $1-O(p^{-1}+n^{-\delta/8})$),

$$|(\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\omega}_i)|_\infty \le Cs_p\log p/n. \tag{12}$$

Thus,

$$|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\omega}_i|_\infty \le CM_p\sqrt{\frac{\log p}{n}}.$$

This proves (6). By (9) and the inequality $\|\hat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\|_F^2 \le 2\sum_{j=1}^p |\hat{\boldsymbol{\beta}}_i - \boldsymbol{\omega}_i|_2^2$, we obtain (7). Theorem 3 (i) follows from the proof of Theorem 1. Theorem 3 (ii) follows from Theorem 2 and the lower bound condition on $\min_{(i,j)\in\Psi}|\omega_{ij}|$. ∎

Proof of Theorem 4. Let

$$\hat{\boldsymbol{\beta}}_i = \arg\min_{\boldsymbol{\beta}\in I\!R^p} \Big\{ \frac{1}{2}\boldsymbol{\beta}^T \hat{\Sigma}_1^1 \boldsymbol{\beta} - \boldsymbol{e}_i^T \boldsymbol{\beta} + \lambda_{ni}|\boldsymbol{\beta}|_1 \Big\}$$

with the theoretical $\lambda_{ni} = C\sqrt{\log p/n} \in \{\lambda_i, 1 \le i \le N\}$ and C is sufficiently large. Then by the proofs of Theorem 1 and 2, we have with probability greater than $1 - O(p^{-1})$,

$$\max_{1\le i\le p} |\hat{\boldsymbol{\beta}}_i - \boldsymbol{\omega}_i|_2^2 \le C s_p \frac{\log p}{n}.$$

By the definition of $\hat{\boldsymbol{\beta}}_i^1$ with the cross validated $\hat{\lambda}_i$, we have

$$\frac{1}{2}(\hat{\boldsymbol{\beta}}_i^1)^T \hat{\boldsymbol{\Sigma}}_2^1 \hat{\boldsymbol{\beta}}_i^1 - \boldsymbol{e}_i^T \hat{\boldsymbol{\beta}}_i^1 \le \frac{1}{2}(\hat{\boldsymbol{\beta}}_i)^T \hat{\boldsymbol{\Sigma}}_2^1 \hat{\boldsymbol{\beta}}_i - \boldsymbol{e}_i^T \hat{\boldsymbol{\beta}}_i.$$

Set $\boldsymbol{D}_i = \hat{\boldsymbol{\beta}}_i^1 - \boldsymbol{\omega}_i$ and $\boldsymbol{D}_i^o = \hat{\boldsymbol{\beta}}_i - \boldsymbol{\omega}_i$. This implies that

$$\begin{aligned} &\langle (\hat{\boldsymbol{\Sigma}}_2^1 - \boldsymbol{\Sigma})\boldsymbol{D}_i, \boldsymbol{D}_i\rangle + \langle \boldsymbol{\Sigma}\boldsymbol{D}_i, \boldsymbol{D}_i\rangle + 2\langle \hat{\boldsymbol{\Sigma}}_2^1 \boldsymbol{\omega}_i - \boldsymbol{e}_i, \hat{\boldsymbol{\beta}}_i^1 - \hat{\boldsymbol{\beta}}_i\rangle \\ &\quad \le \langle (\hat{\boldsymbol{\Sigma}}_2^1 - \boldsymbol{\Sigma})\boldsymbol{D}_i^o, \boldsymbol{D}_i^o\rangle + \langle \boldsymbol{\Sigma}\boldsymbol{D}_i^o, \boldsymbol{D}_i^o\rangle. \end{aligned}$$

Lemma 4 proved later yields that

$$|\langle (\hat{\boldsymbol{\Sigma}}_2^1 - \boldsymbol{\Sigma})\boldsymbol{D}_i, \boldsymbol{D}_i\rangle| = O_P(1)|\boldsymbol{D}_i|_2^2 \sqrt{\frac{\log N}{n}}$$

and

$$\langle \hat{\boldsymbol{\Sigma}}_2^1 \boldsymbol{\omega}_{\cdot i} - \boldsymbol{e}_i, \hat{\boldsymbol{\beta}}_i^1 - \hat{\boldsymbol{\beta}}_i\rangle = O_P(1)|\hat{\boldsymbol{\beta}}_i^1 - \hat{\boldsymbol{\beta}}_i|_2 \sqrt{\frac{\log N}{n}}.$$

Thus,

$$|\boldsymbol{D}_i|_2^2 \le O_P\Big(\sqrt{\frac{\log N}{n}}\Big)(|\boldsymbol{D}_i|_2 + |\hat{\boldsymbol{\beta}}_i - \boldsymbol{\omega}_i|_2) + |\boldsymbol{D}_i^o|_2^2.$$

This proves the theorem. ∎

The following lemma is needed for proving Theorem 4.

Lemma 4 *For any vector* $\boldsymbol{v}$ *with* $|\boldsymbol{v}|_2 = 1$*, we have*

$$\max_{1\le i\le N} |\langle(\hat{\boldsymbol{\Sigma}}_2^1 - \boldsymbol{\Sigma})\boldsymbol{v}, \boldsymbol{v}\rangle| = O_P\left(\sqrt{\frac{\log N}{n}}\right) \tag{13}$$

and

$$\max_{1\le i\le N} |\langle\hat{\boldsymbol{\Sigma}}_2^1\boldsymbol{\omega}_i - \boldsymbol{e}_i, \boldsymbol{v}\rangle| = O_P\left(\sqrt{\frac{\log N}{n}}\right). \tag{14}$$

Proof. We will use the following identity

$$\begin{aligned}\langle(\hat{\boldsymbol{\Sigma}}_2^1 - \boldsymbol{\Sigma})\boldsymbol{v}, \boldsymbol{v}\rangle &= \langle(\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\Sigma}}_2^1\boldsymbol{\Sigma}^{-1/2} - \boldsymbol{I})\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}, \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\rangle \\ &= \langle(\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{\Sigma}}_2^1\boldsymbol{\Sigma}^{-1/2} - \boldsymbol{I})\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}, \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\rangle + (\boldsymbol{v}^T\bar{\boldsymbol{X}} - \boldsymbol{v}^T\boldsymbol{\mu})^2,\end{aligned}$$

where $\tilde{\boldsymbol{\Sigma}}_2^1 = \frac{1}{n_2}\sum_{k=1}^{n_2}(\boldsymbol{X}_k - \boldsymbol{\mu})(\boldsymbol{X}_k - \boldsymbol{\mu})^T$. We have

$$\langle(\tilde{\boldsymbol{\Sigma}}_2^1 - \boldsymbol{\Sigma})\boldsymbol{v}, \boldsymbol{v}\rangle = \frac{1}{n_2}\sum_{k=1}^{n_2}(\boldsymbol{v}^T(\boldsymbol{X}_k - \boldsymbol{\mu}))^2 - \boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}.$$

By (C3) and the exponential inequality in Lemma 1, for any $M > 0$, there exists some $C > 0$ such that

$$\max_{1\le i\le N} P\Big(\Big|\frac{1}{n_2}\sum_{k=1}^{n_2}(\boldsymbol{v}^T(\boldsymbol{X}_k - \boldsymbol{\mu}))^2 - \boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}\Big| \ge C\sqrt{\frac{\log N}{n}}\Big) = O(N^{-M}),$$

$$\max_{1\le i\le N} P\Big(|\boldsymbol{v}^T\bar{\boldsymbol{X}} - \boldsymbol{v}^T\boldsymbol{\mu}| \ge C\sqrt{\frac{\log N}{n}}\Big) = O(N^{-M}).$$

Hence, (13) is proved. (14) follows from the exponential inequality in Lemma 2. ∎

Proof of Proposition 1. The objective is equivalent to (after neglecting constant terms with respect to β_p)

$$\beta_p\boldsymbol{\beta}_{-p}^T\hat{\boldsymbol{\Sigma}}_{12} + \frac{1}{2}\beta_p^2\hat{\Sigma}_{22} - \beta_p 1\{p = i\} + \lambda|\beta_p|.$$

The minimizer then should have a subgradient equal to zero,

$$\boldsymbol{\beta}_{-p}^T\hat{\boldsymbol{\Sigma}}_{12} + \beta_p\hat{\Sigma}_{22} - 1\{p = i\} + \lambda\partial|\beta_p| = 0.$$

Thus the solution is the thresholding rule

$$\beta_p = \mathcal{T}\left(1\{p = i\} - \boldsymbol{\beta}_{-p}^T\hat{\boldsymbol{\Sigma}}_{12}, \lambda\right)/\hat{\Sigma}_{22}.$$

∎

References

[1] Borjabad, A., Morgello, S., Chao, W., Kim, S.-Y., Brooks, A.I., Murray, J., Potash, M.J., and Volsky, D.J. (2011). Significant effects of antiretroviral therapy on global gene expression in brain tissues of patients with HIV-1-associated neurocognitive disorders. *PLoS Pathog* 7(9): e1002213.

[2] Cai, T., Liu, W. and Luo, X. (2011), A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106, 594-607.

[3] Cai, T., Liu, W. and Zhou, H.H. (2011). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation, *Annals of Statistics*, to appear.

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