

## Weighted covariance matrix estimation

Yang, Guangren; Liu, Yiming; Pan, Guangming

2019

Yang, G., Liu, Y., & Pan, G. (2019). Weighted covariance matrix estimation. *Computational Statistics & Data Analysis*, 139, 82–98. doi:10.1016/j.csda.2019.04.017

<https://hdl.handle.net/10356/144459>

<https://doi.org/10.1016/j.csda.2019.04.017>

---

© 2019 Elsevier B.V. All rights reserved. This paper was published in *Computational Statistics & Data Analysis* and is made available with permission of Elsevier B.V.

*Downloaded on 04 Dec 2023 01:03:48 SGT*

# Weighted covariance matrix estimation

Guangren Yang<sup>a</sup>, Yiming Liu<sup>b</sup>, Guangming Pan<sup>c,\*</sup>

<sup>a</sup>Department of Statistics, School of Economics, Jinan University, Guangzhou, China 510632

<sup>b</sup>School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, 637371

<sup>c</sup>School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, 637371

---

## Abstract

The paper proposes a cross-validated linear shrinkage estimation for population covariance matrices. Moreover we also propose a novel weighted estimator based on the thresholding and shrinkage methods for high dimensional datasets. It is applicable to a wider scope of different structures of covariance matrices. Some theoretical results about the cross-validated shrinkage method and weighted covariance estimation methods are also developed. The finite-sample performance of the proposed methods is illustrated through extensive simulations and real data analysis.

*Keywords:* Thresholding, Shrinkage, Adaptive thresholding, Weighted, Bridge function.

*2010 MSC:* 62H05, 62H12, 62H99

---

## 1. Introduction

Estimation of the covariance matrix or its inverse plays a significant role in multivariate statistical analysis. There exists numerous literature investigating such a problem. It is well known that the sample covariance matrix is a consistent estimator of the population covariance matrix when the dimension  $p$  is fixed. However, the dimension of data in an era of big data increases rapidly among many areas, including gene data, e.g., Rothman et al. [14], economy data, e.g., Xue et al. [21], Fan et al. [8] and climate data, e.g., Bickel and Levina [2], etc. In those cases, the convention methods based on the low dimension setting are no longer applicable, especially when  $p$  is much larger than the sample size  $n$ . There are two main approaches in finding a proper covariance matrix estimation in the high dimensional setting.

The first approach is the shrinkage method pioneered by Ledoit and Wolf [10] without assuming a particular structure for the population covariance matrices. There are a series of methods proposed since their paper. The linear shrinkage method proposed by Ledoit and Wolf [10] is to find an optimal trade-off between the identity matrix  $I$  and the sample covariance matrix  $S_n$  so that the expected quadratic loss between their linear combination and the population covariance matrix is minimum. Moreover, using random matrix theory, Ledoit and Wolf ([11, 12]) proposed a class of rotation-equivariant covariance estimator. This class of estimators was motivated by the idea in Stein [15], i.e., pulling up the small eigenvalues of the sample covariance matrix and pulling down the large ones by an amount that is determined individually for each eigenvalue. Fisher and Sun [9] developed an improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance. One can see more extensions in Ledoit and Wolf [13], Wang et al. [19], etc.

The second approach is to propose tailored estimators for the population covariance matrices with some special structures. Among others, Bickel and Levina [1] and Wu and Pourahmadi [20] proposed the regularized estimation for either banding or tapering covariance matrices. Bickel and Levina [2] and El Karoui [7] further proposed the hard thresholding estimation for sparse covariance matrices. Based on the seminal work of Bickel and Levina [2], Rothman et al. [14] consider thresholding estimation with some general thresholding functions. Cai and Liu [4] constructed an

---

\*Corresponding author

Email address: gmpan@ntu.edu.sg (Guangming Pan)

33 adaptive thresholding estimator within a less restrictive set for sparse covariance matrices. For easier presentation, we  
 34 refer to these two kinds of methods as shrinkage and thresholding methods, respectively.

35 Both types of methods have pros and cons. Theoretically, the thresholding method only keeps the large covari-  
 36 ances and omits the small ones, so that the variation of the covariance estimates can be reduced. the corresponding  
 37 consistency in terms of the spectral norm ( $L_2$ ) and the entry-wise maximum norm ( $L_\infty$ ) can be established. While,  
 38 for the shrinkage method, it is applicable for more general settings. For example, the linear shrinkage method aims  
 39 to find an optimal tradeoff between bias and variance. In this case, its consistency in terms of  $L_\infty$  norm is hard to  
 40 achieve. From our simulation results, we also observe that when the sparsity is strong enough (one may see  $\Sigma_2$  defined  
 41 in the section of the simulation studies), the thresholding estimators proposed by Rothman et al. [14] and Cai and Liu  
 42 [4] perform much better than the shrinkage estimators. However, without any special structures, as in Table 4 below  
 43 shows, both linear (Lediot and Wolf [10]) and nonlinear (Lediot and Wolf ([11, 12])) shrinkage estimators outperform  
 44 the thresholding based estimators. However, which type of method is preferred when the underlying covariance ma-  
 45 trices have some weak structure such as  $\Sigma = \Sigma_1 + \Sigma_2$ , where  $\Sigma_1$  has strong sparsity and  $\Sigma_2$  does not have any special  
 46 structure? Is there an alternative method that can be applied in such a case?

47 The main contribution of this article is to propose a cross-validated linear shrinkage estimation for population  
 48 covariance matrices and a new weighted estimator based on the thresholding and shrinkage methods such that it is  
 49 applicable in a wider area. Furthermore, we propose a cross-validated linear shrinkage estimation. Our proposed linear  
 50 shrinkage method performs better than some other shrinkage methods in many cases and enjoys fast computation  
 51 time than nonlinear shrinkage method. In addition, we also propose an implementable algorithm in finding such an  
 52 estimator by cross validation.

53 The rest of the article is organized as follows. Section 2 provides the methodology and the implementable algo-  
 54 rithm. Section 3 introduces the theoretical properties of proposed estimators. Simulation results and real data analysis  
 55 are illustrated in Section 4. We relegate all the proof details to the Appendix.

## 56 2. Methodolgy

57 Before introducing the main results, we first introduce some basic notations for easier description. Let  $X_1, \dots, X_n \in$   
 58  $\mathbb{R}^p$  be i.i.d. random vectors with mean 0 and covariance matrix  $\Sigma$ ,  $X = (X_1, \dots, X_n)^\top \in \mathbb{R}^{n \times p}$  be the corresponding data  
 59 matrix, and  $S_n = X^\top X/n = (s_{ij})$  be the sample covariance matrix. Let  $D = (\sigma_{ii}) = \text{diag}(\Sigma)$  and  $D_n = (s_{ii}) = \text{diag}(S_n)$ ,  
 60 where  $\sigma_{ii}$  and  $s_{ii}$  are the  $i$ -th diagonal entry of  $\Sigma$  and  $S_n$ , respectively. Denote the population correlation matrix and  
 61 the sample correlation matrix respectively by

$$R^\Sigma = (r_{ij}) = D^{-1/2} \Sigma D^{-1/2}, \quad R_n^\Sigma = (r_{n,ij}) = D_n^{-1/2} S_n D_n^{-1/2}.$$

62 Define  $s_0 = \max_{1 \leq i \leq p} \sum_{j=1}^p I(\sigma_{ij} \neq 0)$  and  $s_R = \max_{1 \leq i \leq p} \sum_{j=1}^p I(r_{ij} \neq 0)$  to be the maximum number of nonzero  
 63 elements on each row of the population covariance matrix and the correlation matrix, respectively.

64 Moreover, for any  $A_1, A_2 \in \mathbb{R}^{p \times p}$ ,  $\|A_1\|_F$ ,  $\|A_1\|_2$  and  $\langle A_1, A_2 \rangle = \text{tr}(A_1 A_2^\top)$ , are the Frobenius norm, the spectral  
 65 norm and the inner product of matrices, respectively. Use  $I$  to denote the identity matrix with different dimension in  
 66 different situations. We also set  $\mathcal{T} = \{(i, j) | \sigma_{ij} \neq 0, i \neq j\}$ . Note that under Condition (C3) below  $\mathcal{T}$  is also the same  
 67 as  $\{(i, j) | r_{ij} \neq 0, i \neq j\}$ , and  $s_0$  and  $s_R$  are the same.

### 68 2.1. Cross-validated linear shrinkage

69 This section is to propose a cross-validated based linear shrinkage method. Unlike Ledoit and Wolf [10] we would  
 70 like to propose the shrinkage target to be a diagonal matrix  $D_n$ , which consists of the respective variances of the entries  
 71 of the population vector instead of the identity matrix as in Touloumis [16].

72 Thus, we aim at finding an optimal trade-off between  $D_n$  and  $S_n$  (recall the definition of  $D_n$  in section 2). More  
 73 explicitly we consider the following linear shrinkage strategy:

$$\arg \min_{0 < \tau < 1 - \varepsilon'} E[\|(1 - \tau)D_n + \tau S_n - \Sigma\|_F^2],$$

74 where  $\varepsilon'$  is a small positive constant. In our simulation studies, we take  $\varepsilon' = 0.05$ . From Proposition 1 in Appendix,  
 75 the above optimization problem is equivalent to

$$\arg \min_{0 < \tau < 1 - \varepsilon'} E[\|\tau S_{n,\text{off}} - \Sigma_{\text{off}}\|_F^2], \quad (1)$$

76 where  $A_{\text{off}}$  represents the matrix of  $A$  with the diagonal entries removed. Rewriting (1), it becomes

$$\arg \min_{0 < \tau < 1 - \varepsilon'} \sum_{i \neq j} E(\sigma_{ij} - \tau s_{ij})^2. \quad (2)$$

77 Solving it yields the oracle estimator of  $\tau$ ,

$$\tau^o = \frac{\sum_{i \neq j} \sigma_{ij} E s_{ij}}{\sum_{i \neq j} E s_{ij}^2} = \frac{\sum_{(i,j) \in \mathcal{T}} \sigma_{ij} E s_{ij}}{\sum_{i \neq j} E s_{ij}^2}. \quad (3)$$

78 Hence, the oracle estimator of  $\Sigma$  is given as

$$\Sigma_{sh}^o = D_n + \tau^o S_{n,\text{off}}. \quad (4)$$

79 We use the idea of cross validation to estimate  $\tau^o$ . Denote the set  $X_1, \dots, X_n$  as  $U_1, \dots, U_{n_1}, V_1, \dots, V_{n_2}$ , where  
 80  $n_1 + n_2 = n$ ,  $n_1 = \lfloor cn \rfloor$ ,  $n_2 = n - n_1$  and  $0 < c < 1$  is a constant. We set  $c = 0.6$ , and one can also find  $c$  by  
 81 cross validation. Let  $S_n^{(1)} = (s_{ij}^{(1)}) = \sum_{k=1}^{n_1} U_k U_k^\top / n_1$  and  $S_n^{(2)} = (s_{ij}^{(2)}) = \sum_{k=1}^{n_2} V_k V_k^\top / n_2$ . We propose the following  
 82 optimization problem:

$$\hat{\tau} = \arg \min_{i \neq j} \sum_{i \neq j} (s_{ij}^{(2)} - \tau s_{ij}^{(1)})^2. \quad (5)$$

83 Thus, if  $\hat{\tau} \leq 0$ , then  $\hat{\tau}^e = 0$ , and if  $\hat{\tau} \geq 1 - \varepsilon'$ , then  $\hat{\tau}^e = 1 - \varepsilon'$ , else

$$\hat{\tau}^e = \frac{\sum_{(i,j) \in \mathcal{T}} s_{ij}^{(2)} s_{ij}^{(1)}}{\sum_{i \neq j} (s_{ij}^{(1)})^2}, \quad (6)$$

84 and

$$\widehat{\Sigma}_{sh}^e = D_n + \hat{\tau}^e S_{n,\text{off}}, \quad (7)$$

85 where the set  $\mathcal{T}$  is easy to be estimated by the thresholding method. Indeed, from the proof of the Theorem 2 in  
 86 Rothman et al. (2009) one can see that the set  $\mathcal{T}$  can be detected with probability 1 under the condition  $\sqrt{n}(\tau - \lambda) \rightarrow \infty$ ,  
 87 where  $|\sigma_{ij}| > \tau$  for all  $i, j$  and  $\lambda = O(\sqrt{\log p/n})$ . Our Condition C3 also ensures this condition. This condition does  
 88 not require the signal to be strong and it could tend to zero with a slow convergence rate.

89 In order to make the empirical estimator more precisely, we also give an improved estimator of  $\tau^o$  by permuting  
 90 the data  $X$ . To be specific, let  $X_m := (U_m^\top, V_m^\top)$ , where  $U_m^\top := (U_{1,1}^m, \dots, U_{1,n_1}^m)$  is obtained by randomly sampling  
 91 without replacement  $n_1$  column from original  $X^\top$  and  $V_m^\top$  is the remaining  $n_2$  columns from  $X^\top$ , and hence denote the  
 92 corresponding sample covariance matrix by  $S_{n,m}^{(1)} = (s_{ij,m}^{(1)})$  and  $S_{n,m}^{(2)} = (s_{ij,m}^{(2)})$ , respectively. Define

$$\tilde{\tau}_{sh}^e = \frac{1}{M} \sum_{m=1}^M \frac{\sum_{(i,j) \in \mathcal{T}} s_{ij,m}^{(2)} s_{ij,m}^{(1)}}{\sum_{i \neq j} (s_{ij,m}^{(1)})^2} := \frac{1}{M} \sum_{m=1}^M \hat{\tau}_m^e \quad (8)$$

93 as the average over  $M$  different resamples. In our simulation studies, we take  $M = 50$ . Also, we have the empirical  
 94 estimator of  $\Sigma$ ,

$$\tilde{\Sigma}_{sh}^e = D_n + \tilde{\tau}_{sh}^e S_{n,\text{off}}. \quad (9)$$

95 Compared with the linear shrinkage method (Ledoit and Wolf [10]), nonlinear shrinkage (Ledoit and Wolf [11]) and  
 96 the method proposed in Touloumis [16], our proposed estimator (9) performs much better. One can see more in Table  
 97 3 below.

98 2.2. *Weighted method*

99 This section is to propose a new weighted method based on both the shrinkage and thresholding methods to  
100 estimate the population covariance matrices.

101 We first consider the population correlation matrix as in Cai and Liu [4] and Cui et al. [3]. To this end, define the  
102 set of correlation matrices

$$\mathcal{U}_R^*(c(p), q) = \left\{ R^\Sigma : R^\Sigma > 0, \max_i \sum_{j=1}^p |r_{ij}|^q \leq c(p) \right\}, \quad (10)$$

103 for  $0 \leq q < 1$ . It is easy to see that  $\mathcal{U}_R^*(c(p), q)$  covers the sets proposed by Rothman et al. [14] and Cai and Liu [4].  
104 One may use the thresholding methods if a population correlation matrix belongs to the set  $\mathcal{U}_R^*(c(p), q)$ .

105 The idea behind our weighted method is as follows. Write

$$R^\Sigma = \rho R^\Sigma + (1 - \rho)R^\Sigma, \quad (11)$$

106 where  $0 < \rho < 1$ . We hope to find an appropriate  $\rho$  such that  $\rho R^\Sigma \in \mathcal{U}_R^*(c(p), q)$  even if a population correlation matrix  
107  $R^\Sigma$  does not belong to  $\mathcal{U}_R^*(c(p), q)$  for a given  $c(p)$  and  $q$ . In this manner, once  $\rho$  is found out we can estimate  $\rho R^\Sigma$  and  
108  $(1 - \rho)R^\Sigma$  by the thresholding based method and the shrinkage based method, respectively.

109 To this end, in practice, one has to find a consistent estimator of  $\max_i \sum_{j=1}^p |r_{ij}|^q$  to decide whether a population cor-  
110 relation matrix  $R^\Sigma$  or  $\rho R^\Sigma$  belongs to the set  $\mathcal{U}_R^*(c(p), q)$ . Two natural estimators of  $\max_i \sum_{j=1}^p |r_{ij}|^q$  are  $\max_i \sum_{j=1}^p |r_{n,ij}|^q$   
111 and  $\max_i \sum_{j=1}^p |s_\lambda(r_{n,ij})|^q$ , the corresponding sample estimator and the thresholding version of the sample estimator,  
112 respectively. Here the thresholding function  $s_\lambda : \mathbb{R} \rightarrow \mathbb{R}$  satisfies three conditions  $|s_\lambda(z)| \leq |z|$ ;  $s_\lambda(z) = 0$  for  $|z| \leq \lambda$   
113 and  $|s_\lambda(z) - z| \leq \lambda$  for all  $z \in \mathbb{R}$  (one may see Rothman et al. [14]).

We conduct simulations to investigate the performance of these two estimators, which turn out bad estimators. To  
this end, we consider estimation of two (near) sparsity models  $R^\Sigma = (r_{ij})$  with  $r_{ij} = 0.7^{i-j}$  or  $r_{ij} = (1 - \frac{|i-j|}{5})_+$  when  
 $(p, n) = (100, 100)$ , both of which belong to  $\mathcal{U}_R^*(c(p), q)$  and can use the thresholding method. Table 1 below displays  
the estimated errors of these two estimators, i.e.,

$$\left| \max_i \sum_{j=1}^p |r_{ij}|^q - \max_i \sum_{j=1}^p |r_{n,ij}|^q \right|$$

and

$$\left| \max_i \sum_{j=1}^p |r_{ij}|^q - \max_i \sum_{j=1}^p |s_\lambda(r_{n,ij})|^q \right|.$$

114 From Table 1, we see that for different  $0 < q < 1$ , all the estimated values are not precise enough, especially for small  
 $q$ . Moreover, the selection of  $q$  is also difficult to determine in practice.

Table 1: The estimated errors for different  $q$

	$q$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	$r_{ij} = 0.7^{i-j}$	70.65	55.27	43.56	34.76	28.02	22.80	18.73	15.54	13.07
	$r_{ij} = (1 - \frac{ i-j }{5})_+$	33.22	37.38	34.33	29.85	25.52	21.74	18.57	15.94	13.78
2	$r_{ij} = 0.7^{i-j}$	9.67	6.51	4.37	4.58	4.11	3.49	2.92	2.43	2.09
	$r_{ij} = (1 - \frac{ i-j }{5})_+$	22.69	16.08	11.34	7.92	5.44	3.80	2.77	2.18	2.03

115 We have seen that these two estimators are not precise although using the thresholding estimator may not increase  
116 the order of  $c(p)$  when  $c(p)$  is replaced with  $(c(p) + 2.03)$  if  $q = 0.9$  is used from Table 1. We next see whether these  
117 two estimators can help select an appropriate  $\rho$  so that  $\rho R^\Sigma \in \mathcal{U}_R^*(c(p), q)$ . Recall that Rothman et al. [14] obtained  
118 the convergence rate  $\|s_\lambda(\Sigma_n) - \Sigma\|_2 = O_P(c(p)(\frac{\log p}{n})^{\frac{1-q}{2}})$ . We set below

$$c(p) = C \left( \frac{n}{\log p} \right)^{\frac{1-q}{2} - \epsilon}, \quad (12)$$

120 for  $0 \leq q < 1$ , where  $C > 0$  is a constant and  $0 < \epsilon < \frac{1-q}{2}$  is an arbitrarily small value. This  $c(p)$  is used throughout  
 121 the paper. Define

$$\bar{\rho} = \left( \frac{c(p)}{\max_i \sum_{j=1}^p |r_{ij}|^q} \right)^{\frac{1}{q}}. \quad (13)$$

122 One should note that  $\bar{\rho}$  is the maximum value of  $\rho$  which belongs to  $0 \leq \rho \leq 1$  such that  $\rho R^\Sigma \in \mathcal{U}_R^*(c(p), q)$ . We still  
 123 use  $\max_i \sum_{j=1}^p |r_{n,ij}|^q$  and  $\max_i \sum_{j=1}^p |s_{\lambda}(r_{n,ij})|^q$  to estimate  $\max_i \sum_{j=1}^p |r_{ij}|^q$  so that we are able to obtain the estimators  
 124 of  $\bar{\rho}$  via (13), denoted by  $\rho_1$  and  $\rho_2$ , respectively. Use the same setting as in the example in Table 1. Ideally,  $\rho_1$  and  $\rho_2$   
 125 should be close to one in view of (11) and the (near) sparsity structures of two population covariance matrices used.  
 126 However, from Table 2 one can see that the estimator  $\rho_1$  is still close to 0, and  $\rho_2$  performs slightly better than  $\rho_1$ , but  
 127 still far from one. This indicates the above two estimators fail in selecting an appropriate an appropriate  $\rho$ .

Table 2: The estimator of  $\rho$ .

$q$		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\rho_1$	$r_{ij} = 0.7^{i-j}$	0	0	0	0	0.01	0.03	0.05	0.07	0.09
	$r_{ij} = (1 - \frac{i-j}{5})_+$	0	0	0	0	0.01	0.03	0.05	0.06	0.09
$\rho_2$	$r_{ij} = 0.7^{i-j}$	0	0	0.04	0.09	0.14	0.19	0.23	0.27	0.29
	$r_{ij} = (1 - \frac{i-j}{5})_+$	0	0	0	0.03	0.07	0.13	0.18	0.23	0.26
$\rho_3$	$r_{ij} = 0.7^{i-j}$	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	$r_{ij} = (1 - \frac{i-j}{5})_+$	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87

128 Based on these observations, we see that it may be hard to estimate such a  $\rho$  directly. To avoid the estimation  
 129 of  $\max_i \sum_{j=1}^p |r_{ij}|^q$ , we construct a “bridge function”  $g(s)$ , which depends on  $s_R$ , the maximum of the numbers of the  
 130 nonzero elements in each row of  $R^\Sigma$  such that

$$\max_i \rho^q \sum_{j=1}^p |r_{ij}|^q \leq \rho^q \cdot g(s_R). \quad (14)$$

We treat  $\rho^q \cdot g(s_R)$  as the term  $c(p)$  involved in  $\mathcal{U}_R^*(c(p), q)$ . Thus, we set

$$\rho = \left( \frac{c(p)}{g(s_R)} \right)^{\frac{1}{q}},$$

131 where  $g(s_R) > c(p)$ . In Table 2, we denote the estimator of  $\bar{\rho}$  through “bridge” function by  $\rho_3$ , and we see that it  
 132 performs much better than estimating  $\sum_{j=1}^p |r_{ij}|^q$  directly. To obtain  $\rho_3$  in practice, one can refer to (25) or (27) below.

133 Ideally, we hope to find an optimal  $\rho$  to minimize

$$E \|R^\Sigma - \rho \widehat{R}_1^\Sigma - (1 - \rho) \widehat{R}_2^\Sigma\|_F^2, \quad (15)$$

134 where  $\widehat{R}_1^\Sigma$  and  $\widehat{R}_2^\Sigma$  are the estimator of  $R^\Sigma$  obtained by the thresholding method and new proposed linear shrinkage  
 135 method, respectively.

### 136 2.3. Estimation of the weighted method

137 Let  $L(s_R) = \left( \frac{c(p)}{g(s_R)} \right)^{\frac{1}{q}}$ . In view of (15) we propose the following optimization problem to obtain the corresponding  
 138 oracle estimator of  $g(s)$ :

$$\begin{aligned} g^o(s_R) &= \arg \min_g E \left\| R^\Sigma - L(s_R) \widehat{R}_1^\Sigma - [1 - L(s_R)] \widehat{R}_2^\Sigma \right\|_F^2 \\ \text{subject to} & \quad g(s_R) \geq c(p), \end{aligned} \quad (16)$$

139 where  $\widehat{R}_1^\Sigma = (r_{1,ij})$  is a thresholding based estimator of the correlation matrix and the shrinkage estimator

$$\widehat{R}_2^\Sigma = D_n^{-\frac{1}{2}} \widehat{\Sigma}_{sh}^e D_n^{-\frac{1}{2}} \quad (17)$$

140 with  $\widehat{\Sigma}_{sh}^e$  is given in (7). For definiteness, we use the soft thresholding method, i.e.,

$$r_{1,ij} = s_\lambda(r_{n,ij}) = \text{sign}(r_{n,ij})(|r_{n,ij}| - \lambda)_+, \quad (18)$$

141 where  $\lambda = O(\sqrt{(\log p)/n})$ . Other methods can be similarly discussed. (16) is equivalent with

$$L^o(s_R) = \arg \min_{0 \leq L(s_R) < 1} E \|R^\Sigma - \widehat{R}_2^\Sigma - L(s_R) \cdot (\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma)\|_F^2. \quad (19)$$

142 Solving (19) yields,

$$L^o(s_R) = \frac{E \langle R^\Sigma - \widehat{R}_2^\Sigma, \widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma \rangle}{E \|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2}. \quad (20)$$

143 Recalling (11), we can set the oracle estimator of  $\rho$  to be

$$\rho^o(s_R) = \begin{cases} 1 & \text{if } s_R \leq c(p), \\ L^o(s_R) & \text{else,} \end{cases} \quad (21)$$

144 so that the oracle estimator of  $\Sigma$  is defined as

$$\widehat{\Sigma}_w^o = \rho^o(s_R) D_n^{1/2} \widehat{R}_1^\Sigma D_n^{1/2} + (1 - \rho^o(s_R)) \widehat{\Sigma}_{sh}^e. \quad (22)$$

145 **Remark 1.** For any population correlation matrices  $R$ , we denote its thresholding estimator and new shrinkage  
 146 estimator by  $\widehat{R}_1$  and  $\widehat{R}_2$ , respectively. It is easy to see  $\rho \widehat{R}_2$  is still the unique shrinkage estimator of  $\rho R$  once  $\rho$  is given.  
 147 While, for the thresholding estimator,  $\rho \widehat{R}_1 = s_{\rho\lambda}(\rho R)$ , may not be the optimal one. This is because the threshold  $\rho\lambda$   
 148 may be slightly different. Thus, we also consider the following optimization problem

$$(\rho^o, \widehat{R}_1^o) = \arg \min_{\rho, \widehat{R}_1} \|R - \rho \widehat{R}_2 - \widehat{R}_1\|_F^2 + \lambda \|\widehat{R}_1\|_1,$$

149 which is able to select the threshold automatically. By plenty of simulations, we find that it performs worse than (19)  
 150 in practice.

151 We next proceed to find an estimator of  $\rho^o$ . Motivated by the idea of cross validation we propose the following  
 152 approach to estimate it. Denote  $X_1, \dots, X_n$  by  $U_1, \dots, U_{n_1}, V_1, \dots, V_{n_2}$ , where  $n_1 + n_2 = n$ ,  $U$  and  $V$  have no overlap.  
 153 Use  $R_n^{(1)} = (r_{ij}^{(1)})$  and  $R_n^{(2)} = (r_{ij}^{(2)})$  to denote the corresponding sample correlation matrix according to  $U$  and  $V$ , respec-  
 154 tively. Based on the subset  $U$ , we construct estimators of  $R^\Sigma$  by the thresholding method and the new proposed linear  
 155 shrinkage method, and denoted by  $\widehat{R}_1^{\Sigma^{(1)}}$  and  $\widehat{R}_2^{\Sigma^{(1)}}$ , respectively. Thus, we propose following optimization problem:

$$\widehat{L}^e = \arg \min_L \|R_n^{(2)} - \widehat{R}_2^{\Sigma^{(1)}} - L(\widehat{R}_1^{\Sigma^{(1)}} - \widehat{R}_2^{\Sigma^{(1)}})\|_F^2, \quad (23)$$

156 and solve it to yield

$$\widehat{L}^e = \frac{E \langle R_n^{(2)} - \widehat{R}_2^{\Sigma^{(1)}}, \widehat{R}_1^{\Sigma^{(1)}} - \widehat{R}_2^{\Sigma^{(1)}} \rangle}{E \|\widehat{R}_1^{\Sigma^{(1)}} - \widehat{R}_2^{\Sigma^{(1)}}\|_F^2}. \quad (24)$$

157 It follows that

$$\widehat{\rho}^e(\widehat{\delta}) = \begin{cases} 1 & \text{if } \widehat{\delta} \leq c(p), \\ \widehat{L}^e & \text{else,} \end{cases} \quad (25)$$

158 where  $\widehat{\delta} = \max_{1 \leq i \leq p} \sum_{j=1}^p I(s_\lambda(r_{n,ij}) \neq 0)$  and  $s_\lambda(\cdot)$  is defined in (18). One can see more from Rothman et al. [14]. It is  
 159 easy to see that equations (23)-(25) are the empirical versions corresponding to equations (19)-(21). Thus,

$$\widehat{\Sigma}_w^e = \widehat{\rho}^e(\widehat{\delta}) D_n^{\frac{1}{2}} \widehat{R}_1^\Sigma D_n^{\frac{1}{2}} + (1 - \widehat{\rho}^e(\widehat{\delta})) \widehat{\Sigma}_{sh}^e, \quad (26)$$

160 where  $\widehat{R}_1^{\Sigma}$  and  $\widehat{\Sigma}_{sh}^e$  are given in (18) and (7), respectively. Moreover we can improve the estimator of  $L^o(s_R)$  by taking  
 161 an average of the  $M$  resamples. Let

$$\widetilde{L}^e = \frac{1}{M} \sum_{l=1}^M \widehat{L}_l^e, \quad (27)$$

162 where  $\widehat{L}_l^e$  is obtained from (23) the  $l$ th resamples, and hence  $\widetilde{\rho}^e(s)$  is obtained. The corresponding average version of  
 163  $\widehat{\Sigma}_w^e$  is given as

$$\widetilde{\Sigma}_w^e = \widetilde{\rho}^e(\hat{s}) D_n^{\frac{1}{2}} \widehat{R}_1^{\Sigma} D_n^{\frac{1}{2}} + (1 - \widetilde{\rho}^e(\hat{s})) \widetilde{\Sigma}_{sh}^e, \quad (28)$$

164 where  $\widehat{R}_1^{\Sigma}$  and  $\widetilde{\Sigma}_{sh}^e$  are obtained from whole samples defined as in (18) and (9) below and  $\hat{s}$  has the same definition as  
 165 in (25). In the simulation, we usually take  $n_1 = [0.6n]$ . One can also select the split location,  $n_1$ , by the following  
 166 criterion:

$$h(n_1) = \left\| \frac{1}{M} \sum_{l=1}^M \left( R_{n,l}^{(2)} - \widehat{R}_{2,l}^{\Sigma^{(1)}} - L(\widehat{R}_{1,l}^{\Sigma^{(1)}} - \widehat{R}_{2,l}^{\Sigma^{(1)}}) \right) \right\|_F^2, \quad (29)$$

167 where  $\widehat{R}_{1,l}^{\Sigma^{(1)}}$  and  $\widehat{R}_{2,l}^{\Sigma^{(1)}}$  are the estimators by the thresholding method and the new proposed linear shrinkage estimator  
 168 below according to  $n_1$  samples in the  $l$ -th trial, respectively, and  $R_{n,l}^{(2)}$  is the sample covariance matrix of the remaining  
 169  $n_2 = n - n_1$  samples in the  $l$ -th trail. We suggest to select  $n_1$  over  $(0.4n, 0.5n, 0.6n, 0.7n, 0.8n)$ .

#### 170 2.4. Implementable algorithm

171 This part is to propose an implementable algorithm for the weighted approach. Since we want to obtain (28), we  
 172 conduct  $M = 50$  permutations. For simplicity, in the  $l$ -th permutation, we denote  $\{X_1, \dots, X_n\}$  by  $P$  and randomly  
 173 choose subsets  $P_1 = \{X_1^{(l)}, \dots, X_{n_1}^{(l)}\}$ ,  $P_2 = \{X_{n_1+1}^{(l)}, \dots, X_n^{(l)}\}$  from  $P$ , where  $P_1 \cup P_2 = P$  and  $n_1 = 0.6n$ .

##### 174 Algorithm

- 175 1. By the method proposed by Rothman et al. [14], the estimators of  $\mathcal{T}$  and  $s_0$  can be obtained, denoted by  $\widehat{\mathcal{T}}$  and  
 176  $\hat{s}$ . If  $\hat{s} < c(p)$ , let  $\widetilde{\rho}^e = 1$  and plug it into (28), where  $c(p) = C(n/(\log p))^{(1-q)/2-\epsilon}$  as in (12). In practice, we set  
 177  $C = 2$ ,  $q = 0.1$  and  $\epsilon = 0.05$ .
- 178 2. According to (9) and (18), we can have the corresponding  $\widetilde{\tau}_{sh}^e$ ,  $\widetilde{\Sigma}_{sh}^e$  and  $\widehat{R}_1^{\Sigma}$  from whole samples  $\{X_1, \dots, X_n\}$ .  
 179 For  $l$ -th iteration:
- 180 3. Let  $S_{n,l}^{(1)}$  be the sample covariance matrix based on  $P_1$ , and  $\widetilde{\Sigma}_{2,l}^e = D_{n,l}^{(1)} + \widetilde{\tau}_{sh}^e S_{n,l,off}^{(1)}$  where  $D_{n,l}^{(1)}$  and  $S_{n,l,off}^{(1)}$  is  
 181 the diagonal and the off-diagonal matrix of  $S_{n,l}^{(1)}$  respectively. Hence,  $\widehat{R}_{2,l}^{\Sigma^{(1)}} = (D_{n,l}^{(1)})^{-1/2} \widetilde{\Sigma}_{2,l}^e (D_{n,l}^{(1)})^{-1/2}$  is obtained.  
 182 Also, Let  $R_{n,l}^{(1)} = (D_{n,l}^{(1)})^{-1/2} S_{n,l}^{(1)} (D_{n,l}^{(1)})^{-1/2}$  and  $\widehat{R}_{1,l}^{\Sigma^{(1)}} = s_{\lambda}(R_{n,l}^{(1)})$ , where  $s_{\lambda}(\cdot)$  is the soft thresholding function defined  
 183 in (18).
- 184 4. Let  $S_{n,l}^{(2)}$  be the sample covariance matrix based on  $P_2$ , and denote the corresponding sample correlation matrix  
 185 by  $R_{n,l}^{(2)}$ . Solving  $\widehat{L}_l^e = \arg \min_L \|R_{n,l}^{(2)} - \widehat{R}_{2,l}^{\Sigma^{(1)}} - L(\widehat{R}_{1,l}^{\Sigma^{(1)}} - \widehat{R}_{2,l}^{\Sigma^{(1)}})\|_F^2$ , so that  $\hat{\rho}_l^e$  can be obtained as in (25).
- 186 5. We repeat Step 3 to Step 4 for  $M = 50$  permutations, so that  $\widetilde{\rho}^e = \frac{1}{M} \sum_{l=1}^M \hat{\rho}_l^e$ . Hence, the estimator of the  
 187 population covariance matrix  $\widetilde{\Sigma}_w^e$  is given as in (28).

188 **Remark 2.** One can also use adaptive threshold method in Step 3 to obtain the corresponding estimator. From  
 189 the above algorithm, we see that the estimators  $\hat{s}$ ,  $\widehat{R}_1^{\Sigma}$ ,  $\widehat{R}_2^{\Sigma}$ ,  $\widetilde{\tau}_{sh}^e$  and  $\widetilde{\rho}^e$  is obtained separately. Thus, it is fast when  
 190 implemented this algorithm.

191 **Remark 3.** Since we mainly focus on the covariance matrices estimation we have not mentioned the inverse corre-  
 192 lation matrices estimation in the subsequent parts. But we think the weighted method can be also applicable to the  
 193 inverse correlation matrices estimation. This is because Ledoit and Wolf [10] and Ledoit and Wolf [11] proposed the  
 194 shrinkage method to estimate the inverse correlation matrices, which do not depend on any special structures. On the  
 195 other hand, Cai et al. [5] proposed a kind of CLIME estimator for the inverse correlation matrices that depend on  
 196 some special structures. Thus one can also propose a corresponding method to do the inverse correlation matrices  
 197 estimation by the weighted idea.



198 **3. Theoretical Properties**

199 This section is to consider the asymptotic properties of the new proposed shrinkage estimator and the weighted  
 200 estimator defined in (7) and (26), respectively. Throughout the paper, for two sequence of real numbers  $\{a_n\}$  and  
 201  $\{b_n\}$ , we write  $a_n = O(b_n)$  if there exists a positive constant  $C$  such that  $|a_n| \leq C|b_n|$  holds for all sufficiently large  $n$ ,  
 202  $a_n = o(b_n)$  if  $\lim_{n \rightarrow \infty} a_n/b_n = 0$ ,  $a_n = \Omega(b_n)$  if there exists a constant  $c$  such that  $c|b_n| \leq |a_n|$  holds for all sufficiently  
 203 large  $n$  and  $a_n \asymp b_n$  if there exists a positive constant  $c$  and  $C$  such that  $c|b_n| \leq |a_n| \leq C|b_n|$  holds for all sufficiently  
 204 large  $n$ . Also, the subscript  $P$  in above notation means the corresponding conclusion hold with high probability, i.e.,  
 205 with probability  $1 - o(1)$  hold.

206 We first specify some necessary conditions for further analysis: For  $X_i = (X_{i1}, \dots, X_{ip})^\top$  define  $Y_{ij} = X_{ij} / \sqrt{\text{var}(X_{ij})}$ .

207 **Conditions**

(C1) (Exponential-type tails) Suppose that there exists some  $\eta > 0$  such that

$$E \exp(tY_{ij}) \leq K_1$$

208 for all  $|t| < \eta$  and  $i, j$ .

(C2) (Polynomial-type tails) Suppose that  $X_i = \Sigma^{\frac{1}{2}} Z_i$ , where  $Z_i = (z_{i1}, \dots, z_{ip}) \in \mathbb{R}^p$  and  $z_{ij}$  are i.i.d random variables  
 with mean 0 and finite fourth moment. In addition, for some  $\ell \geq 1$ , there is

$$E|Y_{ij}|^{4\ell} \leq K_1$$

209 for all  $i, j$ .

210 (C3)  $p^\alpha \asymp n$  for  $\alpha > 0$ , assume the nonzero off-diagonal entries  $r_{ij} = \Omega(p^{-\xi})$  and the maximum entry  $\sigma_{ij} = O(p^\zeta)$ , for  
 211  $0 < \xi < \alpha/4$ ,  $\zeta \geq 0$  and  $4\xi + 4\zeta < \alpha$ . Further,  $\min \text{var}(X_{ik}X_{jk}) > c > 0$ .

212 (C4) Let  $|\mathcal{T}| \asymp p^{\gamma_1}$ ,  $|\mathcal{T}^c| \asymp p^{\gamma_2}$  and  $|\mathcal{T}| + |\mathcal{T}^c| = p^2 - p$ , where  $0 \leq \gamma_2 \leq 2$  and  $\gamma_1 < 2 - \alpha - 2\xi - 2\zeta$  or  $2 \geq \gamma_1 >$   
 213  $2 - \alpha + 2\xi + 2\zeta$ . Let  $\theta_1 = \alpha - 4\xi - 2\varepsilon_1 - 4\zeta > 0$ ,  $\theta_2 = \alpha + 2 - \gamma_1 - \varepsilon_1 - 4\zeta > 0$  and  $\theta'_2 = \alpha + \gamma_1 - \gamma_2 - 2\xi - \varepsilon_1 - 2\zeta > 0$   
 214 for some small  $\varepsilon_1 > 0$ .

215 (C5)  $\mu_1 \leq C$ , where  $\mu_1$  is the largest eigenvalue of  $R^\Sigma$ .

216 **Remark 4.** Condition (C1) and (C2) are the moment conditions for  $X$ , which are similar to Cai and Liu [4] and Cui  
 217 et al. [3]. We see that under the Condition (C3),  $\mathcal{T}$  and  $s_0$  can be detected with probability tend to 1, that is

$$P(\mathcal{T} = \widehat{\mathcal{T}}) = 1 - o(1), \quad (30)$$

218 where  $\widehat{\mathcal{T}}$  is the set detected by the method proposed by Rothman et al. [14]. For the proof of (30), one can also refer to  
 219 Theorem 2 in Rothman et al. [14]. Further, the constraint  $4\xi + 4\zeta < \alpha$  keeps  $s_R = s_0$ . Condition  $\min \text{var}(X_{ik}X_{jk}) > c > 0$   
 220 is also proposed in Cai and Liu [4]. From Condition (C4), we see that our proposed new linear shrinkage method  
 221 aims at wider scope population covariance matrices. Additionally, it covers the cases of  $|\mathcal{T}| = 0$  and  $|\mathcal{T}| = p^2 - p$ ,  
 222 i.e., the identity and totally dense cases, respectively. Combining Condition (C3) with (C4), we see that the condition  
 223 between  $p$  and  $n$  is much weaker than that in the shrinkage methods. Condition (C5) states that all eigenvalues of  $R^\Sigma$   
 224 is bounded away from infinity.

225 Set  $\mathcal{A} = \{\mathcal{T} = \widehat{\mathcal{T}}\}$ . In the following we consider all the conclusions when the event  $\mathcal{A}$  happens (otherwise use (30)).  
 226 We first consider convergence of the cross-validated linear shrinkage to the oracle linear shrinkage estimator.

227 **Theorem 1.** Suppose that Conditions (C3), (C4) and (C5) hold.

1. Under Condition (C1),  $\hat{\tau}^\varepsilon$  tends to  $\tau^\circ$  with high probability, and

$$\|\widehat{\Sigma}_{sh}^\varepsilon - \Sigma_{sh}^\circ\|_2 = O_P \left( p^{-\varepsilon_1} \max \left( \sqrt{\frac{p}{n}}, 1 \right) \|\Sigma\|_2 \right),$$

228 for some  $\varepsilon_1 > 0$ , where  $\hat{\tau}^\varepsilon$ ,  $\tau^\circ$ ,  $\widehat{\Sigma}_{sh}^\varepsilon$  and  $\Sigma_{sh}^\circ$  are defined in (6), (3), (7) and (4), respectively.

229 2. Under Condition (C2), the above conclusions still holds as long as  $\ell > \max\{2/\theta_1, 2/\theta_2, 2/\theta'_2\}$ .

230 Comparing with the condition  $p/n \leq K$  in Ledoit and Wolf [10] or  $p/n \rightarrow c \in (0, \infty)$  as  $n$  tends to  $\infty$  in Ledoit  
 231 and Wolf ([11, 12]), our result is less restrictive. For the estimators defined in (8) and (9), we also have

232 **Corollary 1.** *Under conditions of Theorem 1, we have  $\tilde{\tau}^e$  tends to  $\tau^o$  with high probability. Further,  $\|\tilde{\Sigma}_{sh}^e - \Sigma_{sh}^o\|_2 =$   
 233  $O_P(p^{-\varepsilon_1} \max(\sqrt{p/n}, 1) \|\Sigma\|_2)$  or some  $\varepsilon_1 > 0$ , where  $\tilde{\Sigma}_{sh}^e$  is given in (9).*

234 From Corollary 1, we see that  $D_n + \tilde{\tau}^e S_{n,off}$  is a proper shrinkage estimator of  $\Sigma$ , hence  $\widehat{R}_2^\Sigma = D_n^{-1/2}(D_n + \tilde{\tau}^e S_{n,off})D_n^{-1/2}$ .  
 235 Also, by soft thresholding method, we have the estimator  $s_\lambda(R_n^\Sigma)$ , denoted as  $\widehat{R}_1^\Sigma$ . In order to obtain the results of  
 236 weighted estimator, we propose one more condition:

237 **(C6)** Let  $\theta_1^* := \alpha - 8\xi - 4\zeta - 2\varepsilon_1 > 0$ , and  $\theta_2^* := 2 - 4\xi - \gamma_1 - 4\zeta - \varepsilon_1 > 0$  and  $\theta_2^{*'} := 2 - 4\xi - \gamma_2 - 4\zeta - \varepsilon_1 > 0$  for  
 238 some small  $\varepsilon_1 > 0$ .

239 By solving (19), we have the following theorem:

240 **Theorem 2.** *Suppose that Conditions (C3), (C5) and (C6) hold.*

1. *Under Condition (C1),  $\hat{\rho}^e(\hat{\delta})$  tends to  $\rho^o(s_R)$  for any  $s_R \in (c(p), p)$  with high probability. Further,*

$$\|\widehat{\Sigma}_w^e - \widehat{\Sigma}_w^o\|_2 \leq_P C p^{-\varepsilon_1} \left( \max \left\{ \sqrt{\frac{s_R \log p}{n}} \max(\sigma_{ij}), \max(\sigma_{ij}), \sqrt{\frac{p}{n}} \|\Sigma\|_2 \right\} \right).$$

241 Here  $\hat{\rho}^e(\hat{\delta})$ ,  $\rho^o(s_R)$ ,  $\widehat{\Sigma}_w^e$  and  $\widehat{\Sigma}_w^o$  are defined in (25), (21), (26), (22), respectively.

242 2. *Under Conditions (C2), the above conclusions still holds as long as  $\ell > \max\{\frac{2}{\theta_1^*}, \frac{2}{\theta_2^*}, \frac{2}{\theta_2^{*'}}\}$ .*

243 **Corollary 2.** *Under conditions of Theorem 2  $\tilde{\rho}^e$  still converges to  $\rho^o$  with high probability. Further, the bound of*  
 244  $\|\widehat{\Sigma}_w^e - \widehat{\Sigma}_w^o\|_2$  *is the same as that in Theorem 2, where  $\widehat{\Sigma}_w^e$  is given in (26).*

## 245 4. Simulation studies

246 This section is to investigate the finite sample performance of the proposed methods and compare with other  
 247 existing methods in literature. We generate data  $X = \Gamma Z$ , where  $\Sigma = \Gamma \Gamma^\top$ ,  $Z \in R^{n \times p}$  is a random matrix consisting  
 248 of i.i.d random variables either  $N(0, 1)$  or  $t_{20}$ . Similar to Bickel and Levina [2], we consider  $p = 30, 100, 200$  for  
 249  $n = 100$ . As discussed in Rothman et al. [14], there are many options in thresholding functions. In what follows, we  
 250 use the soft thresholding function, i.e.,  $s_\lambda(z) = \text{sign}(z)(|z| - \lambda)_+$ . Here, we use the universal thresholds as in Rothman  
 251 et al. [14] (Proposed 1) and adaptive thresholds as in Cai and Liu [4] (Proposed 2) to test the performances. All the  
 252 simulation studies are replicated for 100 times.

### 253 4.1. Simulation design

254 First, we consider the same covariance model  $\Sigma = U D U^\top := \Sigma_1$ , where  $D$  is a diagonal matrix with 20% 1, 40%  
 255 3 and 40% 10,  $U$  is a random orthogonal matrix, as in Ledoit and Wolf [10, 11], to verify that the new proposed  
 256 shrinkage estimator performs better than the linear and nonlinear shrinkage methods in many cases, denoted as LW1  
 257 and LW2, respectively. In addition, we also compare our method with the method proposed in Touloumis [16] (Tou),  
 258 which is also aim to find a trade off between  $S_n$  and  $D_n$ . In Table 3, we evaluate the spectral norm of the difference  
 259 between the estimated covariance matrix and the true one. From Table 3, one can see that the new proposed linear  
 260 shrinkage method is competitive, and hence can be applied in the next step, i.e., the weighted estimation.

261 In the next step, we also consider the model with strong sparsity, i.e.,  $\Sigma = (1 - \frac{|i-j|}{5})_+ := \Sigma_2$ , which is similar as Cui  
 262 et al. [3]. For comparison, we also consider some other methods proposed in Bickel and Levina [2] (BL), Ledoit and  
 263 Wolf [10] (LW1) Ledoit and Wolf [12] (LW2) and Cai and Liu [4] (CL). For each scenario, we evaluate the Spectral  
 264 norm of the difference between the estimated covariance matrix and the true one. Table 4 illustrates the performances  
 265 when the underlying covariance matrix enjoys strong sparsity ( $\Sigma_2$ ) and without any structure known in advance ( $\Sigma_1$ ).  
 266 One can see that our proposed weighted covariance estimator is able to achieve the same result as the thresholding  
 267 methods when  $\Sigma = \Sigma_2$ , and performs much better than the shrinkage methods in most cases.

Table 3: Performance of new linear shrinkage method

$p$	Proposed	LW1	LW2	Tou	Proposed	LW1	LW2	Tou
		$N(0, 1)$				$t_{20}$		
30	5.050	5.077	4.469	8.399	4.838	5.503	5.204	8.354
100	5.682	5.957	6.291	7.763	5.767	6.578	7.635	7.613
200	5.722	5.723	5.964	6.986	6.033	6.654	8.840	6.881

Table 4:  $\Sigma = \Sigma_1$  and  $\Sigma = \Sigma_2$ 

$p$	Proposed 1	Proposed 2	BL	LW1	LW2	CL
			$\Sigma_1, N(0, 1)$			
30	5.174	5.155	5.824	5.125	4.647	5.769
100	5.840	5.866	7.222	5.962	6.311	7.455
200	6.086	6.072	7.847	5.826	5.927	7.529
$\Sigma_1, t_{20}$						
30	5.030	4.953	6.151	5.495	5.290	5.813
100	5.846	5.763	8.210	6.711	7.931	7.297
200	7.006	6.809	11.598	6.993	10.833	7.350
$\Sigma_2, N(0, 1)$						
30	0.815	0.838	0.816	0.937	0.795	0.839
100	0.919	0.973	0.917	1.100	1.183	1.005
200	1.087	1.075	1.098	1.128	1.152	1.090
$\Sigma_2, t_{20}$						
30	0.785	0.787	0.796	1.021	0.939	0.787
100	0.946	0.946	0.963	1.170	1.321	0.961
200	1.056	1.017	1.108	1.246	1.854	1.060

Furthermore, in order to demonstrate the merit of the proposed weighted methods, the cases with weak sparsity are also taken into consideration. Specifically, we also consider the covariance model  $\Sigma = a\Sigma_1 + b\Sigma_2$ , where  $(a, b)$  is randomly chosen from sequence  $(1, 1.1, \dots, 5)$ . Since the results for different choice of  $(a, b)$  are similar, we only take  $(a, b) = (1.4, 2.3)$  and  $(a, b) = (4.6, 1.2)$  in simulation results, and denote the corresponding covariance matrix by  $\Sigma_3$  and  $\Sigma_4$ , respectively. Table 5 reports the performance of those methods as Table 1 and it indicates that it performs better than the remaining methods for  $t_{20}$  and is competitive to the shrinkage method proposed by Ledoit and Wolf [12] for normal distribution.

Last but not least, as in Ledoit and Wolf [10, 11], we define Percentage Relative Improvement in Average Loss (PRIAL):

$$\text{PRIAL}(\widehat{\Sigma}_n) = 100 \times \left\{ 1 - \frac{E\|\widehat{\Sigma}_n - \Sigma\|_F^2}{E\|S_n - \Sigma\|_F^2} \right\} \%,$$

where  $\widehat{\Sigma}_n$  is an arbitrary estimator of  $\Sigma$ . For simplicity, we only consider the case of  $\Sigma = \Sigma_4$  and  $t_{20}$  when  $p/n$  equal to  $1/3, 3$  and  $10$ , respectively. Figures 1 - 3 display that our proposed methods perform much better than others.

From above analysis, we see that our proposed methods is competitive in most cases, especially in the case of  $t$ -distributions. What's more, comparing with the nonlinear shrinkage methods, the proposed methods are much faster.

## Acknowledgments

Yang's research was supported by the National Nature Science Foundation of China grants 11471086 and 11871173, the National Social Science Foundation of China grants 16BTJ032, the National Statistical Scientific Center 2015LD02 and the Fundamental Research Funds for the Central Universities of Jinan University Qimingxing Plan 15JNQM019.

Table 5:  $\Sigma = \Sigma_3$  and  $\Sigma = \Sigma_4$

$p$	Proposed 1	Proposed 2	BL	LW1	LW2	CL
$\Sigma_3, N(0, 1)$						
30	8.168	8.210	9.709	8.044	7.955	9.630
100	9.515	9.585	14.197	9.045	9.332	11.927
200	9.620	9.733	21.643	9.072	9.068	12.258
$\Sigma_3, t_{20}$						
30	8.276	8.079	12.204	8.562	9.263	9.516
100	9.775	9.554	23.383	9.882	13.504	11.596
200	10.902	10.494	36.282	10.245	19.012	11.951
$\Sigma_4, N(0, 1)$						
30	23.711	24.124	31.850	23.939	22.263	27.270
100	27.305	27.349	66.225	27.448	28.955	34.791
200	26.394	26.322	103.984	26.459	27.139	35.204
$\Sigma_4, t_{20}$						
30	23.808	23.191	42.767	25.493	25.292	27.288
100	27.888	27.090	94.614	30.876	37.605	34.057
200	32.893	31.698	154.365	32.149	51.812	34.355

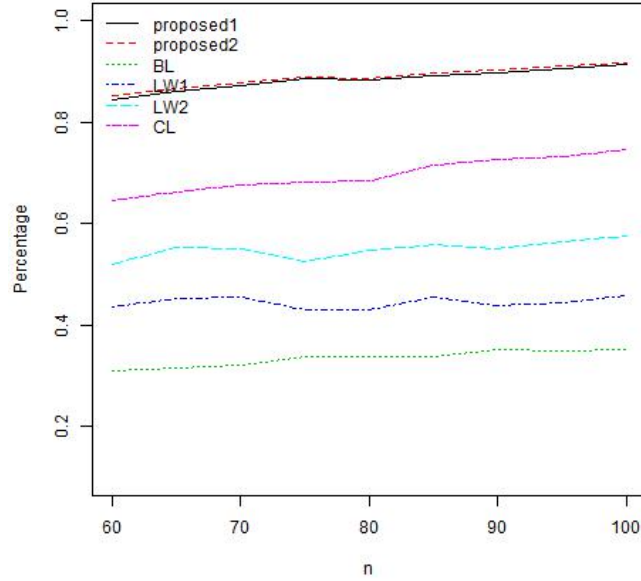


Figure 1: Prial when  $p/n = 1/3$ .

283 **Appendix**

284 For easier presentation, set  $\mathcal{T} = \{(i, j) | \sigma_{ij} \neq 0, i \neq j\}$ ,  $\mathcal{T}^c = \{(i, j) | \sigma_{ij} = 0, i \neq j\}$ , let  $\mathcal{D} = \{(i, j) | i = j\}$ ,  $|\mathcal{T}|$ ,  $|\mathcal{T}^c|$  and  
 285  $|\mathcal{D}|$  be the corresponding cardinalities. We also use  $C$  to denote the positive constants, which can change from line to  
 286 line.

287 Before introducing the proof of the main theorems, we present some useful lemmas:

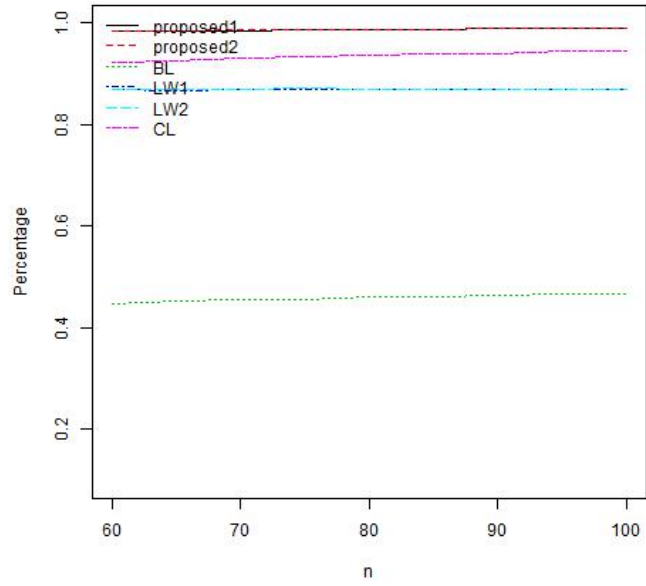


Figure 2: PRIAL when  $p/n = 3$ .

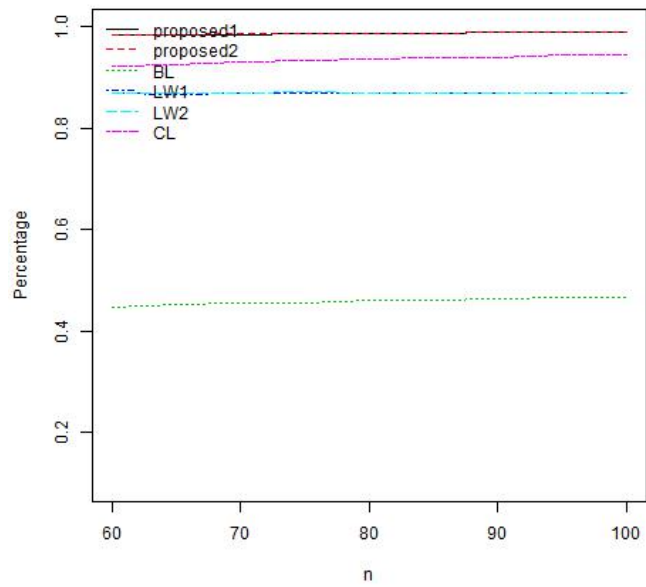


Figure 3: PRIAL when  $p/n = 10$ .

**Lemma 1.** Under Conditions (C1) or (C2), we have for some constant  $K$ ,

$$P\left(\max_{ij} |s_{ij} - \sigma_{ij}| > K \sqrt{\frac{\log p}{n}} \max_{ij}(\sigma_{ij})\right) = o(1).$$

**Lemma 2.** Under Conditions (C1) or (C2), we have for some constant  $K$ ,

$$P\left(\max_{ij} |r_{n,ij} - r_{ij}| > K \sqrt{\frac{\log p}{n}}\right) = o(1).$$

288 The proofs of Lemmas 1-2 follow from Cui et al. [3], and we omit them.

289 **Lemma 3.** Under Condition (C1),  $\|S_n - \Sigma\| \leq \max(\delta, \delta^2)\|\Sigma\|$  with probability at least  $1 - \exp(-ct^2)$ , where  $\delta =$   
290  $K\sqrt{p/n} + t/\sqrt{n}$ ,  $K$  and  $c$  are constant only depend on  $X_i$ .

291 *Proof.* See in Vershynin [17]. □

292 **Lemma 4.** Under Condition (C2),  $\|S_n - \Sigma\|_2 \leq \sqrt{\frac{p}{n}}\|\Sigma\|$  with high probability.

293 *Proof.* By  $\|S_n - \Sigma\|_2 \leq \|\Sigma\| \cdot \|\Sigma^{-1/2} S_n \Sigma^{-1/2} - I\|_2$  and the results in Yin and Li [22] and Chen and Pan [6], the conclusion  
294 follows. □

295 **Lemma 5.** Under Condition (C1) or (C2) and (C3), for  $(i, j) \in \mathcal{T}$ , we have  $|E(s_{ij}^2) - s_{ij}^2| = O(p^{-(2\xi+\varepsilon_1)})$  with probability  
296 at least  $1 - p^{-\ell(\alpha-4\xi-4\zeta-2\varepsilon_1)}$ .

297 *Proof.* Recall  $Y_{ij} = X_{ij}/\sqrt{\text{var}(X_{ij})}$ . Write

$$\begin{aligned} |(s_{12})^2 - E(s_{12})^2| &= \left| \left( \frac{1}{n} \sum_{k=1}^n X_{1k} X_{2k} \right)^2 - E \left( \frac{1}{n} \sum_{k=1}^n X_{1k} X_{2k} \right)^2 \right| \\ &\leq \frac{1}{n^2} \left| \left[ \sum_{k=1}^n (X_{1k} X_{2k} - EX_{1k} X_{2k}) \right]^2 - E \left[ \sum_{k=1}^n (X_{1k} X_{2k} - EX_{1k} X_{2k}) \right]^2 \right| \\ &\quad + 2 \frac{1}{n} \left| \sum_{k=1}^n (X_{1k} X_{2k} - EX_{1k} X_{2k}) \right| \cdot \left| \frac{1}{n} \sum_{k=1}^n EX_{1k} X_{2k} \right| \\ &\leq \frac{\sigma_{11}^2 \sigma_{22}^2}{n^2} \left| \sum_{k=1}^n Y_{1k} Y_{2k} - EY_{1k} Y_{2k} \right|^2 - E \left| \sum_{k=1}^n Y_{1k} Y_{2k} - EY_{1k} Y_{2k} \right|^2 \\ &\quad + 2 \frac{|\sqrt{\sigma_{11} \sigma_{22} \sigma_{12}}|}{n} \left| \sum_{k=1}^n Y_{1k} Y_{2k} - EY_{1k} Y_{2k} \right| \\ &\triangleq \sigma_{11}^2 \sigma_{22}^2 A + 2 |\sqrt{\sigma_{11} \sigma_{22} \sigma_{12}}| B. \end{aligned}$$

298 Let  $\alpha_k = Y_{1k} Y_{2k} - EY_{1k} Y_{2k}$ . Note that  $\alpha_k$  are independent with mean 0. For term  $A$ , by Chebyshev's inequality and  
299 Rosenthal's inequality,

$$P\left(\left| \sum_{k=1}^n \alpha_k \right|^2 - E \left| \sum_{k=1}^n \alpha_k \right|^2 > n^2 \epsilon\right) \leq C \frac{E \left| \sum_{k=1}^n \alpha_k \right|^{2\ell}}{n^{2\ell} \epsilon^\ell} \leq \frac{C \sum_{k=1}^n E |\alpha_k|^{2\ell} + C \left( \sum_{k=1}^n E \alpha_k^2 \right)^\ell}{n^{2\ell} \epsilon^\ell} = O\left(\frac{1}{n^\ell \epsilon^\ell}\right). \quad (31)$$

300 Similarly, for term  $B$ , we have

$$P\left(\left| \sum_{k=1}^n \alpha_k \right| > n\epsilon\right) \leq \frac{C \cdot E \left[ \left| \sum_{k=1}^n \alpha_k \right|^{2\ell} \right]}{n^{2\ell} \epsilon^{2\ell}} \leq \frac{C \sum_{k=1}^n E |\alpha_k|^{2\ell} + C \left( \sum_{k=1}^n E \alpha_k^2 \right)^\ell}{n^{2\ell} \epsilon^{2\ell}} = O\left(\frac{1}{n^\ell \epsilon^{2\ell}}\right). \quad (32)$$

For  $(i, j) \in \mathcal{T}$ , i.e.,  $\sigma_{12} \neq 0$ . Recall  $\max(\sigma_{ij}) = O(p^\zeta)$  in Condition (C3). We let  $\epsilon = p^{-(2\xi+4\zeta+\varepsilon_1)}$  for some small  $\varepsilon_1 > 0$  in (31). There is

$$P(A > \epsilon) = O(p^{-\ell(\alpha-2\xi-2\zeta-\varepsilon_1)}).$$

Let  $\epsilon = p^{-(2\xi+2\zeta+\varepsilon_1)}$  for some small  $\varepsilon_1 > 0$  in (32). We have

$$P(B > \epsilon) = O(p^{-\ell(\alpha-4\xi-4\zeta-2\varepsilon_1)}).$$

301 Thus, when  $(i, j) \in \mathcal{T}$ , there is  $|E(s_{ij}^2) - s_{ij}^2| = O(p^{-(2\xi+\varepsilon_1)})$ , with probability as least  $1 - p^{-\ell(\alpha-4\xi-2\varepsilon_1-4\zeta)}$ .  $\square$

302 **Lemma 6.** Under Condition (C1) or (C2) and (C3),

303 **I:** for  $(i, j) \in \mathcal{T}$ , we have  $|E(r_{n,ij}^2) - r_{n,ij}^2| = O(p^{-2\xi-\varepsilon_1})$  with probability at least  $1 - p^{-\ell(\alpha-4\xi-2\varepsilon_1-4\zeta)}$ .

304 **II:** for  $(i, j) \in \mathcal{T}^c$ , we have  $|E(r_{n,ij}^2) - r_{n,ij}^2| = O(p^{2-\alpha-2\xi-2\zeta-\gamma_1-\varepsilon_1})$  for some small  $\varepsilon_1 > 0$  with probability at least  
 305  $1 - p^{-\ell(2-4\xi-\varepsilon_1-4\zeta-\gamma_1)}$ , when  $2 - \alpha > \gamma_1$ . Similarly, for  $2 - \alpha < \gamma_1$ , there is  $|E(r_{n,ij}^2) - r_{n,ij}^2| = O(p^{\gamma_1-2\xi-2\zeta-\gamma_2-\varepsilon_1})$  for some  
 306 small  $\varepsilon_1 > 0$  with probability at least  $1 - p^{-\ell(2-4\xi-\varepsilon_1-4\zeta-\gamma_2)}$ .

307 *Proof.* (I) For  $(i, j) \in \mathcal{T}$ , write

$$|E(r_{n,ij}^2) - r_{n,ij}^2| = \left| \frac{s_{ij}^2}{s_{ii}s_{jj}} - E \frac{s_{ij}^2}{s_{ii}s_{jj}} \right| \leq \left| \frac{s_{ij}^2}{s_{ii}s_{jj}} - \frac{\sigma_{ij}^2}{\sigma_{ii}\sigma_{jj}} \right| + \left| E \left[ \frac{s_{ij}^2}{s_{ii}s_{jj}} - \frac{\sigma_{ij}^2}{\sigma_{ii}\sigma_{jj}} \right] \right| := A + B.$$

308 By Lemma 1 and Condition (C3), there is  $|\sigma_{ii}\sigma_{jj} - s_{ii}s_{jj}| \leq C(\sqrt{\log p} \cdot p^{-(\alpha/2-2\zeta)})$  with high probability. Under  
 309 Condition (C3), we can obtain  $s_{ii}s_{jj} = \Omega_P(p^{-2\xi})$ . Also, following the strategy of the proof of Lemma 5, it is easy to  
 310 obtain that  $|s_{ij}^2 - E s_{ij}^2| = O_P(-p^{(4\xi+\varepsilon_1)})$  under Condition (C1) and  $|s_{ij}^2 - E s_{ij}^2| = O_P(-p^{(4\xi+\varepsilon_1)})$  under Condition (C2) as  
 311 long as  $\ell > 2/(\alpha - 8\xi - 4\zeta - 2\varepsilon_1)$ . Moreover, by the upper bound of  $\sigma_{ij}$  in Condition (C3),  $E s_{ij}^2 - \sigma_{ij}^2 = O(p^{4\xi}/n) =$   
 312  $O(p^{-(\alpha-4\zeta)})$ . Therefore,

$$\begin{aligned} A &\leq \frac{|s_{ij}^2 - \sigma_{ij}^2|}{|s_{ii}s_{jj}|} + \frac{|\sigma_{ii}\sigma_{jj} - s_{ii}s_{jj}|\sigma_{ij}^2}{|s_{ii}s_{jj}\sigma_{ii}\sigma_{jj}|} \leq \frac{|s_{ij}^2 - \sigma_{ij}^2|}{|s_{ii}s_{jj}|} + \frac{|\sigma_{ii}||\sigma_{jj} - s_{jj}| + |s_{jj}||\sigma_{ii} - s_{ii}|}{|s_{ii}s_{jj}|} \\ &\leq C(p^{-(2\xi+\varepsilon_1)} + \sqrt{\log p} \cdot p^{-(\alpha/2-\xi-\zeta)}) \leq C(p^{-(2\xi+\varepsilon_1)}), \end{aligned}$$

313 with probability at least  $1 - o(p^{-2})$ , where the last inequality follows from Condition (C6). For term B, we define  
 314  $\mathcal{B}_i = \{|s_{ii}| > \frac{1}{2}\sigma_{ii}\}$  and  $\mathcal{B}_i^c = \{|s_{ii}| \leq \frac{1}{2}\sigma_{ii}\}$ . Write

$$B \leq \left| E \left[ \left( \frac{s_{ij}^2}{s_{ii}s_{jj}} - \frac{\sigma_{ij}^2}{\sigma_{ii}\sigma_{jj}} \right) I(\mathcal{B}_i \cap \mathcal{B}_j) \right] \right| + \left| E \left[ \left( \frac{s_{ij}^2}{s_{ii}s_{jj}} - \frac{\sigma_{ij}^2}{\sigma_{ii}\sigma_{jj}} \right) I(\mathcal{B}_i^c \cup \mathcal{B}_j^c) \right] \right| \quad (33)$$

315 The first term in (33) can be similarly handled as in the term A. While, for the second term in (33), by the fact of  
 316 boundedness of  $\left| \frac{s_{ij}^2}{s_{ii}s_{jj}} - \frac{\sigma_{ij}^2}{\sigma_{ii}\sigma_{jj}} \right|$ , we have

$$\left| E \left[ \left( \frac{s_{ij}^2}{s_{ii}s_{jj}} - \frac{\sigma_{ij}^2}{\sigma_{ii}\sigma_{jj}} \right) I(\mathcal{B}_i^c \cup \mathcal{B}_j^c) \right] \right| \leq C(P(\mathcal{B}_i) + P(\mathcal{B}_j)) \leq CP \left( |s_{ii} - \sigma_{ii}| \geq \frac{1}{2}\sigma_{ii} \right) \leq Cp^{-(2\xi+\varepsilon_1)}.$$

317 **II:** If  $2 - \alpha > \gamma_1$ , for  $(i, j) \in \mathcal{T}^c$ , by the same method, it is easy to obtain that  $|E(r_{n,ij}^2) - r_{n,ij}^2| = O(p^{2-\alpha-2\xi-2\zeta-\gamma_1-\varepsilon_1})$  with  
 318 probability at least  $1 - o(p^{-2})$  under Condition (C1). Also, the bound of  $|E(r_{n,ij}^2) - r_{n,ij}^2|$  achieves the same order under  
 319 Condition (C2) as long as  $\ell > 2/(2 - 4\xi - \varepsilon_1 - 4\zeta - \gamma_1)$ . If  $2 - \alpha < \gamma_1$ , we can also obtain similar conclusion.  $\square$

**Lemma 7.** Under conditions of Theorem 2, there is

$$E\|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2 \geq C_1 p^{2-\alpha-2\xi-2\zeta} + C_2 p^{\gamma_1-2\xi-2\zeta}.$$

320 *Proof.* Let  $\mathcal{T}_+ = \{(i, j) \in \mathcal{T} | r_{n,ij} > 0\}$  and  $\mathcal{T}_- = \{(i, j) \in \mathcal{T} | r_{n,ij} < 0\}$ . Recalling definition of  $\widehat{R}_1^\Sigma$  in (18) and  $\widehat{R}_2^\Sigma$  in (17),  
 321 we write

$$\begin{aligned}
 E\|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2 &= \sum_{(i,j) \in \mathcal{T}} E[r_{1,ij} - r_{n,ij} + (1-\tau)r_{n,ij}]^2 + \tau^2 \sum_{(i,j) \in \mathcal{T}^c} Er_{n,ij}^2 + \sum_{i=1}^p E(r_{1,ii} - 1)^2 \\
 &= \sum_{(i,j) \in \mathcal{T}_+} E[(1-\tau)r_{n,ij} - \lambda]^2 + \sum_{(i,j) \in \mathcal{T}_-} E[(1-\tau)r_{n,ij} + \lambda]^2 + \tau^2 \sum_{(i,j) \in \mathcal{T}^c} Er_{n,ij}^2 + \lambda^2 p \\
 &\geq \sum_{(i,j) \in \mathcal{T}_+} E[(1-\tau)r_{n,ij} - \lambda]^2 + \sum_{(i,j) \in \mathcal{T}_-} E[(1-\tau)r_{n,ij} + \lambda]^2 + \lambda^2 p \\
 &\geq (1-\tau)^2 \sum_{(i,j) \in \mathcal{T}} Er_{n,ij}^2 + \tau^2 \sum_{(i,j) \in \mathcal{T}^c} Er_{n,ij}^2 + \lambda^2 p + H_1,
 \end{aligned} \tag{34}$$

322 where  $H_1$  is a negligible term compared with the first two terms. Expanding  $Er_{n,ij}^2$ , there is

$$Er_{n,ij}^2 \geq E\left[\frac{s_{ij}^2}{s_{ii}s_{jj}} I(|s_{ii}s_{jj} - \sigma_{ii}\sigma_{jj}| \leq \sqrt{\log p} \cdot p^{-\alpha/2-2\xi})\right] \geq C \frac{\frac{1}{n} EX_{ik}^2 X_{jk}^2 + \frac{n^2-n}{n^2} \sigma_{ij}^2}{\sigma_{ii}\sigma_{jj}}.$$

323 Hence, we have (34)  $\geq C_1 p^{2-\alpha-2\xi-2\zeta} + C_2 p^{\gamma_1-2\xi-2\zeta}$ . □

324 **Proposition 1.** The optimization problem

$$\arg \min_{0 < \tau < 1 - \varepsilon'} E[\|(1-\tau)D_n + \tau S_n - \Sigma\|_F^2]$$

is equivalent to

$$\arg \min_{0 < \tau < 1 - \varepsilon'} E[\|\tau S_{n,\text{off}} - \Sigma_{\text{off}}\|_F^2].$$

325 *Proof.* Expanding  $E[\|(1-\tau)D_n + \tau S_n - \Sigma\|_F^2]$ , we obtain

$$E[\|(1-\tau)D_n + \tau S_n - \Sigma\|_F^2] = E[\|D_n - \text{diag}(\Sigma)\|_F^2] + E[\|\tau S_{n,\text{off}} - \Sigma_{\text{off}}\|_F^2] + E\langle D_n - \text{diag}(\Sigma), \tau S_{n,\text{off}} - \Sigma_{\text{off}} \rangle,$$

326 where the third term above is equal to 0. Hence, the conclusion follows. □

327 **Proof of Theorem 1.** For the first statement, write

$$|\hat{\tau}^e - \hat{\tau}^o| \leq \left| \frac{\sum_{(i,j) \in \mathcal{T}} s_{ij}^{(2)} s_{ij}^{(1)}}{\sum_{i \neq j} (s_{ij}^{(1)})^2} - \frac{\sum_{(i,j) \in \mathcal{T}} \sigma_{ij}^2}{\sum_{i \neq j} (s_{ij}^{(1)})^2} \right| + \left| \frac{\sum_{(i,j) \in \mathcal{T}} \sigma_{ij}^2}{\sum_{i \neq j} (s_{ij}^{(1)})^2} - \frac{\sum_{(i,j) \in \mathcal{T}} \sigma_{ij}^2}{\sum_{i \neq j} E s_{ij}^2} \right| := \text{I} + \text{II}.$$

328 We first consider I:

$$\begin{aligned}
 \text{I} &\leq \frac{\sum_{(i,j) \in \mathcal{T}} |s_{ij}^{(2)} s_{ij}^{(1)} - \sigma_{ij}^2|}{\sum_{i \neq j} (s_{ij}^{(1)})^2} \leq \frac{\sum_{(i,j) \in \mathcal{T}} \{|s_{ij}^{(2)}| |s_{ij}^{(1)} - \sigma_{ij}| + |\sigma_{ij}| |s_{ij}^{(2)} - \sigma_{ij}|\}}{\sum_{i \neq j} (s_{ij}^{(1)})^2} \\
 &\leq \frac{C \sqrt{\frac{\log p}{n}} |\mathcal{T}| \cdot \max |\sigma_{ij}|^2}{|\mathcal{T}| \cdot p^{-2\xi}} = O(\sqrt{\log p} \cdot p^{-(\alpha/2-2\xi-2\zeta)})
 \end{aligned}$$

329 where the third inequality is obtained by Lemma 1 and Condition (C3).

330 Next consider II. Recalling  $|\mathcal{T}| \asymp p^{\gamma_1}$ ,  $|\mathcal{T}^c| \asymp p^{\gamma_2}$  and  $|\mathcal{T}| + |\mathcal{T}^c| = p^2 - p$  in Condition (C4), we consider following  
 331 three cases:

- 332 1.  $\gamma_1 < 2 - \alpha - 2\xi - 2\zeta$  and  $\gamma_2 = 2$ .
- 333 2.  $2 > \gamma_1 > 2 - \alpha + 2\xi + 2\zeta$  and  $\gamma_2 = 2$ .
- 334 3.  $\gamma_1 = 2$  and  $\gamma_2 \in [0, 2]$ .



335 For Cases 1, 2 and 3, we rewrite

$$\Pi = \left( \sum_{(i,j) \in \mathcal{T}} \sigma_{ij}^2 \right) \cdot \left| \frac{\sum_{i \neq j} |Es_{ij}^2 - (s_{ij}^{(2)})^2|}{[\sum_{i \neq j} Es_{ij}^2][\sum_{i \neq j} (s_{ij}^{(1)})^2]} \right| = \left( \frac{\sum_{(i,j) \in \mathcal{T}} \sigma_{ij}^2}{\sum_{i \neq j} Es_{ij}^2} \right) \cdot \left| \frac{\sum_{i \neq j} |Es_{ij}^2 - (s_{ij}^{(2)})^2|}{\sum_{i \neq j} (s_{ij}^{(1)})^2} \right|. \quad (35)$$

336 Since  $\text{var}(X_{ik}^2 X_{jk}^2) > c > 0$ , it is easy to obtain  $n^{-1} \sum_{k=1}^n X_{ik}^2 X_{jk}^2 = EX_{ik}^2 X_{jk}^2 + o_p(EX_{ik}^2 X_{jk}^2)$ . Also, for  $(i, j) \in \mathcal{T}$ , there is  
 337  $n^{-2} \sum_{k_1 \neq k_2} X_{ik_1} X_{jk_1} X_{ik_2} X_{jk_2} = \sigma_{ij}^2 + o_p(\sigma_{ij}^2)$ . Therefore, write

$$\left| \frac{\sum_{i \neq j} |Es_{ij}^2 - (s_{ij}^{(2)})^2|}{\sum_{i \neq j} (s_{ij}^{(1)})^2} \right| \leq C \left( \left| \frac{\sum_{(i,j) \in \mathcal{T}} |E(s_{ij}^{(2)})^2 - (s_{ij}^{(2)})^2|}{\frac{1}{n} \sum_{i \neq j} EX_{ik}^2 X_{jk}^2 + \frac{n^2-n}{n^2} \sum_{(i,j) \in \mathcal{T}} \sigma_{ij}^2} \right| + \left| \frac{\sum_{(i,j) \in \mathcal{T}^c} |E(s_{ij}^{(2)})^2 - (s_{ij}^{(2)})^2|}{\frac{1}{n} \sum_{i \neq j} EX_{ik}^2 X_{jk}^2 + \frac{n^2-n}{n^2} \sum_{(i,j) \in \mathcal{T}} \sigma_{ij}^2} \right| \right) \quad (36)$$

338 with high probability. The first term in the right hand side of (36) is bounded by  $O_p(p^{-\varepsilon_1})$  for all these three cases by  
 339 Lemma 5. Now, consider the second term of (36). Under Case 1, according to  $\sum_{i \neq j} Es_{ij}^2 = n^{-1} \sum_{i \neq j} EX_{ik}^2 X_{jk}^2 + (n^2 -$   
 340  $n)n^{-2} \sum_{(i,j) \in \mathcal{T}} \sigma_{ij}^2$  and Condition (C3), there is  $\sum_{i \neq j} Es_{ij}^2 = \Omega(p^{2-\alpha-2\xi})$ , and hence

$$\lim_{n \rightarrow \infty} \left( \frac{\sum_{(i,j) \in \mathcal{T}} \sigma_{ij}^2}{\sum_{i \neq j} Es_{ij}^2} \right) = O(p^{\gamma_1 - 2 - \alpha - 2\xi + 2\zeta}).$$

341 By setting  $\epsilon = p^{2-\gamma_1-4\zeta-\varepsilon_1}$  in (31), we can obtain that  $\Pi = O_p(p^{-\varepsilon_1})$ .

342 Under Cases 2 and 3, there is

$$\lim_{n \rightarrow \infty} \left( \frac{\sum_{(i,j) \in \mathcal{T}} \sigma_{ij}^2}{\sum_{i \neq j} Es_{ij}^2} \right) = 1.$$

343 It is easy to obtain that  $(n^2 - n)(n^{-2}) \sum_{(i,j) \in \mathcal{T}} \sigma_{ij}^2 \gg n^{-1} \sum_{i \neq j} EX_{ik}^2 X_{jk}^2$  under Cases 2 and 3. Thus, taking  $\epsilon =$   
 344  $p^{\gamma_1 - \gamma_2 - 2\zeta - 2\xi - \varepsilon_1}$  in (31), we can also have  $\Pi = O_p(p^{-\varepsilon_1})$ .

345 Thus, we conclude that  $|\hat{\tau}^e - \tau^o| = O(p^{-\varepsilon_1})$ , with high probability.

346 Further, by Lemma 3, the definitions in (4) and (7),

$$\|\widehat{\Sigma}_{sh}^e - \Sigma_{sh}^o\|_2 \leq p^{-\varepsilon_1} \|S_{n,\text{off}}\|_2 \leq Cp^{-\varepsilon_1} \max\left(\sqrt{\frac{p}{n}}, 1\right) \|\Sigma\|_2.$$

347 Thus, the first statement follows.

348 For the second statement, most proof is similar. The only difference is that it requires  $|E(s_{ij}^2) - s_{ij}^2| = O(p^{-(2\xi+\varepsilon_1)})$   
 349 with probability as least  $1 - o(p^{-2})$  in Lemma 5. Thus, recalling  $\theta_1$  and  $\theta_2$  in Condition (C4), we let  $\ell$  in Condition  
 350 (C2) larger than  $\max\{2/\theta_1, 2/\theta_2, 2/\theta'_2\}$ . Combining with Lemma 4, the conclusion follows.  $\square$

351 **Proof of Corollary 1.** From (8), we see  $|\hat{\tau}^e - \tau^o| = |M^{-1} \sum_{m=1}^M (\hat{\tau}_m^e - \tau^o)| \leq M^{-1} \sum_{m=1}^M |(\hat{\tau}_m^e - \tau^o)|$ . Thus, following the  
 352 proof of Theorem 1, we can draw the conclusion.  $\square$

353 **Proof of Theorem 2.** Note that there is no difference between  $\widehat{R}_1^{\Sigma^{(1)}}$  and  $\widehat{R}_1^{\Sigma}$  (or  $\widehat{R}_2^{\Sigma^{(1)}}$  and  $\widehat{R}_2^{\Sigma}$ ) as long as  $R_n^{(2)}$  is  
 354 independent of  $\widehat{R}_1^{\Sigma^{(1)}}$  and  $\widehat{R}_2^{\Sigma^{(1)}}$ , and these two estimators constructed from the same samples. For simplicity, we below  
 355 suppress superscripts from  $\widehat{R}_1^{\Sigma^{(1)}}$  and  $\widehat{R}_2^{\Sigma^{(1)}}$ .

Recall (20) and (24). Rewrite

$$\widehat{L}^e = \frac{\langle R_n^{(2)} - \widehat{R}_2^{\Sigma}, \widehat{R}_1^{\Sigma} - \widehat{R}_2^{\Sigma} \rangle}{\|\widehat{R}_1^{\Sigma} - \widehat{R}_2^{\Sigma}\|_F^2}.$$

356 Note that here  $R_n^{(2)}$  and  $\widehat{R}_1^\Sigma$  or  $R_n^{(2)}$  and  $\widehat{R}_2^\Sigma$  come from different samples, and hence they are independent each other. We  
 357 first consider the difference between  $L^o(s_R)$  and  $\widehat{L}^e$ . Specifically,

$$|L^o(s_R) - \widehat{L}^e| = \left| \frac{E\langle R^\Sigma - \widehat{R}_2^\Sigma, \widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma \rangle \|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2 - \langle R_n^{(2)} - \widehat{R}_2^\Sigma, \widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma \rangle E\|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2}{\|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2 E\|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2} \right|$$

$$\leq \frac{|\langle R_n^{(2)} - \widehat{R}_2^\Sigma, \widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma \rangle - E\langle R^\Sigma - \widehat{R}_2^\Sigma, \widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma \rangle|}{E\|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2} \quad (37)$$

$$+ \frac{|\langle R_n^{(2)} - \widehat{R}_2^\Sigma, \widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma \rangle| \cdot \left| \|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2 - E\|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2 \right|}{\|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2 E\|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2}. \quad (38)$$

358 For the nominator of (37), recalling the definition of  $\mathcal{T}_+$  and  $\mathcal{T}_+$  in Lemma 7,

$$\left| \langle R_n^{(2)} - \widehat{R}_2^\Sigma, \widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma \rangle - E\langle R^\Sigma - \widehat{R}_2^\Sigma, \widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma \rangle \right|$$

$$= \left| \sum_{i \neq j} \left[ (r_{n,ij}^{(2)} - \tau r_{2,ij})(r_{1,ij} - r_{2,ij}) - E(r_{ij} - \tau r_{n,ij})(r_{1,ij} - r_{2,ij}) \right] \right|$$

$$\leq \left| \sum_{(i,j) \in \mathcal{T}_+} (r_{n,ij}^{(2)} - \tau r_{n,ij})(r_{n,ij} - \lambda - \tau r_{n,ij}) - E(r_{ij} - \tau r_{n,ij})(r_{n,ij} - \lambda - \tau r_{n,ij}) \right| \quad (39)$$

$$+ \left| \sum_{(i,j) \in \mathcal{T}_-} (r_{n,ij}^{(2)} - \tau r_{n,ij})(r_{n,ij} + \lambda - \tau r_{n,ij}) - E(r_{ij} - \tau r_{n,ij})(r_{n,ij} + \lambda - \tau r_{n,ij}) \right| \quad (40)$$

$$+ \left| \sum_{(i,j) \in \mathcal{T}^c} \tau (r_{n,ij}^{(2)} - \tau r_{n,ij}) r_{n,ij} + \tau^2 E r_{n,ij}^2 \right|. \quad (41)$$

By the fact of  $|r_{n,ij}^{(2)} r_{n,ij} - r_{ij}^2| \leq |r_{n,ij}^{(2)} - r_{ij}| |r_{n,ij}| + |r_{n,ij} - r_{ij}| |r_{ij}|$ , Lemma 2 and the first statement of Lemma 6, it is easy to obtain that

$$(39) \leq C |\mathcal{T}_+| p^{-(2\xi + \varepsilon_1)},$$

with high probability. By similar arguments, we also have

$$(40) \leq C |\mathcal{T}_-| p^{-(2\xi + \varepsilon_1)}.$$

with high probability. It follows from the second statement of Lemma 6 that

$$(41) \leq C |\mathcal{T}^c| p^{-(\alpha + 2\xi + 2\zeta + \gamma_2 + \varepsilon_1 - 2)} \quad \text{as } 2 - \alpha > \gamma_1,$$

or

$$(41) \leq C |\mathcal{T}^c| p^{-(\gamma_1 - \gamma_2 - 2\xi - 2\zeta - \varepsilon_1)} \quad \text{as } 2 - \alpha < \gamma_1,$$

359 with high probability. Via Lemma 7, there is

$$(37) = O_P(p^{-\varepsilon_1}). \quad (42)$$

360 Now, we consider (38). Similarly, we can obtain that

$$\frac{\left| \|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2 - E\|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2 \right|}{E\|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2} = O_P(p^{-\varepsilon_1}). \quad (43)$$

361 Thus, if there is

$$\frac{|\langle R_n^{(2)} - \widehat{R}_2^\Sigma, \widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma \rangle|}{\|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2} = O_P(1), \quad (44)$$

362 the conclusion follows.

363 Consider (44) now and write

$$\begin{aligned}
\left| E\langle R^\Sigma - \widehat{R}_2^\Sigma, \widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma \rangle \right| &= \left| E \sum_{i \neq j} (r_{ij} - \tau r_{n,ij})(r_{n,ij} - \lambda - \tau r_{n,ij}) \right| \\
&= \left| \sum_{(i,j) \in \mathcal{T}} [(1-\tau)r_{ij}^2 - \tau(1-\tau)Er_{n,ij}^2 + H_1] - \sum_{(i,j) \in \mathcal{T}^c} \tau(1-\tau)Er_{n,ij}^2 \right| \\
&= \left| \sum_{(i,j) \in \mathcal{T}} [(1-\tau)^2 Er_{n,ij}^2 + H_1 + H_2] - \sum_{(i,j) \in \mathcal{T}^c} \tau(1-\tau)Er_{n,ij}^2 \right| \\
&\leq \left| \sum_{(i,j) \in \mathcal{T}} [(1-\tau)^2 Er_{n,ij}^2 + H_1 + H_2] \right| + \left| \sum_{(i,j) \in \mathcal{T}^c} \tau(1-\tau)Er_{n,ij}^2 \right| := H_2 + H_3,
\end{aligned}$$

where  $H_2$  and  $H_3$  are of the lower order terms than  $\sum_{(i,j) \in \mathcal{T}} [Er_{n,ij}^2]$ . For  $(i, j) \in \mathcal{T}^c$ , it is easy to obtain  $Er_{n,ij}^2 = O(p^{-(\alpha-2\xi-4\xi)})$ . Note that  $R_n^{(2)}$  is independent of  $\widehat{R}_1^\Sigma$  and  $\widehat{R}_2^\Sigma$ . Thus, combining with the first statement in Lemma 7 and Condition (C6), we have

$$\frac{\left| E\langle R_n^{(2)} - \widehat{R}_2^\Sigma, \widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma \rangle \right|}{E\|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2} = O(1).$$

364 This, together with (42) and (43), implies (44). By the argument in (37) and the bound of  $\left| \|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2 - E\|\widehat{R}_1^\Sigma - \widehat{R}_2^\Sigma\|_F^2 \right|$ ,  
365 we can draw the first conclusion. Since we have obtained  $|\widehat{L}^e - L^o(s_R)| = O_P(p^{-\varepsilon_1})$ , through the the results in Rothman  
366 et al. [14], there is  $|\rho^e(\hat{\delta}) - \rho^o(s_R)| = O_P(p^{-\varepsilon_1})$ . Recalling the definition in (26) and (22), there is

$$\begin{aligned}
\|\widehat{\Sigma}_w^e - \widehat{\Sigma}_w^o\|_2 &\leq Cp^{-\varepsilon_1} \left( \|D_n^{\frac{1}{2}}(\widehat{R}_1^\Sigma - R^\Sigma)D_n^{\frac{1}{2}}\|_2 + \|D_n^{\frac{1}{2}}R^\Sigma D_n^{\frac{1}{2}}\|_2 \right) + p^{-\varepsilon_1} \left( \|\widehat{\Sigma}_{sh}^e - \widehat{\Sigma}_{sh}^o\|_2 \right) + p^{-\varepsilon_1} \|\widehat{\Sigma}_{sh}^o\|_2 \\
&\leq Cp^{-\varepsilon_1} \left( \max\left\{ \sqrt{\frac{s_R \log p}{n}}, 1 \right\} \max(\sigma_{ij}) \right) + p^{-2\varepsilon_1} \max\left( \sqrt{\frac{p}{n}}, 1 \right) \|\Sigma\|_2 + p^{-\varepsilon_1} \left( \sqrt{\frac{p}{n}} \|\Sigma\|_2 + \max(\sigma_{ij}) \right) \\
&\leq Cp^{-\varepsilon_1} \left( \max\left\{ \sqrt{\frac{s_R \log p}{n}} \max(\sigma_{ij}), \max(\sigma_{ij}), \sqrt{\frac{p}{n}} \|\Sigma\|_2 \right\} \right),
\end{aligned}$$

367 where the second inequality follows from Cui et al. [3], Theorem 1 and Lemma 3 or Lemma 4.

368 The proof of the second statement is similar. The only difference is that it requires  $|E(r_{n,ij}^2) - r_{n,ij}^2| = O(p^{-(2\xi+\varepsilon_1)})$   
369 with probability as least  $1 - o(p^{-2})$  in Lemma 6. Thus, recalling  $\theta_1^*$  and  $\theta_2^*$  in Condition (C6), we let  $\ell$  in Condition  
370 (C2) larger than  $\max\{2/\theta_1^*, 2/\theta_2^*, 2/\theta_2^*\}$ . This and Lemma 4 lead to the conclusion.  $\square$

371 **Proof of Corollary 2.** The proof of Corollary 2 is similar to the Corollary 1, hence omitted.  $\square$

## 372 References

- 373 [1] Bickel, P. & Levina, E., Regularized estimation of large covariance matrices, *Ann. Statist.* 36(2008a) 199–227.  
374 [2] Bickel, P. & Levina, E., Covariance regularization by thresholding, *Ann. Statist.* 36 (2008b)2577–2604.  
375 [3] Cui, Y., Leng, C. & Sun, D., Sparse estimation of high-dimensional correlation matrices, *Comp. Statist. Data Anal.* 93(2016) 390–403.  
376 [4] Cai, T. & Liu, W., Adaptive thresholding for sparse covariance matrix estimation, *J. Am. Statist. Assoc.*, 106(2011) 1–13.  
377 [5] Cai, T., Liu, W. & Luo, X., A Constrained  $l_1$  Minimization Approach to Sparse Precision Matrix Estimation, *J. Am. Statist. Assoc.*, 106(2011)  
378 594–607.  
379 [6] Chen, B. and Pan, G., Convergence of the largest eigenvalue of normalized sample covariance matrices when  $p$  and  $n$  both tend to infinity  
380 with their ratio converging to zero, *Bernoulli*, 18(2012) 1405–1420.  
381 [7] El Karoui, N., Operator norm consistent estimation of large dimensional sparse covariance matrices, *Ann. Statist.*, 36(2008) 2717–2756.  
382 [8] Fan, J., Zhang, J. & Yu, K., Vast portfolio selection with gross exposure constraints, *J. Am. Statist. Assoc.*, 107(2012) 592–606.  
383 [9] Fisher, T.J. & Sun, X., Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix, *Comp.*  
384 *Statist. Data Anal.*, 55(2011) 1909–1918.

- 385 [10] Ledoit, O. & Wolf, M., A well-conditioned estimator for large-dimensional covariance matrices, *J. Multiv. Anal.*, 88(2004) 365–411.
- 386 [11] Ledoit, O. and Wolf, M., Nonlinear shrinkage estimation of large-dimensional covariance matrices, *Ann. Statist.*, 40(2012) 1024–1060.
- 387 [12] Ledoit, O. and Wolf, M., Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions, *J. Multiv.*  
388 *Anal.*, 139(2015)360–384.
- 389 [13] Ledoit, O. and Wolf, M., Optimal estimation of a large-dimensional covariance matrix under Stein’s loss, *Bernoulli*, 24(2018) 3791–3832.
- 390 [14] Rothman, A., Levina, L., & Zhu, J., Generalized thresholding of large covariance matrices, *J. Am. Statist. Assoc.* 104(2009) 177–186.
- 391 [15] Stein, C., Lectures on the theory of estimation of many parameters, *J. Math. Sci.*, 34(1986) 1373–1403.
- 392 [16] Touloumis, A., Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional setting, *Comp. Statist. Data Analysis*,  
393 83(2015) 251–261.
- 394 [17] Vershynin, R., Introduction to the non-asymptotic analysis of random matrices, *Compressed Sensing: Theory and Applications* (2011).
- 395 [18] Vershynin, R., How close is the sample covariance matrix to the actual covariance matrix?, *Journal of Theoretical Probability*, 25(2012)655–  
396 686.
- 397 [19] Wang, C., Pan, G., Tong, T., & Zhu, L., Shrinkage estimation of large dimensional precision matrix using random matrix theory, *Statistica*  
398 *Sinica*, 25(2015)993–1008.
- 399 [20] Wu, W. & Pourahmadi, M., Banding sample autocovariance matrices of stationary processes, *Statistica Sinica*, 19(2009) 1755–1768.
- 400 [21] Xue, L., Ma, S. & Zou, H., Positive definite  $L_1$  penalized estimation of large covariance matrices, *J. Am. Statist. Assoc.* 107(2012) 1480–149.
- 401 [22] Yin, J. & Li, H., Model selection and estimation in the matrix normal graphical model, *J. Multiv. Anal.*,107(2012) 119–140.
- 402 [23] Zhou, H., Li, L. & Zhu, H. ,Tensor regression with applications in neuroimaging data analysis, *J. Am. Statist. Assoc.*, 108(2013) 540–552.