Journal of Econometrics 215 (2020) 118–130

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Ultrahigh dimensional precision matrix estimation via refitted cross validation☆

Luheng Wang [a], Zhao Chen [b,*], Christina Dan Wang [c], Runze Li [d]

[a] School of Statistics, Beijing Normal University, Haidian, Beijing 100875, PR China
[b] School of Data Science, Fudan University, Yangpu District, Shanghai 200433, PR China
[c] New York University Shanghai, Pudong New District, Shanghai 200122, PR China
[d] Department of Statistics and The Methodology Center, The Pennsylvania State University, PA 16802-2111, USA

ARTICLE INFO

Article history:
Received 18 November 2018
Received in revised form 30 July 2019
Accepted 12 August 2019
Available online 25 September 2019

JEL classification:
C13 and C51

Keywords:
Covariance matrix estimation
Precision matrix
Refitted cross validation
Sample splitting
Spurious correlation

ABSTRACT

This paper develops a new estimation procedure for ultrahigh dimensional sparse precision matrix, the inverse of covariance matrix. Regularization methods have been proposed for sparse precision matrix estimation, but they may not perform well with ultrahigh dimensional data due to the spurious correlation. We propose a refitted cross validation (RCV) method for sparse precision matrix estimation based on its Cholesky decomposition, which does not require the Gaussian assumption. The proposed RCV procedure can be easily implemented with existing software for ultrahigh dimensional linear regression. We establish the consistency of the proposed RCV estimation and show that the rate of convergence of the RCV estimation without assuming banded structure is the same as that of those assuming the banded structure in Bickel and Levina (2008b). Monte Carlo studies were conducted to access the finite sample performance of the RCV estimation. Our numerical comparison shows that the RCV estimation outperforms the existing ones in various scenarios. We further apply the RCV estimation for an empirical analysis of asset allocation.

1. Introduction

Precision matrix, the inverse of covariance matrix, plays an important role in statistical inference and statistical learning such as one sample mean test, graphical modeling and linear discrimination analysis. This study is motivated by an empirical analysis of portfolio allocation, in which the estimation of precision matrix is required (see Section 3.2 for more details). In the classic multivariate statistical analysis, the inverse of sample covariance matrix is a natural and consistent estimator of precision matrix. Sample covariance matrix, however, becomes singular for high dimensional data when the sample size n is less than the dimension p of data. Thus, it is of interest to develop alternative methods to estimate precision matrix. Bickel and Levina (2008a), Rothman et al. (2009) and Cai and Liu (2011) approached the estimation problem by applying different thresholding methods to the sample covariance matrix. These methods may not perform well for ultrahigh dimensional data. Yuan and Lin (2007) considered a LASSO-type method for precision matrix and solved the optimization problem by a maxdet algorithm. Due to the important role of precision matrix in Gaussian graphic model, Ren et al. (2015), Fan and Lv (2016) and Ren et al. (2019) studied the estimation and statistical

☆ All authors equally contribute to this work, and the authors are listed in seniority order.

* Corresponding author.
E-mail addresses: wangluheng@mail.bnu.edu.cn (L. Wang), zchen_fdu@fudan.edu.cn (Z. Chen), christina.wang@nyu.edu (C.D. Wang), rzli@psu.edu (R. Li).

https://doi.org/10.1016/j.jeconom.2019.08.004
0304-4076/

inference of sparse precision matrix by assuming conditional normal distribution. However, it is challenging in verifying the normality of high dimensional data in practice. People recast the problem of precision matrix estimation into a high dimensional regularization regression. A constrained ℓ_1 minimization procedure was proposed by Cai et al. (2011) for sparse precision matrix estimation. Rothman et al. (2008) and Lam and Fan (2009) used the penalized likelihood method to estimate sparse precision matrix with ℓ_1 and nonconvex penalty functions, respectively. These methods are easy to implement, and their statistical properties have been studied. However, when $p \gg n$, these methods may perform poorly due to severe spurious correlations. A comprehensive investigation of the impact of spurious collections on error variance estimation in ultrahigh dimensional data can be found in Fan et al. (2012). This paper aims to develop a consistent and easy-to-compute estimator of precision matrix for ultrahigh dimensional data without the Gaussian assumption.

In this paper, we propose an estimation procedure for the ultrahigh dimensional precision matrix by using refitted cross-validation (RCV, Fan et al., 2012). We first parameterize the precision matrix using the modified Cholesky decomposition. The decomposition is valid for any positive definite matrices and does not require the Gaussian assumption. It has been used in some literature for the estimation of covariance or precision matrix. Specifically, the modified Cholesky decomposition has been adopted for estimating the finite-dimensional error covariance matrix in longitudinal regression model (Pourahmadi, 1999, 2000), for the low-dimensional covariance estimation in Huang et al. (2006), and for the high-dimensional precision matrix estimation in Rothman et al. (2010). It is challenging to employ the Cholesky decomposition for ultrahigh dimensional covariance matrix estimation since it leads to ultrahigh dimensional linear regression problems. And the theoretical analyses of these problems are much more difficult than that in finite- or high-dimensional settings. To overcome the difficulty, we proposed the RCV estimation. The RCV estimation procedure for the ultrahigh dimensional covariance matrix can be carried out easily with the existing software for ultrahigh dimensional linear regressions. The resulting estimator of precision matrix is symmetric and positive definite. We show that the estimator is consistent. Moreover, we provide the convergence rate of the newly proposed estimator without assuming banded structure. The convergence rate is the same as the rate of those methods assuming the banded structure in Bickel and Levina (2008b).

The rest of this paper is organized as follows. In Section 2, we propose the RCV estimation procedure for the precision matrix and study its theoretical properties. In Section 3, we present a numerical comparison and an empirical study. Technical conditions and proofs are given in Appendix.

2. An RCV estimator for precision matrix

Let $\mathbf{x} = (X_1, \ldots, X_p)^T$ be a p-dimensional random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Of interest is to estimate the precision matrix $\Omega = \Sigma^{-1}$. Without loss of generality, assume $\boldsymbol{\mu} = \mathbf{0}$ throughout this paper. As in Pourahmadi (1999), the modified Cholesky's decomposition of Ω is defined as follows.

$$\Omega = \mathbf{L}^T \mathbf{D}^{-1} \mathbf{L}, \tag{2.1}$$

where $\mathbf{L}$ is unitriangular matrix (i.e. a lower triangular matrix satisfying that each diagonal element $l_{tt} = 1,\ t = 1, \ldots, p$) and $\mathbf{D} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ is a diagonal matrix. Regardless of the population distribution, the matrix decomposition is always true for the precision matrix. Since $\Omega = \Sigma^{-1}$, (2.1) is equivalent to

$$\mathbf{L}\Sigma\mathbf{L}^T = \mathbf{D},$$

which provides us a nice interpretation of the elements in $\mathbf{L}$ and $\mathbf{D}$. Define

$$\boldsymbol{\varepsilon} = \mathbf{L}\mathbf{x}. \tag{2.2}$$

Let $-\beta_{tk} = l_{tk}$, the (t, k)-th entry of $\mathbf{L}$ for $t < k$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^T$. Then (2.2) is equivalent to that $X_1 = \varepsilon_1$, and for $1 < t \le p$,

$$X_t = \sum_{k=1}^{t-1} \beta_{tk} X_k + \varepsilon_t, \tag{2.3}$$

with $E(\varepsilon_t) = 0$ and $\mathrm{var}(\varepsilon_t) = \sigma_t^2$. Thus, the purpose of estimating Σ and Ω can be achieved by estimating the regression coefficients and error variances in (2.2). Based on this idea, Pourahmadi (1999, 2000) proposed parametrization of error covariance matrix in the analysis of longitudinal data. In the presence of sparsity on β_{tk}, Huang et al. (2006) developed an estimation procedure for covariance matrix Σ using penalized least squares method in the fixed and finite dimensional setting. Rothman et al. (2010) assumed that $\mathbf{L}$ is banded and then applied LASSO-type estimation to estimate the regression coefficients $\boldsymbol{\beta}$'s under the high dimensional setting.

2.1. A naive estimator

To illustrate the challenge of estimating ultrahigh dimensional precision matrix, let us start with a natural extension of the estimation procedure proposed by Huang et al. (2006). Since we will deal with ultrahigh dimensional regressions in (2.3), we impose sparsity assumption on $\boldsymbol{\beta}_t = (\beta_{t1}, \ldots, \beta_{t,t-1})^T$, and apply penalized least squares (PLS) method to

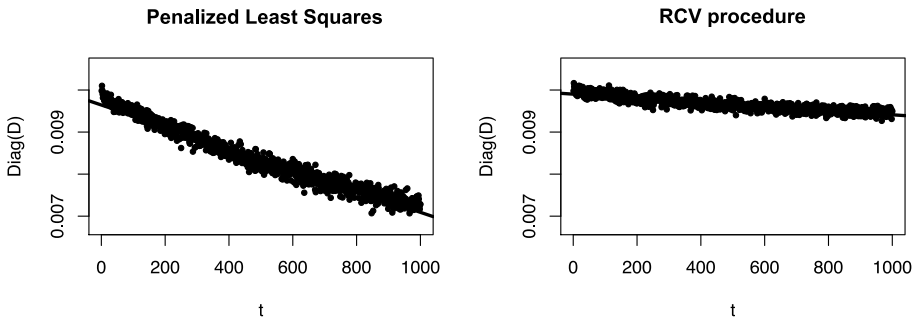

Fig. 1. The means of estimators of diag(**D**) using the naive estimate and the RCV estimate.

estimate $\boldsymbol{\beta}_t$. After estimating $\boldsymbol{\beta}_t$, we use the mean squared errors (MSE) of the corresponding linear models to estimate σ_t^2, the tth diagonal element of **D**.

Suppose that $\mathbf{x}_i = (X_{i1}, \ldots, X_{ip})^T$, $i = 1, \ldots, n$ are the independent and identically distributed samples from the population of **x**. The PLS estimator of $\boldsymbol{\beta}_t$ is

$$\tilde{\boldsymbol{\beta}}_t^{\mathrm{LS}} = \underset{\boldsymbol{\beta}_t}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} \left(X_{it} - \sum_{k=1}^{t-1} \beta_{tk} X_{ik} \right)^2 + \sum_{k=1}^{t-1} p_{\lambda_t}(|\beta_{tk}|), \tag{2.4}$$

where $p_{\lambda_t}(\cdot)$ is a penalty function with tuning parameter λ_t. Various penalty functions can be applied, such as ℓ_1-penalty (Tibshirani, 1996), SCAD penalty (Fan and Li, 2001) and MCP penalty (Zhang, 2010). With the estimate of $\boldsymbol{\beta}_t$, we can estimate σ_t^2 by the MSE

$$\tilde{\sigma}_t^2 = \frac{1}{n - \|\tilde{\boldsymbol{\beta}}_t\|_0} \sum_{i=1}^{n} \left(X_{it} - \sum_{k=1}^{t-1} \tilde{\beta}_{tk}^{LS} X_{ik} \right)^2. \tag{2.5}$$

where $\|\cdot\|_0$ denotes the ℓ_0-norm (i.e., the number of nonzero elements in the vector), and $\tilde{\beta}_{tk}^{LS}$ is the least squares estimator under the model selected by the PLS method.

As the number of parameters increases, the computation consumption increases, while algorithm stability for Eq. (2.4) may decrease, especially for the ultrahigh dimensional cases ($p = O(e^{\gamma n}), 0 \le \gamma < 1$). When $\log(t) = O(n^{\alpha})$ for some constant $\alpha \in (0, 1)$, the penalized least squares (2.4) is difficult to carry out because of computational cost and algorithm stability. To tackle this problem, Fan and Lv (2008) proposed the sure independence screening (SIS) procedure to quickly reduce the ultrahigh dimension to a moderate dimension d such as $d = [n/\log n]$. Consequently, we integrate SIS and PLS procedures estimate the precision matrix, and obtain a *naive estimator* of Ω as follows.

$$\tilde{\Omega} = \tilde{\mathbf{L}}^T \tilde{\mathbf{D}}^{-1} \tilde{\mathbf{L}}, \tag{2.6}$$

where $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{L}}$ are obtained from $\tilde{\sigma}_t^2$, $t = 1, \ldots, p$ and $\tilde{\boldsymbol{\beta}}_t$, $t = 1, \ldots, p$, respectively.

The naive estimator $\tilde{\Omega}$ may perform well in low dimensional case. However, it may not work well under ultrahigh dimensional setting due to the bias of $\tilde{\sigma}_t^2$. Let $\mathcal{M}_t$ be the model selected by the PLS (2.4), $\mathbf{X}_{\mathcal{M}_t}$ be the design matrix for the selected model, and $\mathbf{P}_{\mathcal{M}_t} = \mathbf{X}_{\mathcal{M}_t}(\mathbf{X}_{\mathcal{M}_t}^T \mathbf{X}_{\mathcal{M}_t})^{-1} \mathbf{X}_{\mathcal{M}_t}^T$ be the projection matrix of the linear space spanned by the columns of $\mathbf{X}_{\mathcal{M}_t}$. Define $\zeta_{nt}^2 = \boldsymbol{\varepsilon}_t^T \mathbf{P}_{\mathcal{M}_t} \boldsymbol{\varepsilon}_t / \boldsymbol{\varepsilon}_t^T \boldsymbol{\varepsilon}_t$, where $\boldsymbol{\varepsilon}_t = (\varepsilon_{t1}, \ldots, \varepsilon_{tn})^T$. Under certain regularity conditions, Fan et al. (2012) showed that when $\log t$ and n have the same order, $\tilde{\sigma}_t^2/(1 - \zeta_{nt}^2) \to \sigma_t^2$ in probability and

$$\sqrt{n}\{\tilde{\sigma}_{nt}^2/(1 - \zeta_{nt}^2) - \sigma_t^2\} \to N(0, \mathrm{var}(\varepsilon_t^2)).$$

Thus, $\tilde{\sigma}_{nt}^2$ shrinks σ^2 by the factor $(1 - \zeta_{nt}^2)$. Fan et al. (2012) demonstrated that ζ_{nt}^2 may not tend to zero under ultrahigh dimensional setting. This confirms that $\tilde{\sigma}_t^2$ can be an inconsistent estimator of σ_t^2 for large t. As a result, $\tilde{\mathbf{D}}$ may not be a consistent estimator of **D**, and $\tilde{\Omega}$ may not perform well.

Example 2.1. Before we pursue further, let us illustrate the bias of $\tilde{\mathbf{D}}$ by running a small simulation study. Take $\mathbf{D} = 0.01\mathbf{I}$ and $l_{t,k} = 0.6$ if $k = t - 1$, and 0 otherwise. The corresponding covariance matrix Σ satisfies the AR(1) structure. Set $n = 160$ and $p = 1000$. We run 200 simulations. For each simulation, we obtain $\tilde{\boldsymbol{\beta}}_t$, the PLS with SCAD penalty while the tuning parameter is selected by cross-validation, and then obtain $\tilde{\sigma}_t^2$. The left panel of Fig. 1 presents the means of naive estimators $\tilde{\mathbf{D}}_{tt}$ for $t = 1, \ldots, p$ over 200 replications. The plot clearly shows a significant declining trend as the dimension increases. This implies that the bias of $\tilde{\sigma}_t^2$ becomes more severe as t gets larger. Therefore, we need to develop a new statistical method to eliminate the bias.

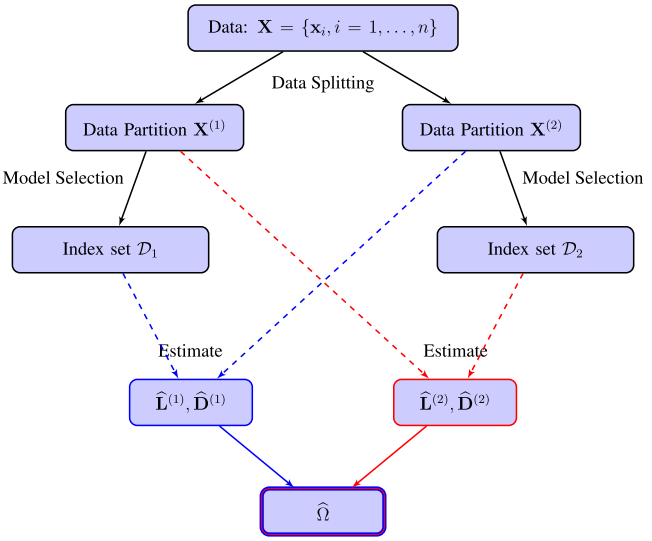

Fig. 2. The flowchart of the RCV procedure for the precision matrix estimation.

2.2. RCV estimator

As shown in Fan et al. (2012), the bias of the naive estimator $\tilde{\sigma}_t^2$ comes from the spuriously correlated covariates in the selected models. To deal with the bias issue, we propose using the RCV method for the estimation of precision matrix. The RCV method has been used to estimate error variance in the ultrahigh dimensional linear regression (Fan et al., 2012) and in ultrahigh dimensional additive models (Chen et al., 2018), but the RCV method is new for estimation of ultrahigh dimensional precision matrix.

RCV procedure for precision matrix estimation consists of the following steps. In the first step, we randomly partition the n samples into two equally sized parts $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. In the second step, we conduct feature screening and variable selection procedures by using dataset $\mathbf{X}^{(1)}$ for the linear model (2.3) and obtain the selected model $\mathcal{D}_{1,t}$, for each $t = 2, \ldots, p$. We refit the data $\mathbf{X}^{(2)}$ to the selected model $\mathcal{D}_{1,t}$, and obtain the regression coefficients $\widehat{\boldsymbol{\beta}}_t^{(1)}$ and the mean squared errors $\widehat{\sigma}_t^{2(1)}$ for $t = 2, \ldots, p$. Based on these estimators $\widehat{\boldsymbol{\beta}}_t^{(1)}$s and $\widehat{\sigma}_t^{2(1)}$s, we can obtain the estimators $\widehat{\mathbf{L}}^{(1)}$ and $\widehat{\mathbf{D}}^{(1)}$ for $\mathbf{L}$ and $\mathbf{D}$, respectively. In the third step, we switch the roles of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. We use $\mathbf{X}^{(2)}$ to select model $\mathcal{D}_{2,t}$ and use $\mathbf{X}^{(1)}$ to refit $\mathcal{D}_{2,t}$ to obtain $\widehat{\boldsymbol{\beta}}_t^{(2)}$ and $\widehat{\sigma}_t^{2(2)}$ for $t = 2, \ldots, p$. We further obtain estimators $\widehat{\mathbf{L}}^{(2)}$ and $\widehat{\mathbf{D}}^{(2)}$. Finally, the RCV estimators of $\mathbf{D}$ and $\mathbf{L}$ are defined as

$$\widehat{\mathbf{D}} = \frac{\widehat{\mathbf{D}}^{(1)} + \widehat{\mathbf{D}}^{(2)}}{2}, \quad \widehat{\mathbf{L}} = \frac{\widehat{\mathbf{L}}^{(1)} + \widehat{\mathbf{L}}^{(2)}}{2}, \tag{2.7}$$

and the RCV estimator for Ω is

$$\widehat{\Omega} = \widehat{\mathbf{L}}^T \widehat{\mathbf{D}}^{-1} \widehat{\mathbf{L}}. \tag{2.8}$$

Since $\widehat{\mathbf{L}}$ is still a lower triangular matrix, the estimator of precision matrix Ω is positive definite. If we partition the data into two parts with different sample sizes n_1 and n_2, $n_1 \neq n_2$, then the RCV estimators of $\mathbf{D}$ and $\mathbf{L}$ are accordingly adjusted to

$$\widehat{\mathbf{L}} = \frac{n_1 \widehat{\mathbf{L}}^{(1)} + n_2 \widehat{\mathbf{L}}^{(2)}}{n_1 + n_2}, \text{ and } \widehat{\mathbf{D}} = \frac{n_1 \widehat{\mathbf{D}}^{(1)} + n_2 \widehat{\mathbf{D}}^{(2)}}{n_1 + n_2}.$$

As a direct comparison, we apply the RCV estimator $\widehat{\mathbf{D}}$ for Example 2.1 and depicts the results in the right panel of Fig. 1, which shows that the two methods present different trends and the proposed method indeed eliminates the bias due to the spurious correlation. Fig. 2 demonstrates the flowchart of the RCV procedure for the precision matrix estimation.

2.3. Asymptotic properties

In this section, we study the asymptotic behavior of the proposed RCV estimator. We first introduce some notation. For a p-dimensional vector $\mathbf{v} = (v_1, \ldots, v_p)^T$, $\|\mathbf{v}\|_\alpha$ stands for the ℓ_α-norm of $\mathbf{v}$. In particular, $\|\mathbf{v}\|_0$ is the number of nonzero elements of $\mathbf{v}$, and $\|\mathbf{v}\|_\infty = \max_j |v_j|$. For a p-by-q matrix $\mathbf{A} = (a_{ij})$, the operator norm is defined by

$$\|\mathbf{A}\|_{(\alpha,\beta)} = \sup_{\|\mathbf{v}\|_\alpha = 1} \|\mathbf{A}\mathbf{v}\|_\beta,$$

where $\mathbf{x}$ is any p-dimensional unit vector and $0 \le \alpha, \beta \le \infty$. In particular,

$$\|\Sigma\|_{(1,1)} = \max_j \sum_i |\sigma_{ij}|, \quad \|\Sigma\|_{(\infty,\infty)} = \max_i \sum_j |\sigma_{ij}|.$$

Define the spectral norm $\|\mathbf{A}\| = \|\mathbf{A}\|_{(2,2)}$. For a $p \times p$ symmetric matrix Σ, it follows that

$$\|\Sigma\| = \lambda_{\max}(\Sigma) \le (\|\Sigma\|_{(1,1)}\|\Sigma\|_{(\infty,\infty)})^{1/2} = \max_j \sum_i |\sigma_{ij}|, \tag{2.9}$$

where $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$ stand for the largest and smallest eigenvalues of Σ, respectively. (See Corollary 2.3.2, Golub and Van Loan (2012)). With a slight abuse of notation, let $\|\Sigma\|_\infty = \max_{i,j} |\sigma_{ij}|$ be the maximum norm of a matrix and $|\mathcal{M}|$ stands for its cardinality of the set $\mathcal{M}$.

Denote $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ and $\mathbf{X}_t$ be a submatrix consisting of the first $(t-1)$ columns of $\mathbf{X}$. Let $\mathcal{M} \subseteq \{1, \ldots, p\}$ be the active index set. $\mathbf{X}_{\mathcal{M},t}$ is the submatrix of $\mathbf{X}_t$ with the columns indexed by $\mathcal{M}$. Define the m-sparse minimal eigenvalue and m-sparse maximal eigenvalue (Meinshausen and Yu, 2009) as

$$\phi_{\min}(m, t) = \min_{\mathcal{M}:|\mathcal{M}|\le m} \lambda_{\min}(n^{-1}\mathbf{X}_{\mathcal{M},t}^T\mathbf{X}_{\mathcal{M},t}),$$

and

$$\phi_{\max}(m, t) = \max_{\mathcal{M}:|\mathcal{M}|\le m} \lambda_{\max}(n^{-1}\mathbf{X}_{\mathcal{M},t}^T\mathbf{X}_{\mathcal{M},t}).$$

We need the following technical conditions to facilitate the proofs, although they may not be the weakest.

(A1) (Exponential tail conditions) X_j and ε_j defined in (2.2), $j = 1, \ldots, p$ satisfy the exponential condition. That is, there exist positive constants c_0 and T such that for $0 < |t| < T$,

$$\mathrm{E}e^{tX_j^2} \le c_0, \text{ and } \mathrm{E}e^{t\varepsilon_j^2} \le c_0, \quad j = 1, \ldots, p. \tag{2.10}$$

(A2) There exist a constant $\lambda_0 > 0$ and constant sequence η_n such that $\eta_n = o(n)$ and $P(\phi_{\min}(\eta_n, t) \ge \lambda_0) = 1$, for all n and $t = 2, \ldots, p$.

(A3) Assume that there exists a constant $\epsilon_0 > 0$ such that $\epsilon_0^{-1} \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) \le \epsilon_0$.

When $\mathbf{x}$ follows a multivariate normal distribution or X_j's are bounded, the exponential tail condition (A1) is satisfied. This assumption is commonly adopted in the literature of high dimensional data modeling (Bickel and Levina, 2008a; Yuan, 2010; Cai et al., 2011; Cai and Liu, 2011).

Condition (A2) is a common assumption for the high dimensional variable selection, which is necessary for LASSO-type method to select all important variables. Here (A2) implies that the selected variables in stage one are not highly correlated. Condition (A3) implies that both the smallest and largest eigenvalues of the precision matrix (and the covariance matrix) are bounded by a constant. Recall notation $\widehat{\mathbf{L}}^{(j)}$, $j = 1, 2$ in Section 2.2, and let $\widehat{\boldsymbol{\ell}}_s^{(j)}$ and $\widehat{\boldsymbol{\ell}}_{(t)}^{(j)T}$ be the sth column and tth row of $\widehat{\mathbf{L}}^{(j)}$, respectively. Denote $k_n = \max_{1\le t\le p} \max_{1\le j\le 2} \max\{\|\widehat{\boldsymbol{\ell}}_t^{(j)}\|_0, \|\widehat{\boldsymbol{\ell}}_{(t)}^{(j)}\|_0\}$.

Theorem 1. *Denote $\mathcal{D}_t^* = \{k : \beta_{tk} \ne 0\}$ and assume that $\mathcal{D}_{k,t}$ satisfies $P(\mathcal{D}_{k,t} \supseteq \mathcal{D}_t^*) = 1$ for $k = 1$ and 2, and $t = 2, \ldots, p$. Suppose that $\log p = O(n^\gamma)$, $\gamma \in [0, 1)$ and $k_n = O\left((n/\log p)^{1/[3(1+\alpha)]}\right)$, for any constant $\alpha > 0$. Under Condition (A1), (A2) and (A3), it follows that*

$$\|\widehat{\Omega} - \Omega\| = O_P\left(\sqrt{\frac{k_n^3 \log p}{n}}\right). \tag{2.11}$$

The condition $\log p = O(n^\gamma)$, $\gamma \in [0, 1)$ is usually adopted by the literature of ultrahigh dimensional data analysis. The order imposed on k_n implies that the number of nonzero elements in each row and each column of $\mathbf{L}^{(j)}$ cannot diverge to infinite too fast. For banded $\mathbf{L}^{(j)}$, we can simply control the number of nonzero elements in $\mathbf{L}^{(j)}$. The convergence rate in Theorem 1 is quite similar to that of banded matrices estimation in Bickel and Levina (2008b). Note the fact that $\mathbf{L}$ of a banded matrix is still banded structure. If $k_n = (n/\log p)^{1/[3(\alpha+1)]}$, for some constant $\alpha > 0$, then the convergence rate of (2.11) becomes $O_P((\log p/n)^{\alpha/[2(\alpha+1)]})$, which is exactly the same as that in Bickel and Levina (2008b). However, the assumption of $\mathbf{L}$ in this paper is more flexible than that in Bickel and Levina (2008b) and Levina et al. (2008). Levina et al. (2008) directly assume the banded structure, and there is no selection procedure involved in their procedure, so the spurious correlation has much less impact on their estimation. In contrast, the proposed method assumes sparsity only rather than the banded structure.

3. Numerical studies

We investigate the finite sample performance of the proposed estimation procedure via Monte Carlo simulation study. We compare the proposed procedure with exiting ones in Section 3.1. We illustrate the proposed methodology by an empirical analysis of real financial market data.

Table 1
Simulation results for precision matrix Ω_1 and Ω_2.

Ω	p	20	50	100	200	500	1000	2000
Ω_1	Sample Covariance	3.30	37.62	534.11	–	–	–	–
	PNL	1.05	3.55	9.25	–	–	–	–
	CLIME	3.49	10.21	22.40	–	–	–	–
	BL	0.50	1.36	2.61	5.4	14.6	28.2	56.9
	Naive-LASSO-AIC	1.42	6.23	17.93	55.2	225.3	670.9	2160.7
	Naive-LASSO-BIC	0.72	1.99	4.40	9.8	27.9	60.9	133.8
	Naive-SCAD-AIC	1.15	5.10	15.71	79.2	1200.0	8804.2	45750.1
	Naive-SCAD-BIC	0.57	1.66	3.97	9.7	31.7	77.1	180.6
	RCV-LASSO-AIC	0.77	2.34	5.36	13.0	45.4	130.9	398.8
	RCV-LASSO-BIC	0.58	1.56	3.54	8.0	24.3	57.6	134.9
	RCV-SCAD-AIC	0.68	2.45	7.66	30.9	181.8	579.9	1579.4
	RCV-SCAD-BIC	0.56	1.54	3.57	8.1	24.3	57.4	136.4
Ω_2	Sample Covariance	3.20	37.06	539.81	–	–	–	–
	PNL	1.05	3.59	9.38	–	–	–	–
	CLIME	3.48	12.79	28.84	–	–	–	–
	BL	3.39	22.46	74.44	193.0	555.8	1185.5	2150.5
	Naive-LASSO-AIC	1.44	6.37	18.74	54.7	233.0	714.6	2338.5
	Naive-LASSO-BIC	0.71	2.06	4.62	10.1	28.9	62.7	137.9
	Naive-SCAD-AIC	1.14	5.27	16.35	78.9	1170.3	8270.2	43641.6
	Naive-SCAD-BIC	0.60	1.73	4.07	10.1	32.9	78.8	185.7
	RCV-LASSO-AIC	0.79	2.43	5.67	13.4	49.2	142.5	433.2
	RCV-LASSO-BIC	0.64	1.69	3.84	8.4	26.2	60.3	141.5
	RCV-SCAD-AIC	0.71	2.49	7.88	31.1	182.7	575.9	1593.4
	RCV-SCAD-BIC	0.62	1.69	3.86	8.5	26.1	60.0	143.2

3.1. Monte Carlo simulation

In this section, we conduct Monte Carlo simulation to examine the finite sample performance of the newly proposed method. We generate random samples from a p-dimensional multi-normal distribution $\mathcal{N}(\mathbf{0}, \Omega^{-1})$. We set the precision matrix Ω through its Cholesky decomposition $\mathbf{L}^T\mathbf{D}^{-1}\mathbf{L}$, and consider four different scenarios.

(a) $\Omega_1 = \mathbf{L}^T\mathbf{D}^{-1}\mathbf{L}$ with $\mathbf{D} = \mathbf{I}$ and $l_{tt} = 1$, $l_{t+1,t} = \rho = 0.5$ with all other $l_{tj} = 0$. This is referred to as the AR(1) covariance matrix in the literature.

(b) $\Omega_2 = \mathbf{L}^T\mathbf{D}^{-1}\mathbf{L}$ with $\mathbf{D} = \mathbf{I}$. We randomly choose an entry $j, j < t$ in the tth row of $\mathbf{L}$, and set $l_{tj} = \rho$ and all other $l_{tj} = 0$.

(c) $\Omega_3 = \mathbf{L}^T\mathbf{D}^{-1}\mathbf{L}$ with $\mathbf{D} = \mathbf{I}$. We randomly choose two entries $j_1, j_2, j_1 < t, j_2 < t$ in the tth row of $\mathbf{L}$, and set $l_{t,j_1} = l_{t,j_2} = \rho$ and all other $l_{tj} = 0$.

(d) $\Omega_4 = \mathbf{L}^T\mathbf{D}^{-1}\mathbf{L}$ with $\mathbf{D} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. $\mathbf{L}$ is set in the same way as that in Ω_3, but $\{\sigma_t^2, t = 1, \ldots, p\}$ are generated by the uniform distribution Unif$[1, 2]$.

In this simulation study, we evaluate estimation procedures by the following quadratic loss

$$\Delta(\widehat{\Omega}, \Omega) = \text{tr}(\Omega^{-1}\widehat{\Omega} - \mathbf{I})^2.$$

Note that the quadratic loss is zero if and only if $\widehat{\Omega} = \Omega$. In our study, we also use different criteria such as the entropy loss (Muirhead, 1982), the Frobenius loss and the Kullback–Leibler divergence. Since the results of numerical comparison based on different measures are almost the same under all of our simulation settings, we only present those based on the quadratic loss to save space.

In our simulation, we set $p = 20, 50, 100, 200, 500, 1000$ and 2000 to investigate the impact of dimensionality, and the sample size $n = 200$. For each combination, we conduct 100 replications. For the purpose of comparison, four existing methods are considered: (1) the inverse of sample covariance matrix when $p < n$, labeled by "Sample Covariance" in tables, (2) the regularized method in Bickel and Levina (2008b), denoted by "BL" for short, (3) the ℓ_1 penalized normal likelihood method in Huang et al. (2006), denoted by "PNL", and (4) the CLIME in Cai et al. (2011), denoted by "CLIME". Since the computation cost of PNL and CLIME are much higher than others, we are able to get their results only when $p < n = 200$.

We also include the naive method in our numerical comparison. For both naive estimator and RCV method, we consider penalized least squares with the LASSO and SCAD penalties with tuning parameter chosen by Akaike information criterion (AIC) or Bayesian information criterion (BIC). This leads to 8 combinations (Naive/RCV-LASSO/SCAD-AIC/BIC).

The simulation results for Ω_1 and Ω_2 are depicted in the top and the bottom panels of Table 1. From the top panel of Table 1, we can see that the BL method preforms the best across all different p dimensions. This is expected because Ω_1 and

Table 2
Simulation results for precision matrix Ω_3 and Ω_4.

Ω	p	20	50	100	200	500	1000	2000
Ω_3	Sample Covariance	3.18	36.65	530.51	–	–	–	–
	PNL	1.53	5.85	16.12	–	–	–	–
	CLIME	8.53	61.55	233.15	–	–	–	–
	BL	3.68	27.64	105.71	345.3	1431.8	3921.5	10462.0
	Naive-LASSO-AIC	1.60	6.39	17.77	48.9	187.7	526.0	1443.1
	Naive-LASSO-BIC	1.04	3.34	7.55	17.2	47.9	103.3	222.0
	Naive-SCAD-AIC	1.05	3.82	9.93	25.9	106.9	347.0	1240.6
	Naive-SCAD-BIC	0.73	2.20	5.01	11.8	37.5	91.9	232.0
	RCV-LASSO-AIC	1.09	3.51	8.08	19.1	55.9	129.7	300.0
	RCV-LASSO-BIC	1.14	3.34	7.00	15.2	38.4	83.3	179.7
	RCV-SCAD-AIC	0.84	2.68	6.62	17.0	60.7	171.8	491.7
	RCV-SCAD-BIC	0.87	2.49	5.57	12.5	34.8	78.5	174.3
Ω_4	Sample	3.23	37.36	537.50	–	–	–	–
	PNL	1.52	6.12	17.69	–	–	–	–
	CLIME	38.25	213.24	710.07	–	–	–	–
	BL	3.71	28.60	107.97	341.2	1439.0	4015.2	10562.6
	Naive-LASSO-AIC	1.68	6.65	17.77	48.5	187.5	528.5	1514.7
	Naive-LASSO-BIC	1.10	3.42	7.78	17.0	47.7	103.4	222.3
	Naive-SCAD-AIC	1.16	4.07	10.30	28.3	121.1	429.2	1710.4
	Naive-SCAD-BIC	0.79	2.28	5.23	12.0	37.9	93.4	233.1
	RCV-LASSO-AIC	1.21	3.68	8.40	18.8	57.1	134.4	323.9
	RCV-LASSO-BIC	1.29	3.49	7.58	15.4	41.5	90.3	203.4
	RCV-SCAD-AIC	0.95	2.90	6.95	17.2	63.7	179.0	515.5
	RCV-SCAD-BIC	0.99	2.68	6.00	12.8	37.2	84.5	194.1

its Cholesky factor **L** are both with banded structure, and 'BL' method is designed for such a setting. Except for 'BL' method, RCV estimator is almost superior to all other estimators. For the naive method, the BIC tuning parameter selector results in significantly smaller quadratic loss than the AIC. This is because the AIC method tends to include more irrelevant variables, which are likely the spuriously correlated predictors. In general, for the naive method and RCV method, the SCAD also has smaller loss than the LASSO for most cases in Table 1. Comparing Naive-LASSO/SCAD-AIC with RCV-LASSO/SCAD-AIC, we can see that the RCV method can effectively eliminate the effect due to spurious correlation.

Comparing the top and bottom panels of Table 1, we can find that the BL method performs much worse for Ω_2 than for Ω_1, while other methods perform similarly for both Ω_1 and Ω_2. This is also expected since the Cholesky factor **L** in Ω_2 does not have regular sparse structure, and thus the banded structure of Ω_2 cannot be guaranteed. This directly leads to the worse performance of BL method for Ω_2. This also implies that the sparsity pattern has less influence on the results from methods other than BL.

Table 2 presents the results for Ω_3 and Ω_4. We can see from Table 2 that the BL method performs poorly for these two precision matrices. Overall patterns in Table 2 are similar to those in the bottom panel of Table 1. For cases in Table 2, the RCV estimators provide the best results, and can effectively improve the estimation accuracy of the naive method when $p > n$.

3.2. Real data analysis

In this section, we illustrate the proposed procedures by an empirical analysis. Portfolio allocation is of great interest in financial econometrics and quantitative finance. In the Markowitz's portfolio theory (Markowitz, 1952), it considers the portfolio allocation with excess returns as an optimization problem

$$\begin{aligned} &\min \ \frac{1}{2}\mathbf{w}^T \Sigma \mathbf{w} \\ &\text{subject to} \ \ \mathbf{w}^T \boldsymbol{\mu} \geq \mu_0, \end{aligned} \tag{3.1}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$ is the mean vector of excess returns of the p assets, Σ is the covariance matrix of assets returns, μ_0 is the expected return. The optimal solution (μ^*, σ^*), the expected returns and the stand deviation, constitutes an efficient frontier (Markowitz, 1952), which theoretically clarifies that higher expected returns always comes with higher risks. Consequently, Sharpe (1966) and Sharpe (1994) proposed using the ratio of the return to the risk as a measure of portfolio allocation. The ratio is called the Sharpe ratio and represents the portfolio return per unit risk. Portfolio optimization still is an active research topic (Cai et al., 2019; Ao et al., 2018).

In this example, we collect monthly excess returns for stocks in S&P 500 index that have complete records from January 1980 to December 2012. The data, collected from the Center for Research in Security Prices (CRSP), contain the returns of 202 stocks with a time span of 396 months. We avoid the period of the subprime crisis starting in 2008 since the market performance was totally different during that period (Fig. 3). We set the returns of the last 6, 12 or 18 months as

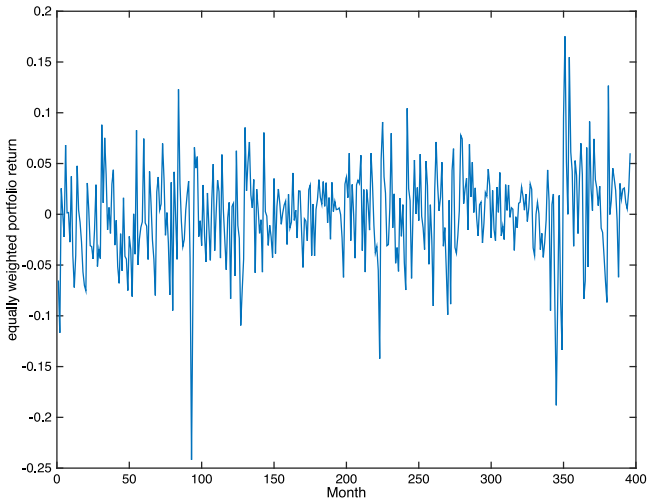

Fig. 3. Equally weighted monthly Returns of portfolio (Jan. 1980∼Dec. 2012).

the testing data and the remaining data as the training data to study the asset allocation. Denote the training data of the period from Jan. 1980 to Dec. 2006 by $\mathbf{X}^{(1)} = \{X_{ij}^{(1)}\},\ i = 1, \ldots, 324, j = 1, \ldots, 202$, and the corresponding testing data by $\mathbf{Y}^{(1)} = \{Y_{ij}^{(1)}\},\ i = 1, \ldots, 12, j = 1, \ldots, 202$.

It is well known that the optimal weights (i.e., the minimizer of (3.1)) is

$$\mathbf{w}_{\text{opt}} = \frac{\mu_0 \Sigma^{-1}\boldsymbol{\mu}}{\boldsymbol{\mu}^T \Sigma^{-1}\boldsymbol{\mu}} = \frac{\mu_0 \Omega \boldsymbol{\mu}}{\boldsymbol{\mu}^T \Omega \boldsymbol{\mu}}, \tag{3.2}$$

where $\Omega = \Sigma^{-1}$. To illustrate the proposed procedure, we consider the annualized return rates (ARR) at 10%, 20% and 30%, respectively. By using the theory of fixed income securities, the corresponding monthly return rates are 0.8%, 1.53% and 2.21% respectively. To obtain the optimal weights, Ω and $\boldsymbol{\mu}$ should be estimated through the training data. The mean vector $\boldsymbol{\mu}$ of the assets returns is estimated by the sample mean vector $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_p)^T = (n^{-1}\sum_{i=1}^n X_{i1}, \ldots, n^{-1}\sum_{i=1}^n X_{ip})^T$.

The inverse of the sample covariance matrix $\mathbf{S}^{-1}$ does not perform very well due to the small sample size. Thus, we apply the naive method and RCV method to estimate the portfolio allocation weights through the training data and to investigate the returns of the newly constructed portfolios through the testing data. The results are presented as the returns, the risks and the return–risk ratios. This ratio can be considered as the measure of excess return per unit of the investment risk, which is the realized Sharpe ratio (Sharpe, 1966, 1994).

For the purpose of comparison, we use the generalized inverse of the sample covariance matrix as the benchmark, and consider other estimates: (1) the ℓ_1-regularization. The tuning parameter are chosen by AIC or BIC; (2) The regularization with SCAD penalty. The tuning parameter is chosen by AIC or BIC; (3) The BL estimator (Bickel and Levina, 2008b). All regularization estimators are improved by RCV technique.

We first focus on the comparison between the naive method and the RCV method with different regularization methods. Table 3 presents the mean return, the risk and the Sharpe ratio. It clearly shows that, in terms of the Sharpe ratio, the RCV methods perform better than the naive methods when these two methods are implemented with the same regularization method for most cases. The RCV with SCAD-BIC performs the best among the four combinations of RCV methods (LASSO-AIC, SCAD-AIC, LASSO-BIC and SCAD-BIC).

Next, we compare the performance of the RCV-SCAD-BIC estimation, the generalized inverse of sample covariance matrix (labeled as Sample Cov in Table 3) and the BL estimation. From Table 3, the RCV method performs the best among these four methods in terms of the ratio.

4. Discussions

In this paper, we have proposed the RCV estimation for ultrahigh dimensional precision matrix, and studied the asymptotical properties of the RCV estimator. The RCV method may be used to estimate covariance matrix too since the Cholesky decomposition has been used to estimate covariance matrix (Pourahmadi, 1999, 2000; Huang et al., 2006). It will be of interest to examine the performance of the RCV estimation procedure for covariance matrix in the future study.

As one referee points out that the sparsity of linear regression models resulted from Cholesky decomposition depends on the order of variables. For this point, we regard it as a minor weakness of the proposed estimation procedure. The

Table 3
Estimated mean return and risk of the S&P 500 portfolios analysis by using excess returns.

		6-month prediction with 30% ARR		12-month prediction with 10% ARR		18-month prediction with 20% ARR	
		Naive	RCV	Naive	RCV	Naive	RCV
LASSO-AIC	Return	−0.0011	0.0062	0.0010	0.0013	−0.0010	0.0001
	Risk	0.0209	0.0254	0.0062	0.0067	0.0108	0.0131
	Ratio	−0.0517	0.2456	0.1583	0.2014	−0.0896	0.0104
SCAD-AIC	Return	0.0023	0.0086	0.0017	0.0013	0.0006	−0.0005
	Risk	0.0263	0.0265	0.0070	0.0068	0.0120	0.0130
	Ratio	0.0884	0.3260	0.2457	0.1842	0.0463	−0.0409
LASSO-BIC	Return	0.0022	0.0064	0.0014	0.0016	0.0007	0.0014
	Risk	0.0172	0.0144	0.0052	0.0054	0.0095	0.0105
	Ratio	0.1290	0.4460	0.2722	0.2932	0.0728	0.1314
SCAD-BIC	Return	0.0025	0.0071	0.0014	0.0016	0.0010	0.0016
	Risk	0.0177	0.0154	0.0055	0.0055	0.0096	0.0100
	Ratio	0.1421	0.4615	0.2532	0.2965	0.1017	0.1623
Sample Cov	Return	0.0029		0.0048		0.0003	
	Risk	0.0317		0.0198		0.0104	
	Ratio	0.0902		0.2450		0.0284	
BL	Return	0.0024		0.0017		0.0003	
	Risk	0.0104		0.0098		0.0049	
	Ratio	0.2302		0.1750		0.0586	

reason is that banded structure on covariance matrix or precision matrix has been assumed in the literature and the banded structure also depends on the order of variables. In addition, we can deal with this problem by sorting the variables according to the magnitude of their variance from the smallest to the largest in practice. This is implemented in our real data analysis.

Acknowledgments

Chen's work was supported by National Natural Science Foundation of China (NNSFC) grants 11690014 and 11690015, the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education KLATASDS1807. Li's work was supported by US National Science Foudation (NSF) grant DMS 1820702 and NIDA, NIH grant P50 DA039838. The content is solely the responsibility of the authors and does not necessarily represent the official views of NNSFC, NSF, NIH or NIDA. The authors thank the AE, and the reviewers for their constructive comments, which have led to a significant improvement of the earlier version of this article.

Appendix

For ease of presentation, let $\mathbf{y}$ the t-column of $\mathbf{X}$ defined in Section 2.3, and $\mathbf{Z} = \mathbf{X}_t$. Rewrite Eq. (2.3) in the matrix form:

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{A.1}$$

where we also compress the subscript in $\boldsymbol{\beta}_t$. For an index set $\mathcal{M}$, denote $\mathbf{Z}_{\mathcal{M}} = \mathbf{X}_{\mathcal{M},t}$, and $\phi_{\min}(m)$ and $\phi_{\max}(m)$ stand for $\phi_{\min}(m, t)$ and $\phi_{\max}(m, t)$, defined in Section 2.3, respectively.

Define the event $\mathcal{E}_{n,t} = \{\max_{1\le j\le t-1} |\mathbf{Z}_j^T \boldsymbol{\varepsilon}| \le c\sqrt{n \log p}\}$, where $\mathbf{Z}_j$ is the jth column of the sample matrix $\mathbf{Z}$ and c is a positive constant. To prove Theorem 1, we need the following two lemmas.

Lemma A.1. *Suppose Condition **(A1)** hold. If $\log p/n = O(1)$, under model (A.1) it follows that $P(\mathcal{E}_{n,t}) \to 1$.*

Proof. By using the exponential tail condition (A1), we have, for any $s \ge 1$,

$$\begin{aligned} &\mathrm{P}\left(|Z_{ij}\varepsilon_i| \ge s\right) \le \mathrm{P}\left(|Z_{ij}| \ge s^{1/2}\right) + \mathrm{P}\left(|\varepsilon_i| \ge s^{1/2}\right) \\ &\le \mathrm{E}\exp\{Z_{ij}^2\}\exp\{-s\} + \mathrm{E}\exp\{\varepsilon_i^2\}\exp\{-s\} \\ &\le c_0 \exp\{-s\}. \end{aligned} \tag{A.2}$$

By inequality (A.2) and using the integration by parts, there exists a constant $c_1 > 0$ such that $\mathrm{E}\exp\{2^{-1}|Z_{ik}\varepsilon_i|\} \le 2 + 2c_1$. For any $m \ge 2$, we have

$$\begin{aligned} \mathrm{E}\left|Z_{ij}\varepsilon_i\right|^m &\le 2^m m!\,\mathrm{E}\exp\left\{2^{-1}|Z_{ik}\varepsilon_i|\right\} \\ &\le 2^m m!\,(2 + 2c_1) \\ &= 2^{-1}\left(8(2 + 2c_1)\right) 2^{m-2} m! \end{aligned} \tag{A.3}$$

From Bernstein's inequality (Theorem 2.2.11 in Vaart and Wellner (1996)),

$$P(|\mathbf{Z}_j^T\boldsymbol{\varepsilon}| \geq s) \leq 2\exp\left\{-\frac{s^2}{4(nL_1+s)}\right\}, \tag{A.4}$$

where $L_1 = 4(2+2c_1)$. Thus, it follows that

$$\begin{aligned} &\mathrm{P}\left(\max_{1\leq j\leq t-1}\left|\mathbf{Z}_j^T\boldsymbol{\varepsilon}\right| \geq c_2\sqrt{n\log p}\right) \leq (t-1)\,\mathrm{P}\left(\left|\mathbf{Z}_j^T\boldsymbol{\varepsilon}\right| \geq c_2\sqrt{n\log p}\right) \\ &\leq 2\exp\left\{\log(t-1)\left[1-\frac{1}{4(L_1c_2^{-2}+c_2^{-1}\sqrt{\log p/n})}\right]\right\}. \end{aligned} \tag{A.5}$$

Since $t < p$, $\log p/n = o(1)$, when c_2 is large enough, $4L_1c_2^{-2} + 4c_2^{-1}\sqrt{\log p/n} < 1$. It turns out that $P(\mathcal{E}_{n,t}^c) \to 0$, that is $P(\mathcal{E}_{n,t}) \to 1$.

Lemma A.2. *Under Conditions in Theorem 1, it follows that*

$$\begin{aligned} &\|\widehat{\mathbf{L}}^{(j)} - \mathbf{L}\|_\infty = O_P\left(\sqrt{k_n\log p/n}\right), \\ &\|\widehat{\mathbf{D}}^{(j)} - \mathbf{D}\|_\infty = O_P\left(n^{-1/2}\log^{1/2} p\right), \end{aligned} \tag{A.6}$$

for $j = 1, 2$.

Proof. It is sufficient to show that (A.6) holds for $j = 2$. For model (A.1), we obtain by data splitting that $\mathbf{y}_t^{(j)}$, $\mathbf{Z}_t^{(j)}$ and $\varepsilon_t^{(j)}$, $j = 1$ and 2, emphasize their dependence on t. Consider the linear model

$$\mathbf{y}_t^{(2)} = \mathbf{Z}_t^{(2)}\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t^{(2)}. \tag{A.7}$$

By their definitions, the tth row of $\mathbf{L}$ is $-\boldsymbol{\beta}_t^T$, and the tth diagonal element of $\mathbf{D}$ is σ_t^2, the error variance in (A.7). Let $\mathcal{D}_{1t}$ is the index of significant predictors that are selected from $X_1, \ldots, X_{t-1}$ based on $\mathbf{Z}_t^{(1)}$, and consider

$$\mathbf{y}_t^{(2)} = \mathbf{Z}_{\mathcal{D}_{1t}}^{(2)}\boldsymbol{\beta}_{\mathcal{D}_{1t}} + \boldsymbol{\varepsilon}_t^{(2)}. \tag{A.8}$$

Let $\widehat{\boldsymbol{\beta}}_{\mathcal{D}_{1t}}^{(2)}$ be the least squares estimate from (A.8), and $\widehat{\sigma}_t^2$ is the mean squares errors from (A.8). We set the t-the diagonal element of $\widehat{\mathbf{D}}$ to be $\widehat{\sigma}_t^2$, and estimate $\boldsymbol{\beta}_{\mathcal{D}_{1t}}$ by $\widehat{\boldsymbol{\beta}}_{\mathcal{D}_{1t}}^{(2)}$, and estimate $\boldsymbol{\beta}_{\mathcal{D}_{1t}^c}$ by $\mathbf{0}$. To show (A.6), we establish the concentration inequality for $\widehat{\boldsymbol{\beta}}_{\mathcal{D}_{1t}}^{(2)}$ and $\widehat{\sigma}_t^2$. To this end, we first derive the following tail probability

$$\mathrm{P}\left(\left\|\widehat{\boldsymbol{\beta}}_{\mathcal{D}_{1t}}^{(2)} - \boldsymbol{\beta}_{\mathcal{D}_{1t}}\right\|_\infty \geq c_1\sqrt{\log p/n}\right) \tag{A.9}$$

for some constant c_1, and

$$P\left(\max_{1\leq t\leq p}|\widehat{\sigma}_t^2 - \sigma_t^2| \geq c_2\sqrt{\log p/n}\right) \tag{A.10}$$

for some constant c_2.

By the assumption $\mathrm{P}(\mathcal{D}_{1t} \supseteq \mathcal{D}^*) = 1$ in Theorem 1 and using the proof of (A.5), it follows that

$$\mathrm{P}\left(\max_{k\in\mathcal{D}_{1t}}\left|\mathbf{Z}_k^{(2)T}\boldsymbol{\varepsilon}_t^{(2)}\right| \geq c\sqrt{n\log k_n}\right) \leq 2\exp\left\{\log k_n\left[1-\frac{1}{4(L_1c^{-2}+c^{-1}\sqrt{\log k_n/n})}\right]\right\} \tag{A.11}$$

When c is large enough, $4(L_1c^{-2} + c^{-1}\sqrt{\log k_n/n}) < 1$, and the exponential part of the last equation in (A.5) is negative.

Note that

$$\widehat{\boldsymbol{\beta}}_{\mathcal{D}_{1t}}^{(2)} = \boldsymbol{\beta}_{\mathcal{D}_{1}t} + (\mathbf{Z}_{\mathcal{D}_{1t}}^{(2)T}\mathbf{Z}_{\mathcal{D}_{1t}}^{(2)})^{-1}\mathbf{Z}_{\mathcal{D}_{1t}}^{(2)T}\boldsymbol{\varepsilon}_t^{(2)}. \tag{A.12}$$

Then

$$\begin{aligned} &\mathrm{P}\left(\left\|\widehat{\boldsymbol{\beta}}_{\mathcal{D}_{1t}}^{(2)} - \boldsymbol{\beta}_{\mathcal{D}_{1t}}\right\|_\infty \geq c_1\sqrt{\log p/n}\right) \\ &= \mathrm{P}\left(\left\|(\frac{2}{n}\mathbf{Z}_{\mathcal{D}_{1t}}^{(2)T}\mathbf{Z}_{\mathcal{D}_{1t}}^{(2)})^{-1}\frac{2}{n}\mathbf{Z}_{\mathcal{D}_{1t}}^{(2)T}\boldsymbol{\varepsilon}_t^{(2)}\right\|_\infty \geq c_1\sqrt{\log p/n}\right) \end{aligned}$$

Let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_p)^T$ be a p-by-p positive definite matrix and $\mathbf{b}$ be a p-dimension vector. Then

$$\|\mathbf{Ab}\|_\infty = \max_{k\in\{1,\ldots,p\}}\left|\mathbf{a}_k^T\mathbf{b}\right| \leq \|\mathbf{b}\|_\infty \max_k\sum_{i=1}^p|\mathbf{a}_{ki}| \leq \|\mathbf{b}\|_\infty\|\mathbf{A}\|_{(1,1)} \leq \sqrt{p}\,\|\mathbf{b}\|_\infty\|\mathbf{A}\|.$$

By using Condition (A2) and (A.11), we have

$$\begin{aligned}
&\mathrm{P}\left(\left\|\widehat{\boldsymbol{\beta}}^{(2)}_{\mathcal{D}_{1t}}-\boldsymbol{\beta}_{\mathcal{D}_{1t}}\right\|_\infty \geq c_1\sqrt{k_n\log k_n/n}\right)\\
&\leq \mathrm{P}\left(\left\|(\frac{2}{n}\mathbf{Z}^{(2)T}_{\mathcal{D}_{1t}}\mathbf{Z}^{(2)}_{\mathcal{D}_{1t}})^{-1}\right\|_2\left\|\frac{2}{n}\mathbf{Z}^{(2)T}_{\mathcal{D}_{1t}}\boldsymbol{\varepsilon}^{(2)}_t\right\|_\infty \geq c_1k_n^{-1/2}\sqrt{k_n\log k_n/n}\right)\\
&\leq \mathrm{P}\left(\left\|\frac{2}{n}\mathbf{Z}^{(2)T}_{\mathcal{D}_{1t}}\boldsymbol{\varepsilon}^{(2)}_t\right\|_\infty \geq \varepsilon_0c_1\sqrt{\log k_n/n}\right)\\
&= \mathrm{E}\left\{\mathrm{P}\left(\left\|\mathbf{Z}^{(2)T}_{\mathcal{D}_{1t}}\boldsymbol{\varepsilon}^{(2)}_t\right\|_\infty \geq \varepsilon_0c_1\sqrt{n\log k_n}\,\middle|\,\mathbf{X}^{(1)}\right)\right\}\\
&\leq 2\exp\left\{\log k_n\left[1-\frac{1}{4(L_1c_1^{-2}\varepsilon_0^{-2}+c_1^{-1}\varepsilon_0^{-1}\sqrt{\log k_n/n})}\right]\right\}.
\end{aligned}\tag{A.13}$$

Since $k_n^2=o(n)$, when c_1 is large enough, the right hand side of the above inequality goes to 0. Finally we have

$$\begin{aligned}
&\mathrm{P}\left(\left\|\widehat{\mathbf{L}}^{(2)}-\mathbf{L}\right\|_\infty \geq c_1\sqrt{k_n\log k_n\log p/n}\right)\\
&= \mathrm{P}\left(\max_{t\in\{1,\ldots,p\}}\left\|\widehat{\boldsymbol{\beta}}^{(2)}_{\mathcal{D}_{1t}}-\boldsymbol{\beta}_{\mathcal{D}_{1t}}\right\|_\infty \geq c_1\sqrt{k_n\log k_n\log p/n}\right)\\
&\leq p\mathrm{P}\left(\left\|\widehat{\boldsymbol{\beta}}^{(2)}_{\mathcal{D}_{1t}}-\boldsymbol{\beta}_{\mathcal{D}_{1t}}\right\|_\infty \geq c_1\sqrt{k_n\log p/n}\right)\\
&\leq 2\exp\left\{\log p\left[1-\frac{1}{4(L_1c^{-2}\varepsilon_0^{-2}+c_1^{-1}\varepsilon_0^{-1}\sqrt{\log p/n})}\right]\right\}\to 0.
\end{aligned}$$

Notice that $\log p=O(n^{\alpha_0})$, $k_n=O(n^{(1-\alpha_0)/2})$, $\alpha_0\in[0,1)$. Thus, the first equation in (A.6) holds.

Next we consider $\|\widehat{\mathbf{D}}^{(2)}-\mathbf{D}\|_\infty$. Denote $\widehat{d}_t^{(2)}$ and d_t to be the tth element of $\widehat{\mathbf{D}}^{(2)}$ and $\mathbf{D}$. Thus $\widehat{d}_t^{(2)}=\widehat{\sigma}_t^2$, the mean squared error of the least squares fit of model (A.7), and $d_t=\sigma_t^2$, the error variance of model (A.7). As a result, it can be written as, for $t=1,\ldots,p$,

$$\begin{aligned}
\widehat{d}_t^{(2)}-d_t &= (n/2)^{-1}\boldsymbol{\varepsilon}^{(2)T}(\mathbf{I}-\mathbf{P}^{(2)}_{\mathcal{D}_{1t}})\boldsymbol{\varepsilon}^{(2)}-d_t\\
&= (n/2)^{-1}(\boldsymbol{\varepsilon}^{(2)T}\boldsymbol{\varepsilon}^{(2)}-2^{-1}nd_t)-(n/2)^{-1}\boldsymbol{\varepsilon}^{(2)T}\mathbf{P}^{(2)}_{\mathcal{D}_{1t}}\boldsymbol{\varepsilon}^{(2)}
\end{aligned}\tag{A.14}$$

where $\mathbf{P}^{(2)}_{\mathcal{D}_{1t}}$ is the projection matrix consisting of variables indexed by $\mathcal{D}_{1t}$. Define the events $\mathcal{A}_{n_1}=\{\mathcal{D}^*\subset\mathcal{D}_{1t}\}$ and $\mathcal{A}_{n_2}=\{\mathcal{D}^*\subset\mathcal{D}_{2t}\}$, where $\mathcal{D}^*$ is the true significant variable set of the jth regression model (A.14). For convenience, we still abbreviate the subscript j in the above equations.

$$\begin{aligned}
&\mathrm{P}\left(\left|\widehat{d}_t^{(2)}-d_t\right|\geq\nu\right)\\
&\leq \mathrm{P}\left((n/2)^{-1}\left|\boldsymbol{\varepsilon}^{(2)T}\boldsymbol{\varepsilon}^{(2)}-2^{-1}nd_t\right|\geq\nu/2\right)\\
&\quad+\mathrm{P}\left((n/2)^{-1}\boldsymbol{\varepsilon}^{(2)T}\mathbf{P}^{(2)}_{\mathcal{D}_{1t}}\boldsymbol{\varepsilon}^{(2)}\geq\nu/2\right).
\end{aligned}\tag{A.15}$$

Similar to Eq. (A.4), the first term follows that

$$\mathrm{P}\left((n/2)^{-1}\left|\boldsymbol{\varepsilon}^{(2)T}\boldsymbol{\varepsilon}^{(2)}-2^{-1}nd_t\right|\geq\nu/2\right)\leq 2\exp\left\{-\frac{n^2\nu^2}{16(nL_2\kappa_t+n\nu)}\right\},\tag{A.16}$$

where κ_t is the fourth moment of ε_{it}. For the second term, we have

$$\mathrm{P}\left((n/2)^{-1}\boldsymbol{\varepsilon}^{(2)T}\mathbf{P}^{(2)}_{\mathcal{D}_{1t}}\boldsymbol{\varepsilon}^{(2)}\geq\nu/2\right)\leq\ \mathrm{P}\left((n/2)^{-1}\boldsymbol{\varepsilon}_t^{(2)T}\mathbf{Z}^{(2)}_{\mathcal{D}_{1t}}(\mathbf{Z}^{(2)T}_{\mathcal{D}_{1t}}\mathbf{Z}^{(2)}_{\mathcal{D}_{1t}})^{-1}\mathbf{Z}^{(2)T}_{\mathcal{D}_{1t}}\boldsymbol{\varepsilon}^{(2)}_t\geq\nu/4\right).\tag{A.17}$$

By using the condition (A2), it follows that

$$\begin{aligned}
&\mathrm{P}\left(\boldsymbol{\varepsilon}_t^{(2)T}\mathbf{Z}^{(2)}_{\mathcal{D}_{1t}}(\mathbf{Z}^{(2)T}_{\mathcal{D}_{1t}}\mathbf{Z}^{(2)}_{\mathcal{D}_{1t}})^{-1}\mathbf{Z}^{(2)T}_{\mathcal{D}_{1t}}\boldsymbol{\varepsilon}^{(2)}_t\geq\nu\right)\\
&\leq \mathrm{P}\left(\left\|(\mathbf{Z}^{(2)T}_{\mathcal{D}_{1t}}\mathbf{Z}^{(2)}_{\mathcal{D}_{1t}})^{-1}\right\|\left\|\mathbf{Z}^{(2)T}_{\mathcal{D}_{1t}}\boldsymbol{\varepsilon}^{(2)}_t\right\|_2^2\geq\nu\right)\leq\mathrm{P}\left(\left\|\mathbf{Z}^{(2)T}_{\mathcal{D}_{1t}}\boldsymbol{\varepsilon}^{(2)}_t\right\|_2^2\geq\lambda_0^{-1}n\nu\right)
\end{aligned}\tag{A.18}$$

Since $\mathcal{D}_{1t}$ comes from anther dataset $\mathcal{I}_1$, it is independent of $(\mathbf{Z}_t^{(2)},\boldsymbol{\varepsilon}_t^{(2)})$. Recall the definition of k_n. With the probability one, we have

$$\left\|\mathbf{Z}^{(2)T}_{\mathcal{D}_{1t}}\boldsymbol{\varepsilon}^{(2)}_t\right\|_2^2=\sum_{j\in\mathcal{D}_{1t}}(\mathbf{x}_j^{(2)T}\boldsymbol{\varepsilon}_t)^2\leq k_n\max_{j\in\mathcal{D}_{1t}}\left|\mathbf{x}_j^{(2)T}\boldsymbol{\varepsilon}_t\right|^2,\tag{A.19}$$

where $\mathbf{x}_j^{(2)}$ are the observations in $\mathcal{I}_2$ of covariate X_j. Since X_j and ε_t are uncorrelated, $\mathrm{E}\mathbf{x}_j^{(2)T}\boldsymbol{\varepsilon}_t^{(2)} = 0$. Next consider the concentration inequality for the term $\left|\mathbf{x}_j^{(2)T}\boldsymbol{\varepsilon}_t^{(2)}\right|$. Provided that $\mathbf{x}_j$ and ε_t satisfy the condition (A1) and are uncorrelated, we still have the Bernstein's inequality

$$\mathrm{P}\left(\max_{j\in\mathcal{D}_{1t}}\left|\mathbf{x}_j^{(2)T}\boldsymbol{\varepsilon}_t^{(2)}\right| > \nu_1/2\right) \le k_n\mathrm{P}\left(\left|\sum_{i=1}^{n} X_{ij}^{(2)}\varepsilon_{it}^{(2)}\right| > \nu_1/2\right) \le 2k_n\exp\left\{-\frac{\nu_1^2}{16(nL_1+\nu_1/2)}\right\},$$

where L_1 is similarly defined in the lemma A.1. The last inequality holds due to Eq. (A.3) and (A.4). Consequently, we obtain that

$$\begin{aligned}
&\mathrm{P}\left(\boldsymbol{\varepsilon}_t^{(2)T}\mathbf{P}_{\mathcal{D}_{1t}}^{(2)}\boldsymbol{\varepsilon}_t^{(2)} \ge \nu\right)\\
&\le \mathrm{P}\left(\max_{j\in\mathcal{D}_{1t}}\left|\mathbf{x}_j^{(2)T}\boldsymbol{\varepsilon}_t^{(2)}\right|^2 > \lambda_0 k_n^{-1} n\nu\right)\\
&\le 2k_n\exp\left\{-\frac{\lambda_0 k_n^{-1} n\nu}{4nL_1 + 4\sqrt{\lambda_0 k_n^{-1} n\nu}}\right\}\\
&= 2\exp\left\{\log k_n\left(1-\left(4L_1\frac{nk_n\log k_n}{\lambda_0 n\nu} + 4\sqrt{\frac{k_n(\log k_n)^2}{\lambda_0 n\nu}}\right)^{-1}\right)\right\}.
\end{aligned} \tag{A.20}$$

Taking $\nu = c_3 k_n \log k_n$, when c_3 is large enough, the last equation in (A.20) will exponentially converge to 0. Together with (A.15), (A.16) and (A.20), the leading term $\boldsymbol{\varepsilon}_t^{(2)T}\boldsymbol{\varepsilon}_t^{(2)}$ dominates the residual term $\boldsymbol{\varepsilon}_t^{(2)T}\mathbf{P}_{\mathcal{D}_{t1}}^{(2)}\boldsymbol{\varepsilon}_t^{(2)}$. Furthermore, we apply the similar arguments to study the diagonal matrix $\widehat{\mathbf{D}}$. Then under event $\mathcal{A}_n = \mathcal{A}_{n_1}\bigcap\mathcal{A}_{n_2}$, taking $\nu = c_4\sqrt{\log p/n}$, provided $k_n = o(n^{1/2})$, we get that

$$\begin{aligned}
&\mathrm{P}\left(\max_{t\in\{1,\ldots,p\}}\left|\widehat{d}_t^{(2)} - d_t\right| \ge c_4\sqrt{\log p/n}\right)\\
&\le p\ \mathrm{P}\left(\left|\widehat{d}_t^{(2)} - d_t\right| \ge c_4\sqrt{\log p/n}\right)\\
&\le 2p\exp\left\{-\frac{c_4^2 n\log p}{16(L_2\kappa_k n + c_4\sqrt{n\log p})}\right\} + 2pk_n\exp\left\{-\frac{c_4\lambda_0 k_n^{-1} n\sqrt{n\log p}}{4L_1 n + 4(c_4\lambda_0 k_n^{-1} n\sqrt{n\log p})^{1/2}}\right\}.
\end{aligned} \tag{A.21}$$

Since $k_n = O((n/\log p)^{1/[3(1+\alpha)]})$, $\alpha > 0$ and $\log p = n^{\gamma}$, $0 \le \gamma < 1$, the last two terms in (A.21) go to 0. Therefore,

$$\|\widehat{\mathbf{D}}^{(2)} - \mathbf{D}\|_\infty = \max_{t\in\{1,\ldots,p\}}\left|\widehat{d}_t^{(2)} - d_t^{(2)}\right| = O_P\left((n^{-1}\log p)^{1/2}\right). \tag{A.22}$$

Proof of Theorem 1. By assumption $0 < \epsilon_0^{-1} \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) \le \epsilon_0$, it follows that $\|\mathbf{L}\| = \|\mathbf{D}\| = O(1)$. Since $\mathbf{D}$ is a diagonal matrix and RCV estimator $\widehat{\mathbf{D}} = (\widehat{\mathbf{D}}^{(1)} + \widehat{\mathbf{D}}^{(2)})/2$, it follows that

$$\left\|\widehat{\mathbf{D}} - \mathbf{D}\right\| = \left\|\widehat{\mathbf{D}} - \mathbf{D}\right\|_\infty = O_p(\sqrt{\log p/n}). \tag{A.23}$$

Because of $\left|\widehat{d}_t^{-1} - d_t^{-1}\right| = \left|\widehat{d}_t - d_t\right| / \left|\widehat{d}_t d_t\right|$ and $\mathrm{E}\widehat{d}_t = d_t$, $\left\|\widehat{\mathbf{D}}^{-1} - \mathbf{D}^{-1}\right\| \asymp \left\|\widehat{\mathbf{D}} - \mathbf{D}\right\|$. Using Hölder's inequality for matrix norm, $\|\mathbf{A}\| \le \sqrt{\|\mathbf{A}\|_{(1,1)}\|\mathbf{A}\|_{(\infty,\infty)}}$ and Lemma A.2., it follows that $\left\|\widehat{\mathbf{L}} - \mathbf{L}\right\| \le k_n\left\|\widehat{\mathbf{L}} - \mathbf{L}\right\|_\infty = O_P\left(k_n\sqrt{k_n\log p/n}\right)$. Use an inequality in Bickel and Levina (2008a),

$$\begin{aligned}
&\|\mathbf{A}_1\mathbf{A}_2\mathbf{A}_3 - \mathbf{B}_1\mathbf{B}_2\mathbf{B}_3\|\\
&\le \sum_{j=1}^{3}\|\mathbf{A}_j - \mathbf{B}_j\|\prod_{k\ne j}\|\mathbf{B}_k\| + \sum_{j=1}^{3}\|\mathbf{B}_j\|\prod_{k\ne j}\|\mathbf{A}_k - \mathbf{B}_k\| + \sum_{j=1}^{3}\|\mathbf{A}_j - \mathbf{B}_j\|.
\end{aligned} \tag{A.24}$$

Plugging $\mathbf{A}_1 = \mathbf{A}_3^T = \widehat{\mathbf{L}}, \mathbf{A}_2 = \widehat{\mathbf{D}}^{-1}$, and $\mathbf{B}_1 = \mathbf{B}_3^T = \mathbf{L}, \mathbf{B}_2 = \mathbf{D}^{-1}$ into (A.24). By Assumption (A3), $\|\mathbf{L}\| = O(1)$ and $\|\mathbf{D}^{-1}\| = O(1)$. Thus, $\left\|\widehat{\Omega} - \Omega\right\| = \left\|\widehat{\mathbf{L}}^T\widehat{\mathbf{D}}^{-1}\widehat{\mathbf{L}} - \mathbf{L}^T\mathbf{D}^{-1}\mathbf{L}\right\| = O_P(k_n\sqrt{k_n\log p/n})$. The proof of Theorem 1 is completed.

References

Ao, M., Li, Y., Zheng, X., 2018. Approaching mean-variance efficiency for large portfolios. Rev. Financ. Stud. 32 (7), 2890–2919.

Bickel, P.J., Levina, E., 2008a. Covariance regularization by thresholding. Ann. Statist. 36 (6), 2577–2604.

Bickel, P.J., Levina, E., 2008b. Regularized estimation of large covariance matrices. Ann. Statist. 36 (1), 199–227.

Cai, T., Hu, J., Li, Y., Zheng, X., 2019. High-dimensional minimum variance portfolio estimation based on high-frequency data. J. Econometrics (in press).

Cai, T., Liu, W., 2011. Adaptive thresholding for sparse covariance matrix estimation. J. Amer. Statist. Assoc. 106 (494), 672–684.

Cai, T., Liu, W., Luo, X., 2011. A constrained l(1) minimization approach to sparse precision matrix estimation. J. Amer. Statist. Assoc. 106 (494), 594–607.
Chen, Z., Fan, J., Li, R., 2018. Error variance estimation in ultrahigh-dimensional additive models. J. Amer. Statist. Assoc. 113 (521), 315–327.
Fan, J., Guo, S., Hao, N., 2012. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. J. R. Stat. Soc. Ser. B Stat. Methodol. 74 (1), 37–65.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.
Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (5), 849–911.
Fan, Y., Lv, J., 2016. Innovated scalable efficient estimation in ultra-large Gaussian graphical models. Ann. Statist. 44 (5), 2098–2126.
Golub, G.H., Van Loan, C.F., 2012. Matrix Computations, vol. 3. JHU Press.
Huang, J.Z., Liu, N., Pourahmadi, M., Liu, L., 2006. Covariance matrix selection and estimation via penalised normal likelihood. Biometrika 93 (1), 85–98.
Lam, C., Fan, J., 2009. Sparsistency and rates of convergence in large covariance matrix estimation. Ann. Statist. 37 (6B), 4254–4278.
Levina, E., Rothman, A., Zhu, J., 2008. Sparse estimation of large covariance matrices via a nested lasso penalty. Ann. Appl. Stat. 2 (1), 245–263.
Markowitz, H., 1952. Portfolio selection. J. Finance 7 (1), 77–91.
Meinshausen, N., Yu, B., 2009. Lasso-type recovery of sparse representations for high-dimensional data. Ann. Statist. 37 (1), 246–270.
Muirhead, R., 1982. Aspects of Multivariate Statistical Theory. Wiley, New York.
Pourahmadi, M., 1999. Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. Biometrika 86 (3), 677–690.
Pourahmadi, M., 2000. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. Biometrika 87 (2), 425–435.
Ren, Z., Kang, Y., Fan, Y., Lv, J., 2019. Tuning-free heterogeneous inference in massive networks. J. Amer. Statist. Assoc. 1–34.
Ren, Z., Sun, T., Zhang, C.-H., Zhou, H.H., 2015. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. Ann. Statist. 43 (3), 991–1026.
Rothman, A., Bickel, P., Levina, E., Zhu, J., 2008. Sparse permutation invariant covariance estimation. Electron. J. Stat. 2, 494–515.
Rothman, A.J., Levina, E., Zhu, J., 2009. Generalized thresholding of large covariance matrices. J. Amer. Statist. Assoc. 104 (485), 177–186.
Rothman, A.J., Levina, E., Zhu, J., 2010. A new approach to Cholesky-based covariance regularization in high dimensions. Biometrika 97 (3), 539–550.
Sharpe, W.F., 1966. Mutual fund performance. J. Bus. 39 (1), 119–138.
Sharpe, W.F., 1994. The sharpe ratio. J. Portf. Manag. 21 (1), 49–58.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.
Vaart, A.W.v.d., Wellner, J.A., 1996. Weak Convergence and Empirical Processes: With Application to Statistics. Springer, New York.
Yuan, M., 2010. High dimensional inverse covariance matrix estimation via linear programming. J. Mach. Learn. Res. 11, 2261–2286.
Yuan, M., Lin, Y., 2007. Model selection and estimation in the Gaussian graphical model. Biometrika 94 (1), 19–35.
Zhang, C.-H., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. 38 (2), 894–942.