UNIVERSITY OF CAMBRIDGE
Faculty of Economics

JANEWAY INSTITUTE CAMBRIDGE

CAMBRIDGE WORKING PAPERS IN ECONOMICS
JANEWAY INSTITUTE WORKING PAPERS

Should We Augment Large Covariance Matrix Estimation with Auxiliary Network Information?

Shuyi Ge
University of Nankai

Shaoran Li
Peking University

Oliver Linton
University of Cambridge

Weiguang Liu
University of Cambridge

Wen Su
University of Oxford

Abstract

In this paper, we propose two novel frameworks to incorporate auxiliary information about connectivity among entities (i.e., network information) into the estimation of large covariance matrices. The current literature either completely ignores this kind of network information (e.g., thresholding and shrinkage) or utilizes some simple network structure under very restrictive settings (e.g., banding). In the era of big data, we can easily get access to auxiliary information about the complex connectivity structure among entities. Depending on the features of the auxiliary network information at hand and the structure of the covariance matrix, we provide two different frameworks correspondingly —the Network Guided Thresholding and the Network Guided Banding. We show that both Network Guided estimators have optimal convergence rates over a larger class of sparse covariance matrix. Simulation studies demonstrate that they generally outperform other pure statistical methods, especially when the true covariance matrix is sparse, and the auxiliary network contains genuine information. Empirically, we apply our method to the estimation of the covariance matrix with the help of many financial linkage data of asset returns to attain the global minimum variance (GMV) portfolio.

Reference Details

2427 Cambridge Working Papers in Economics
2416 Janeway Institute Working Paper Series

Published 20 May 2024

Keywords Big Data, Network, Large Covariance Matrix, Thresholding, Banding
JEL-codes C13, C58, G11

Websites www.econ.cam.ac.uk/cwpe
www.janeway.econ.cam.ac.uk/working-papers

Should We Augment Large Covariance Matrix Estimation with Auxiliary Network Information? *

Shuyi Ge†, Shaoran Li‡, Oliver Linton§, Weiguang Liu¶, and Wen Su‖

May 20, 2024

Abstract

In this paper, we propose two novel frameworks to incorporate auxiliary information about connectivity among entities (i.e., network information) into the estimation of large covariance matrices. The current literature either completely ignores this kind of network information (e.g., thresholding and shrinkage) or utilizes some simple network structure under very restrictive settings (e.g., banding). In the era of big data, we can easily get access to auxiliary information about the complex connectivity structure among entities. Depending on the features of the auxiliary network information at hand and the structure of the covariance matrix, we provide two different frameworks correspondingly —the

*The authors would like to thank Xiaohong Chen, Hashem Pesaran, Cheng Hsiao, Harrison Zhou, Seok Young Hong, Jeroen Dalderop for their useful comments.

†School of Finance, University of Nankai. Author email:sg751@cam.ac.uk

‡School of Economics, Peking University. Author email:sl736@cam.ac.uk

§Faculty of Economics, University of Cambridge. Author email:obl20@cam.ac.uk

¶Faculty of Economics, University of Cambridge. Author email:wl342@cam.ac.uk

‖Mathematical Institude, University of Oxford. Author email: wen.su@maths.ox.ac.uk

Network Guided Thresholding and the Network Guided Banding. We show that both Network Guided estimators have optimal convergence rates over a larger class of sparse covariance matrix. Simulation studies demonstrate that they generally outperform other pure statistical methods, especially when the true covariance matrix is sparse, and the auxiliary network contains genuine information. Empirically, we apply our method to the estimation of the covariance matrix with the help of many financial linkage data of asset returns to attain the global minimum variance (GMV) portfolio.

Keywords: Big data; network; large covariance matrix; thresholding; banding.

JEL Classification: C13, C58, G11

1 Introduction

Covariance matrix estimation is an important problem in statistics and econometrics. When the dimensionality N of the random vector $\boldsymbol{X}_t = (X_{1t}, \ldots, X_{Nt})^\intercal$ under inspection is large, estimating its covariance matrix is challenging. It is well known that the sample covariance matrix is ill-conditioned when the dimension exceeds the sample size. In that case, some structures need to be imposed on the covariance matrix, and regularization techniques need to be applied for consistent estimation.

In the era of big data, we are gaining access to more and more auxiliary information in addition to the observations of $\{\boldsymbol{X}_t\}_{t=1}^T$, which could potentially help us learn about the underlying structure of the covariance matrix (i.e., connectivity among entities). Consider the case of equity return covariance. Israelsen (2016) finds that stocks covered by similar sets of analysts co-move a lot. Ge et al. (2022) find that stocks co-mentioned in business news exhibit excess co-movement beyond risk factors. Applying textual analysis to firms' 10-K reports, Hoberg and Phillips (2016) define peer groups within which firms are fundamentally similar. All of these aforementioned auxiliary network information could help us to learn about the connectivity among the stocks. However, the current literature either completely ignores this kind of auxiliary information or uses part of it under some very restrictive settings.

In this paper, we incorporate auxiliary information about connectivity among entities (i.e., network information) into the estimation of large covariance matrices. Depending on the features of the auxiliary network information at hand and the structure of the covariance matrix, we provide two separate avenues for application and derive their theories accordingly.

The first method we propose is called Network Guided Thresholding. The method is applicable when auxiliary information identifies the location of "significant" elements in the covariance matrix while staying silent about the relative importance of neighbors for each node. Industry information is a good example of such auxiliary information as it implies a block-diagonal net-

work where every node is equal within an industry. The original series of thresholding methods (Bickel and Levina (2008a), Cai and Liu (2011), Fan et al. (2013)) keep the "significant" elements in sample covariance and shrink the rest based on statistical information only under the assumption of sparsity (or conditional sparsity). These thresholding estimators require no knowledge about the location information. We, on the other hand, use auxiliary network information to identify the location of these "significant" elements. We keep the "significant" elements in the sample covariance and then apply generalized thresholding to the rest. The work closest to this method is Fan et al. (2016), where the authors apply location-based thresholding utilizing sector information. However, the factor model residual correlation structure is not as simple as a block diagonal assumed by Fan et al. (2016), and our method accommodates more complex structures. We derive the theoretical properties of the Network Guided Thresholding estimator. Compared with Bickel and Levina (2008a), we consider a larger class of sparse covariance matrices as we distinguish "large" and "small" elements using the auxiliary information and we quantify their behaviors separately. We show the consistency of the estimator in the operator norm as $(\log N)/T \to 0$ uniformly over the class of matrices that satisfy our sparsity condition. Next, we show that the Network Guided Thresholding estimator achieves optimal rate as in Bickel and Levina (2008a) over a larger parameter space.

The second method we propose is called Network Guided Banding. Bickel and Levina (2008b) show that uniformly over the class of the "approximately bandable" matrices, the banding estimator shows a superior convergence rate. However, according to their definition, the elements become smaller in magnitude as one moves away from the diagonal. This definition is appropriate for applications with natural orderings of variables, such as time series, climatology, and spectroscopy. Unfortunately, in most cases, such orderings do not exist, which means that the banding estimator cannot be applied. In this paper, we propose a theoretical framework that expands the class of "bandable" matrices, making this method applicable to a broader range

of scenarios. One of the key features of this new method is that it is permutation-invariant, while the original banding estimator performs poorly on a permutated ordering. Unlike the first method, we require auxiliary information to reveal the relative importance of neighbors for each node to make this method applicable. For example, analyst co-coverage (Israelsen (2016)), news co-mentioning (Ge et al. (2022)), and text-based product similarity (Hoberg and Phillips (2016)) all provide degrees of similarities among entities, according to which we could rank the relative importance of neighbors for each node. Taking news co-mentioning for illustration, firms co-mentioned by the same piece of news are treated as linked, and the frequency of co-mentioning could be used to measure the strength of linkages and thus rank their relative importance. We derive the theoretical properties of the Network Guided Banding estimator. We show the consistency of the estimator in the operator norm as $(\log N)/T \to 0$, uniformly over the class of matrices that satisfy our sparsity and "bandable" condition. We also show that the Network Guided Banding estimator achieves optimal rate as in Bickel and Levina (2008b) over a larger parameter space.

Practically, we apply proposed estimators to construct and test Global Minimum Variance (GMV) portfolios by estimating the covariance matrix of the idiosyncratic shocks by incorporating the auxiliary information.

We assume that asset returns obey a Fama-French factor framework. In Monte Carlo experiments, we compare our estimation accuracy with other benchmark competitors, including conventional thresholding, linear and nonlinear shrinkage approaches. The asset returns follow a factor model where the true covariance matrix of asset-specific shocks is sparse. For the Network Guided Thresholding approach, we introduce auxiliary information of varying quality, simulating both type I (false positives) and type II (false negatives) errors. The findings reveal superior finite sample performance over all compared statistical methods, provided the auxiliary network information is of reasonable quality, with minimal type I and II errors. As for the

Network Guided Banding, the generated neighborhood information can reveal the true network structure with different probabilities. When the probability of accurately revealing the network structure is not negligible, the performance of our estimator surpasses that of all benchmark models. Finally, the relative performance of the two proposed estimators depends on the genuine structure of the underlying covariance matrix and the quality of the auxiliary information.

Our empirical analysis utilizes real-world data from the Chinese stock market, adopting the Chinese factor model as outlined in Liu et al. (2019) to analyze asset returns. We explore various sources of auxiliary information that reveal linkages among listed stocks. A primary focus is the news co-mention network, as elaborated in Ge et al. (2023), which delineates two specific types of linkages: co-mentions within the same passage and within the same sentence. This categorization not only enriches the connectivity data but also serves as a critical input for both our Network Guided Thresholding and Banding estimators, given its ability to quantify the frequency of co-mentions. Furthermore, we explore the analyst co-coverage in China, which plays a pivotal role in understanding the interconnectedness of stocks within this unique market. This auxiliary information quantifies the strength of connectivity through a continuous measure, making it highly applicable to our proposed estimators. Additionally, we incorporate the traditional industry classification as an indicator of connectivity, particularly aligning with the Network Guided Thresholding approach. This established method of identifying linkages serves as a comparative basis for the newer forms of auxiliary information. To assess the practical value of incorporating this auxiliary information, we construct Global Minimum Variance (GMV) portfolios both with and without these additional data sources. These portfolios are then tested for their out-of-sample performance, using conventional statistical methods that lack auxiliary connection information as benchmark models. Our comparison spans different sets of constituent stocks, ranging from 300 to 800. Our findings consistently indicate that the integration of auxiliary information significantly enhances the out-of-sample performance of

GMV portfolios. This improvement underscores the practical benefits of incorporating network linkages and connectivity evidence, derived from both novel and traditional sources, in portfolio construction.

Literature Review: A growing number of works have been proposed in the literature to study the covariance matrix estimation when the dimensionality is large. Bickel and Levina (2008a) develop the theory for universal thresholding, which assumes the diagonal of the covariance matrix is uniformly bounded. Cai and Liu (2011) relax the uniform boundedness assumption and proposes an adaptive thresholding estimator where there are entry-adaptive thresholds. Fan et al. (2013) argue that common factors should be extracted first before applying thresholding when there are "extremely spiked" eigenvalues in the covariance and such a covariance matrix is conditionally sparse. Another strand of literature has tried to correct the spectrum of the sample covariance matrix instead of imposing sparsity on the elements of the matrix. For instance, Ledoit and Wolf (2004) and Ledoit et al. (2012) have proposed linear and nonlinear shrinkage estimators that apply shrinkage to the eigenvalues of the sample covariance matrix. The linear shrinkage does this by finding the linear combination of the sample covariance and a well-conditioned matrix, such as the identity matrix. The nonlinear shrinkage estimator corrects the eigenvalues using the asymptotic Marchenko–Pastur distribution. One of the advantages of shrinkage estimators is that they are well-conditioned, while the estimators based on sparsity often require choosing tuning parameters to guarantee positive definiteness. However, shrinkage estimators may be undesirable when the true covariance matrix is sparse. These aforementioned methods completely ignore the location information implied by auxiliary information that might be out there and rely on observations of $\{\boldsymbol{X}_t\}_{t=1}^T$ only. There is also literature that embraces the usage of very simple location information. Bickel and Levina (2008b) proposes the banding and tapering estimators, where indexes have orderings and elements in the covariance matrix become smaller in magnitude as one moves away from the

diagonal. They show that the banding estimator has a superior convergence rate by utilizing the location information. However, the underlying structure of these "bandable" matrices is very restrictive, which leaves the banding estimator inapplicable in most scenarios.

The novelty of this paper is that we augment the estimation of large covariance matrices with auxiliary network information. Depending on the features of the auxiliary network information at hand, we provide two separate avenues for application. We derive their theories accordingly, and we show that both Network Guided estimators have good theoretical and numerical properties.

Although in this paper, we are applying augmenting network information to the estimation of large static covariance matrices, a similar idea can be extended to the estimation of large dynamic covariance matrices. For example, dynamic network information could be well incorporated into the conditioning information set in Chen et al. (2019).

The remainder of this paper is structured as follows. Section 2 introduces the concepts behind the Network Guided Thresholding estimator and the Network Guided Banding estimator, and provide a comparison against conventional thresholding and banding approaches. In Section 3, we lay down the foundational assumptions and deduce the convergence theorems within the factor model; moreover, this section encompasses a discussion on the adjustments for ensuring positive definiteness. Section 4 is dedicated to presenting our simulation results, comparing the performance of our proposed estimators with other established baseline methodologies. In Section 5, we empirically apply these methodologies to predict the covariance matrix for stock returns in the China market, and evaluate the out-of-sample volatility of the Global Minimum Variance (GMV) portfolio as produced by our estimators as well as other benchmarks. Finally, Section 6 provides a summary of our findings and briefly deliberates on avenues for future research.

Notation: For vector $\boldsymbol{a} \in \mathbb{R}^d$, $\|\boldsymbol{a}\|$ stands for Euclidean norm, i.e., $\|\boldsymbol{a}\| = (a_1^2 + \cdots, a_d^2)^{1/2}$. For matrix $A = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m) \in \mathbb{R}^{m \times d}$, $\|A\|_F$ represents the matrix Frobenius norm, i.e., $\|A\|_F = (\|\boldsymbol{a}_1\|^2 + \cdots + \|\boldsymbol{a}_m\|^2)^{1/2}$; $\|A\|_2$ stands for the matrix 2-norm, which is defined as $\|A\|_2 = \sqrt{\rho_{\max}(A^\top A)}$, where $\rho_{\max}(\cdot)$ returns the maximum eigenvalue of a matrix; the operator norm $\|A\| = \inf\left\{c > 0 : \|Ax\| \leq c\|x\|, \text{ for all } x \in \mathbb{R}^d\right\}$ is typically used in the theoretical proof. For two real-valued sequences $\{a_T\}$ and $\{b_T\}$, $a_T = o(b_T)$ means $a_T/b_T \to 0$ when $T \to \infty$; $a_T = O(b_T)$ means there exists some constant A, s.t. $a_T \leq Ab_T$ for all T. We use $\boldsymbol{J}_{N\times N}$ to represent $N \times N$ unit matrix.

2 Model Setup

Suppose we have observations $\boldsymbol{X}_t = (X_{1t}, \ldots, X_{Nt})^\intercal$[1], $t = 1, \ldots, T$ of a N-dimensional random vector $\boldsymbol{X}$ with mean $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and variance $E((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\intercal) = \boldsymbol{\Sigma}$. The sample covariance estimator is given as follows:

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})^\intercal = [\widehat{\sigma}_{ij}]_{N\times N}, \tag{1}$$

where $\bar{\boldsymbol{X}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{X}_t$. As mentioned in the introduction section, the sample covariance matrix behaves poorly when N is large. Below we propose two theoretical frameworks for augmenting large covariance matrix estimation with auxiliary network information. One may choose the suitable framework depending on the features of the auxiliary network information at hand.

2.1 Network Guided Thresholding

When our auxiliary information identifies the location of "significant" elements in the covariance matrix while staying silent about the relative importance of neighbors for each node, we go for the Network Guided Thresholding method. Recall that in the original thresholding paper

[1]The samples can either be independent or strong α-mixing in our assumptions.

(Bickel and Levina (2008a)), their uniformity class of covariance matrices is given by

$$\mathcal{U}_\tau(q, c_0, M) = \left\{ \mathbf{\Sigma} : \sigma_{ii} \leq M, \sum_{j=1}^{N} |\sigma_{ij}|^q \leq c_0(N), \text{ for all } i \right\}. \tag{2}$$

Here, q plays an important role. Suppose that $q = 0$, then the number of non-zero elements needs to be bounded. On the other hand, when $q \to 1$, the large elements of Σ will dominate, and thus the sum of large elements should be bounded.

In this paper, we consider an extension to their uniformity class. We first define the Location Indicator Matrix

$$L = [L_{ij}]_{N\times N}, \quad L_{ij} = I_{\{|r_{ij}|>l\}} = \begin{cases} 1, & |r_{ij}| > l, \\ 0, & |r_{ij}| \leq l, \end{cases} \tag{3}$$

note this L matrix is defined over the correlation coefficients matrix $R = [r_{ij}]_{N\times N}$, and for $s \in \{0, 1\}$, we define $L^s_{ij} = I_{\{L_{ij}=s\}}$. Thus, L^1 indicates the location of large elements in the correlation matrix, and L^0 indicates the location of small elements (not necessarily zero) in the correlation matrix. It is obvious that $L^1 = L$ and $L^0 = I_{N,N} - L^1$, where $I_{N,N}$ is a unit matrix. We use auxiliary information to estimate L and therefore, L^s for $s \in \{0, 1\}$. We then consider the following uniformity class:

$$\mathcal{U}_1(q, c_0, c_1, M, L) = \left\{ \mathbf{\Sigma} = DRD : \sigma_{ii} \leq M, \sum_j L^1_{ij} \leq c_1(N), \sum_j L^0_{ij} |r_{ij}|^q \leq c_0(N), \text{ for all } i \right\}, \tag{4}$$

where $D = \text{diag}\{\sqrt{\sigma_{11}}, \cdots, \sqrt{\sigma_{NN}}\}$, and we separately state the conditions for large elements ((i, j) pairs such that $L^1_{ij} = 1$) and small elements ((i, j) pairs such that $L^0_{ij} = 1$). Essentially, this uniformity class controls the number of large elements and the growth rate of the remaining small elements. Compared to the uniformity class of covariance matrices considered in Bickel and Levina (2008a), we extend the class of covariance matrices that satisfy the thresholding condition. Consider a covariance matrix $\mathbf{\Sigma}$ that contains a small number of relatively large elements and a large number of small elements. Such a covariance matrix does not satisfy the sparsity assumption from Bickel and Levina (2008a) while it still satisfies our sparsity condition. Thus our method can deal with more scenarios.

Of course, a priori we do not know the location of the large elements. Suppose that we have observations from the auxiliary dataset that allow us to form an estimator $\widehat{L}$ for L, independent of the sample $\boldsymbol{X}$. With the help of $\widehat{L}$, we define the Network Guided Thresholding Estimator to be

$$T_{L,\lambda}\left(\widehat{R}\right) = \left[s_{L,\lambda}\left(\widehat{\sigma}_{ij}/\sqrt{\widehat{\sigma}_{ii}\widehat{\sigma}_{jj}}\right)\right]_{N\times N} \quad \text{with} \quad s_{L,\lambda}\left(r_{ij}\right) = r_{ij}I_{\{L_{ij}=1\}} + s_{\lambda}\left(r_{ij}\right)I_{\{L_{ij}=0\}}, \tag{5}$$

where $s_\lambda(x)$ is the generalized thresholding operator[2]. Thus for covariance matrix, we naturally define $\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_{\widehat{L}} := \widehat{D}T_{L,\lambda}\left(\widehat{R}\right)\widehat{D}$. Then the feasible Network Guided Thresholding Estimator is $\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_{\widehat{L}} := \widehat{D}T_{\widehat{L},\lambda}\left(\widehat{R}\right)\widehat{D}$, where we use the estimated Location Indication Matrix $\widehat{L}$.

2.2 Network Guided Banding

When the auxiliary information reveals the relative importance of neighbors for each node, we prefer the Network Guided Banding method. Recall that the original Banding and Tapering methods work when there is a natural "order" or "distance" among variables; they consider the following uniformity class of covariance matrices:

$$\mathcal{U}_b(\varepsilon,\alpha,c) = \left\{\boldsymbol{\Sigma} : \max_j \sum_{i:|i-j|>k} |\sigma_{ij}| \le ck^{-\alpha} \text{ for all } k, \text{ and } 0 < \varepsilon \le \rho_{\min}\left(\boldsymbol{\Sigma}\right) \le \rho_{\max}\left(\boldsymbol{\Sigma}\right) \le \frac{1}{\varepsilon}\right\}, \tag{6}$$

where $\rho_{\min}(\cdot)$ and $\rho_{\max}(\cdot)$ give the minimal and maximal eigenvalues of a matrix. Bickel and Levina (2008b) shows that when this banding condition is satisfied, a better convergence rate can be achieved by taking advantage of the underlying structure.

However, the original Banding and Tapering methods are only applicable to time series essentially, as variables are not ordered in most cases. We extend their method by allowing a more general underlying connectivity (network) structure, making these methods applicable to a wider range of covariance matrices. We first define a new order $\langle\{1,\cdots,N\},\succ\rangle$ for a N-dimensional vector $\boldsymbol{a} = (a_1,\ldots,a_N)^\intercal$ with distinct elements as follows:

$$i \succ j \;\Leftrightarrow\; a_i > a_j.$$

[2]Commonly used thresholding operators such as hard thresholding, soft thresholding, and SCAD can be applied.

Given a vector of relative importance $\boldsymbol{a} = (a_1, \ldots, a_N)^\intercal$, we can use this order operator to sort the elements from the vector. Then we use a descending (in terms of $\succ$) tuple $(p_1, \ldots, p_N)$ to record the sorted result, where $p_1 \succ p_2 \succ \cdots \succ p_N$. Notice that $(p_1, \ldots, p_N)$ is a permutation of $(1, \ldots, N)$, where p_1 gives the index of the largest element (the most important) and p_N gives the index of the smallest element (the least important). For any positive integer k, define $S_k^a = \{p_1, ..., p_k\}$ as the set of indexes of the k-biggest elements under $\succ$ for vector $\boldsymbol{a}$. For example, if $\boldsymbol{a} = (1, 4, 3, 2)$, then the sorted tuple is $(2, 3, 4, 1)$, $S_2^a = \{2, 3\}$. Next, we generalize the uniformity class considered in Bickel and Levina (2008b) (Equation 6) by directly comparing the relative magnitudes (not a real "distance") of entries for each row of a matrix. We modify the correlation counterpart of Equation 6 instead of itself for fair comparison under heteroskedasticity. To be precise, we consider a generalized uniformity class of covariance matrices:

$$\mathcal{U}_2(\varepsilon, \alpha, b_0, M) = \left\{ \boldsymbol{\Sigma} = DRD : \max_i \sigma_{ii} < M, \sum_{j \notin S_k^{\text{abs}(r_i)}} |r_{ij}| < b_0 (N) k^{-\alpha} \text{ for all } i, k, \text{ and } \rho_{\max}(R) \leq \frac{1}{\varepsilon} \right\}, \quad (7)$$

where r_i is the i-th column (row) of R, and $\text{abs}(r_i) = (|r_{i1}|, \cdots, |r_{iN}|)$ gives the absolute values of the correlation coefficients. $S_k^{\text{abs}(r_i)}$ gives the set of indexes of the k-biggest elements. Notice that when $k = 1$, $S_k^{\text{abs}(r_i)}$ as the self-correlation is always the largest. When $k > 1$, $S_k^{\text{abs}(r_i)}$ includes i itself and the set of $k - 1$ nearest neighbours. Essentially, the correlations between non-neighboring pairs need to be small under Equation 7. Compared with the original banding, this method is permutation-invariant and accommodates a more general connectivity (network) structure.

We use auxiliary information to infer the underlying connectivity structure of the target correlation matrix. Define a relative importance indicator matrix $C = [C_{ij}]_{N \times N}$ with non-negative elements. For each row i (or column), the elements of $C_i = (C_{i1}, \ldots, C_{iN})$ give the relative importance scores and retain the order of importance from $\text{abs}(r_i)$ (i.e., for each i, there exists a non-decreasing function f_i such that $C_{ij} = f_i(|r_{ij}|)$ for all j). Then we can reconstruct

the correlation structure with a Network Guided Banding Estimator as follows

$$\widehat{\mathbf{\Sigma}}_C^{\mathcal{B}} = \widehat{D} B_{C,k}\left(\widehat{R}\right)\widehat{D} \quad \text{with} \quad B_{C,k}\left(\widehat{R}\right) = \left[b_{C,k}\left(\widehat{r}_{ij}\right)\right]_{N\times N},$$
$$b_{C,k}\left(r_{ij}\right) = r_{ij} I_{\left\{i\in S_k^{c_j}, j\in S_k^{c_i}\right\}} = \begin{cases} r_{ij}, & i \in S_k^{c_j} \text{ and } j \in S_k^{c_i}, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

We do not observe the relative importance indicator matrix C, and we use the auxiliary dataset to form an estimator $\widehat{C}$, and the feasible estimator is $\widehat{\mathbf{\Sigma}}_{\widehat{C}}^{\mathcal{B}}$.

It's noteworthy that $\widehat{\mathbf{\Sigma}}_{\widehat{C}}^{\mathcal{B}}$ is not strictly a banding or tapering estimator because the k-neighbour relationship is asymmetric, i.e., $i \in S_k^{c_j} \not\Leftrightarrow j \in S_k^{c_i}$ for certain symmetric matrix C. For example, in a scale-free network, the central node is the neighbor of many nodes connected to it, but the reverse is not true.

3 Main Results

3.1 Factor Structure and Global Minimum Variance (GMV) Portfolio

Consider the following classical factor model

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{f}_{1,t} + \boldsymbol{\beta}_2 \boldsymbol{f}_{2,t} + \cdots + \boldsymbol{\beta}_K \boldsymbol{f}_{K,t} + \boldsymbol{u}_t \\ &= \boldsymbol{\beta}_0 + \boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{u}_t, \end{aligned} \tag{9}$$

$t = 1, 2, \cdots, T$, where $\boldsymbol{y}_t$ is the $N \times 1$ assets return at time t, $\boldsymbol{f}_t$ is the $K \times 1$ observable factors return, $\boldsymbol{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_K)$ is the $N \times K$ factor loading matrix, the mispricing term $\boldsymbol{\beta}_0$ is taken out of $\boldsymbol{B}$ since it has different economic meaning, and $\boldsymbol{u}_t$ is the zero-mean residual term, which may contain cross-section dependency but is uncorrelated with $\boldsymbol{f}_t$. Our goal is to construct a Global Minimum Variance (GMV) portfolio by taking account of the covariance matrix of $\boldsymbol{u}_t$, denoted by $\mathbf{\Sigma} = E\left(\boldsymbol{u}_t \boldsymbol{u}_t^{\intercal}\right) = \left(\sigma_{ij}\right)_{N\times N}$, with the help of auxiliary network information, especially in the case $T < N$.

It is notoriously hard to estimate both the first and second moments of asset returns through past observations, and this motivates us to take the GMV portfolio as a playground to test our methodology. Compared with the mean-variance optimal portfolio as in Markowitz et al. (1952), GMV portfolio avoids the estimation error of the expectation of assets returns, which can mostly reflect the performance of covariance matrix estimators.

Under mild assumptions, the portfolio weights for each asset to construct a GMV portfolio are:

$$\boldsymbol{\omega}^{\text{GMV}} = \frac{\boldsymbol{\Sigma}_y^{-1}\mathbf{1}}{\mathbf{1}^\intercal\boldsymbol{\Sigma}_y^{-1}\mathbf{1}},$$

where $\boldsymbol{\omega}$ is $N \times 1$ vector of portfolio weights, with $\mathbf{1}$ a conforming vector of ones and $\boldsymbol{\Sigma}$ the covariance matrix of assets returns $\boldsymbol{y}_t$. Given the factor structure of assets returns in Equation 9, we have $E(\boldsymbol{y}_t\boldsymbol{y}_t^\intercal) = \boldsymbol{\Sigma}_y = \boldsymbol{B}\boldsymbol{\Sigma}_f\boldsymbol{B}^\intercal + \boldsymbol{\Sigma}_u$. Next, we discuss how to estimate $\boldsymbol{\omega}$ under our framework.

The ordinary least square method gives estimators of $\boldsymbol{\beta}_0$ and $\boldsymbol{B}$, thus we can collect residuals as $\widehat{\boldsymbol{u}}_t = \boldsymbol{y}_t - \widehat{\boldsymbol{\beta}}_0 - \widehat{\boldsymbol{B}}\boldsymbol{f}_t$. The conventional estimator of $\boldsymbol{\Sigma}_u$ is $\widehat{\boldsymbol{\Sigma}}_u = \frac{1}{T}\sum_{t=1}^T \widehat{\boldsymbol{u}}_t\widehat{\boldsymbol{u}}_t^\intercal = (\widehat{\sigma}_{ij})_{N\times N}$. After applying our thresholding or banding technique with auxiliary information to $\widehat{\boldsymbol{\Sigma}}_u$, we attain $\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_{u,\widehat{L}}$ and $\widehat{\boldsymbol{\Sigma}}^{u\mathcal{B}}_{u,\widehat{C}}$ separately. Finally,

$$\widehat{\boldsymbol{\omega}} = \frac{\widehat{\boldsymbol{\Sigma}}_y^{-1}\mathbf{1}}{\mathbf{1}^\intercal\widehat{\boldsymbol{\Sigma}}_y^{-1}\mathbf{1}},$$

with $\widehat{\boldsymbol{\Sigma}}_y = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^\intercal + \widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_{u,\widehat{L}}$ or $\widehat{\boldsymbol{\Sigma}}^{\mathcal{B}}_{u,\widehat{C}}$.

Next, we discuss the assumptions and the corresponding theoretical properties of $\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_{u,\widehat{L}}$ and $\widehat{\boldsymbol{\Sigma}}^{\mathcal{B}}_{u,\widehat{C}}$. In our analysis, both N and T can go to infinity, and N can be larger than T, but we restrict $\frac{\log N}{T} \to 0$. Proofs of all theorems are deferred to the appendix. For simplicity, we may abuse the notation A in the future to represent any large enough constant which does not depend on N and T.

Assumption 1. *(a) Sequence $\{\boldsymbol{u}_t\}$ is strong α-mixing stationary and ergodic, with zero means and covariance matrix $\boldsymbol{\Sigma}$, the mixing coefficients $\left\{\alpha_t^{\text{mixing}}, t \geq 0\right\}$ satisfy $\alpha_t^{\text{mixing}} \leq \exp\left(-\phi_1 t^{\phi_2}\right)$*

for some positive constants ϕ_1 *and* ϕ_2 *not depending on* N *(thus uniformly mixing over* N*), and there are some constants* $\underline{c}, \overline{c}$*, s.t.,* $0 < \underline{c} < \inf_{i,j} Var(u_{it}u_{jt}) < \sup_{i,j} Var(u_{it}u_{jt}) < \overline{c}$*,* $\underline{c} < \rho_{\min}(\mathbf{\Sigma}) < \rho_{\max}(\mathbf{\Sigma}) < \overline{c}$*.*

(b) The tail of the distribution of u_{it} *is uniformly bounded by an exponential-type tail, i.e., for some constant* $\phi_3, \phi_4 > 0$ *not depending on* N*, and any* $x > 0$*, we have* $\sup_i P(|u_{it}| > x) \leq \exp\{-\phi_3 x^{\phi_4}\}$*.*

(c) For some positive sequences $\kappa_1(N,T) = o(1)$ *and* $a_T = o(1)$*, and a constant* A *which does not depend on* N *and* T*,* $P\left(\max_i \frac{1}{T}\sum_{t=1}^{T} |u_{it} - \widehat{u}_{it}|^2 > Aa_T^2\right) \leq O(\kappa_1(N,T))$ *and* $P(\max_{i,t} |u_{it} - \widehat{u}_{it}| > A) = o(1)$*.*

(d) For some $\gamma < 1$*,* $(\log N)^{6/\gamma - 1} = o(T)$*.*

Remark: Condition (a) is common in the econometric research. The first part suggests a weak dependency for the sequence while the second part requires $\mathbf{\Sigma}$ invertible. The tail condition (b) allows large deviation theory to be applied. Condition (c) is important to allow us to apply estimation constructed by $\widehat{\boldsymbol{u}}_t$ when the true values are not observable. In addition, conditions (a), (b) and (c) match the Assumptions 2.1, 2.2 in Fan et al. (2011). Condition (d) is an additional assumption to assure good asymptotic properties, which is proposed in Theorem 2.1 of Fan et al. (2011).

For example, with these assumptions, one can easily show

$$P\left(\max_{i,j} |\widehat{\sigma}_{ij} - \sigma_{ij}| > A\sqrt{\frac{\log N}{T}}\right) = O\left(\frac{1}{N^2}\right)$$

for some large A which does not depend on N and T. The proof can be found in Lemma A.3 of Fan et al. (2011).

Remark: In this paper, we suppose the observed price or observed return is equal to the efficient price or efficient return. However, when the observed price P_t is the sum of efficient price P_t^* and microstructure noise e_t, i.e., $P_t = P_t^* + e_t$, as highlighted by Li and Linton (2022), the microstructure noise component is not directly observed because it is obscured by the

efficient price. In that case, the covariance matrix of the efficient price series is equal to the long run covariance matrix of the observed returns.

3.2 Consistency of Thresholding Estimator

The thresholding estimator in our paper is mainly set for the correlation coefficients matrix, i.e., the indicator matrix $L = (L_{ij})_{N\times N}$ with $L_{ij} = I_{\{|r_{ij}|>l\}}$ for some given l (may change with (N,T)), and $\mathbf{\Sigma} = DRD$ where $D = \text{diag}\left\{\sqrt{\sigma_{11}}, \cdots, \sqrt{\sigma_{NN}}\right\}$. We assume that $\mathbf{\Sigma}$ lies in the class $\mathcal{U}_1(q, c_0, c_1, M, L)$ introduced in Equation 4, and the thresholding estimator is given by $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{\widehat{L}} = \widehat{D}T_{\widehat{L},\lambda}\left(\widehat{R}\right)\widehat{D}$. We emphasize that in Fan et al. (2011), conditions and assumptions are mainly designed for those large elements since the rest elements in their paper are directly zeros. In our paper, additional assumptions on the small elements and network information need to be imposed to derive the convergence theory.

Assumption 2. *(a)* $P\left(\max_{1\le i\le N}\sum_{j=1}^{N} I_{\left\{L_{ij}=1,\widehat{L}_{ij}=0\right\}} > a_T c_1(N)\right) \le O(\kappa_2(N))$, *for some* $\kappa_2(N,T) = o(1)$, $c_1(N)\to\infty$;

(b) The function s_λ *satisfies* $|s_\lambda(t) - t| \le t$ *and* $|s_\lambda(t)| \le \lambda$ *for* $|t| \le \lambda$;

(c) We assume

$$P\left(\max_{1\le i\le N}\sum_{j=1}^{N}\left|\frac{L^0_{ij}}{T}\sum_{t=1}^{T}u_{it}u_{jt}\right| > \max_{1\le i\le N}\sum_{j=1}^{N}L^0_{ij}|\sigma_{ij}| + a_T\right) \le O(\kappa_3(N,T))$$

for some $\kappa_3(N,T) = o(1)$;

(d)$P\left(\max_{1\le i\le N}\sum_{j=1}^{N}\left|\frac{L^0_{ij}}{T}\sum_{t=1}^{T}(\widehat{u}_{it}\widehat{u}_{jt} - u_{it}u_{jt})\right| > a_T\right) \le O(\kappa_3(N,T))$.

Remark: Condition (a) restricts the number of mis-classified large elements. Condition (b) is the condition (iii) in Rothman et al. (2009), which is a basic requirement in thresholding estimation. Condition (c) sets an upper bound for the speed of small elements' growth. Condition (d) argues the total error of the small elements between using $\boldsymbol{u}_t$ and $\widehat{\boldsymbol{u}}_t$ cannot be too large. Note that condition (c) and (d) together allow estimation error on small elements, with comparison to those on large elements in condition (a).

For the asymptotic properties of Network Guided Thresholding estimator, we have the following result.

Theorem 1. *Suppose that Assumption 1 and Assumption 2 hold, and* $l \leq \lambda$ *for all* (N,T)*. For some large* A *which does not depend on* (N,T)*, we have:*

$$P\left(\left\|\widehat{\mathbf{\Sigma}}_{\widehat{L}}^{\mathcal{T}}-\mathbf{\Sigma}\right\|>A\left(c_1(N)\sqrt{\frac{\log N}{T}}+c_0(N)\lambda^{1-q}+a_T\right)\right)=O\left(\frac{1}{N^2}+\kappa_1(N,T)+\kappa_2(N,T)+\kappa_3(N,T)\right),$$

where $\|\cdot\|$ *represents the operator norm, and the rate of convergence* κ_1 *is given in Assumption 1 to describe the relationship of* u_{it} *and* $\widehat{u}_{it}$*, while* κ_2 *and* κ_3 *are the convergence rates of* $\widehat{L}$*, introduced in Assumption 2.*

Remark: The error term caused by large elements estimation is $c_1(N)\sqrt{\frac{\log N}{T}}$, the effect of small elements appears in $c_0(N)\lambda^{1-q}$ and the error a_T is caused by using $\widehat{L}$ rather than true L. When $c_0(N)$ and $c_1(N)$ are both $O(1)$, the best choice of λ is $\lambda_N \asymp \left(\frac{\log N}{T}\right)^{1/2(1-q)}$, which then gives

$$\left\|\widehat{\mathbf{\Sigma}}_{\widehat{L}}^{\mathcal{T}}-\mathbf{\Sigma}\right\|=O_P\left(\sqrt{\frac{\log N}{T}}+a_T\right)=o_P(1).$$

If we assume $\frac{c_0(N)}{c_1(N)}=O\left(\left(\frac{\log N}{T}\right)^{q/2}\right)$, an optimal choice of thresholding parameter (also suggested in Rothman et al. (2009)) is $\lambda_N \asymp \sqrt{\frac{\log N}{T}}$, which yields

$$\left\|\widehat{\mathbf{\Sigma}}_{\widehat{L}}^{\mathcal{T}}-\mathbf{\Sigma}\right\|=O\left(c_0(N)\left(\frac{\log N}{T}\right)^{\frac{1-q}{2}}+a_T\right),$$

and provided $c_0(N)\left(\frac{\log N}{T}\right)^{\frac{1-q}{2}}=o(1)$, one obtains $\left\|\widehat{\mathbf{\Sigma}}_{\widehat{L}}^{\mathcal{T}}-\mathbf{\Sigma}\right\|=o_P(1)$.

3.3 Consistency of Banding Estimator

Recall our banding method on correlation coefficients matrix class as defined in Equation 7, and a network matrix C gives $S_k^{c_i}=S_k^{\mathrm{abs}(r_i)}$ for all i, we easily get $\widehat{\mathbf{\Sigma}}_{\widehat{C}}^{\mathcal{B}}=\widehat{D}B_{\widehat{C},k}\left(\widehat{R}\right)\widehat{D}$ via an estimated network information $\widehat{C}$. The following assumptions on C guarantee the consistency of our estimator.

Assumption 3. *(a) For R and C, there exists b_1, s.t. $\sum_{j=1}^{N} |r_{ij}| I_{\left\{i \notin S_k^{c_j}, j \in S_k^{c_i}\right\}} < b_1(N)$, for all i, k;*

(b) Suppose $\widehat{C}$ is the estimator for C, and there exists a sequence $\kappa_4(N,T) \to 0$ when $T \to \infty$, for some A which does not depend on (N,T),

$$P\left(\frac{1}{k}\sum_{j=1}^{N} I_{\left\{j \in S_k^{c_i}, j \notin S_k^{\widehat{c}_i}\right\}} > A\sqrt{\frac{\log N}{T}}\right) = O\left(\kappa_4(N,T)\right), \quad P\left(\frac{1}{k}\sum_{j=1}^{N} I_{\left\{i \in S_k^{c_j}, i \notin S_k^{\widehat{c}_j}\right\}} > A\sqrt{\frac{\log N}{T}}\right) = O\left(\kappa_4(N,T)\right).$$

Remark: Since C is not symmetric, $i \in S_k^{c_j}$ may not be equivalent to $j \in S_k^{c_i}$. The condition (a) requires the "asymmetry quantity" to be bounded by $b_1(N)$, alternatively speaking, most of the asymmetric parts of C correspond to the small elements. The condition (b) assumes the number of wrong estimates for large elements is bounded by $O\left(k\sqrt{\frac{\log N}{T}}\right)$.

For Network Guided Banding estimator, we have the following asymptotic theorem.

Theorem 2. *Suppose that Assumption 1, Assumption 3 hold and $k = k_N \to \infty$. Then,*

$$P\left(\left\|\widehat{\mathbf{\Sigma}}_{\widehat{C}}^{\mathcal{B}} - \mathbf{\Sigma}\right\| > A\left(k\sqrt{\frac{\log N}{T}} + b_0(N)\, k^{-\alpha} + b_1(N)\right)\right) = O\left(\frac{1}{N^2} + \kappa_1(N,T) + \kappa_4(N,T)\right),$$

for some constant A which does not depend on (N,T), where κ_1 is the rate introduced in Assumption 1, and κ_4 is the convergence rate of $\widehat{C}$ in Assumption 3.

Remark: In the error term, the first two parts $k\sqrt{\frac{\log N}{T}} + b_0(N)\, k^{-\alpha}$ are the same as Bickel and Levina (2008a), while $b_1(N)$ is "asymmetry quantity" introduced in Assumption 3. Additionally, the error caused by using $\widehat{C}$ to replace C is bounded by $O\left(\sqrt{\frac{\log N}{T}}\right)$ (details can be found in the proof of Theorem 2), thus absorbed into the first part in the result. Bickel and Levina (2008a) suggests an optimal choice of k, which is $k_N \asymp \left(\frac{\log N}{T}\right)^{-1/2(\alpha+1)}$, then we get

$$\left\|\widehat{\mathbf{\Sigma}}_{\widehat{C}}^{\mathcal{B}} - \mathbf{\Sigma}\right\| = O_P\left(\left(1 + b_0(N)\right)\left(\frac{\log N}{T}\right)^{\frac{\alpha}{2(\alpha+1)}} + b_1(N)\right). \tag{10}$$

If matrix C is symmetric, then our bound in Equation (10) turns to $O_P\left(\left(1 + b_0(N)\right)\left(\frac{\log N}{T}\right)^{\frac{\alpha}{2(\alpha+1)}}\right)$, which matches the bound in Bickel and Levina (2008a). Therefore, provided $b_0(N)\left(\frac{\log N}{T}\right)^{\frac{\alpha}{2(\alpha+1)}} = o(1)$ and $b_1(N) = o(1)$, one easily obtains $\left\|\widehat{\mathbf{\Sigma}}_{\widehat{C}}^{\mathcal{B}} - \mathbf{\Sigma}\right\| = o_P(1)$.

3.4 Positive Definiteness of $\widehat{\mathbf{\Sigma}}_y$

To ensure the positive definiteness of $\widehat{\mathbf{\Sigma}}_y$, we borrow the modification method of Chen et al. (2019). Namely, for a certain estimator $\widehat{\mathbf{\Sigma}}$ of a $N \times N$ positive definite population covariance matrix $\mathbf{\Sigma}$, let $\widehat{\rho}_1 \geq \widehat{\rho}_2 \geq \cdots \geq \widehat{\rho}_N$ be the eigenvalues of estimator $\widehat{\mathbf{\Sigma}}$. In the case $\widehat{\rho}_N \leq 0$, the estimator $\widehat{\mathbf{\Sigma}}$ is not positive definite, one can follow Chen and Leng (2016) to modify it by constructing

$$\widehat{\mathbf{\Sigma}}_{M_0} = \widehat{\mathbf{\Sigma}} + (m_T - \widehat{\rho}_N) \cdot \boldsymbol{J}_{N\times N}, \tag{11}$$

where $J_{N\times N}$ is the $N \times N$ identity matrix and $m_T > 0$ is a tuning parameter. Clearly, Equation 11 assures the smallest eigenvalues be positive thus $\widehat{\mathbf{\Sigma}}_{M_0}$ becomes invertible. Chen et al. (2019) augment Equation 11 by defining

$$\widehat{\mathbf{\Sigma}}_M = \widehat{\mathbf{\Sigma}} \cdot \mathbf{1}_{\{\widehat{\rho}_N > 0\}} + \widehat{\mathbf{\Sigma}}_{M_0} \cdot \mathbf{1}_{\{\widehat{\rho}_N \leq 0\}} = \widehat{\mathbf{\Sigma}} + (m_T - \widehat{\rho}_N) \cdot \boldsymbol{J}_{N\times N} \cdot \mathbf{1}_{\{\widehat{\rho}_N \leq 0\}}, \tag{12}$$

which indicates we still keep $\widehat{\mathbf{\Sigma}}$ when it is already positive definite while choose to use $\widehat{\mathbf{\Sigma}}_{M_0}$ when non-positive eigenvalues appear.

Now, we apply the Sherman-Morrison-Woodbury formula to $\widehat{\mathbf{\Sigma}}_y$ and get

$$\widehat{\mathbf{\Sigma}}_y^{-1} = \widehat{\mathbf{\Sigma}}_u^{-1} - \widehat{\mathbf{\Sigma}}_u^{-1}\widehat{\boldsymbol{B}}\left(\widehat{\mathbf{\Sigma}}_f^{-1} + \widehat{\boldsymbol{B}}^\intercal\widehat{\mathbf{\Sigma}}_u^{-1}\widehat{\boldsymbol{B}}\right)\widehat{\boldsymbol{B}}^\intercal\widehat{\mathbf{\Sigma}}_u^{-1},$$

where $\widehat{\mathbf{\Sigma}}_f$ is naturally to be invertible in a (finite) factor structure while $\widehat{\mathbf{\Sigma}}_u^{-1}$ may not be well-possessed. We modify $\widehat{\mathbf{\Sigma}}_u$ by Equation 12, however, since

$$\left\|\widehat{\mathbf{\Sigma}}_{uM} - \mathbf{\Sigma}_u\right\| \leq \left\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\right\| + (m_T - \widehat{\rho}_N) \leq O_P\left(\left\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\right\|\right) + m_T + \left|\widehat{\rho}_N\right|.$$

When $\widehat{\rho}_N \leq 0$, Weyl's inequality gives

$$\left|\widehat{\rho}_N\right| \leq \left|\widehat{\rho}_N - \rho_{\min}\left(\mathbf{\Sigma}_u\right)\right| \leq \left\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\right\|,$$

then we suddenly have $\left\|\widehat{\mathbf{\Sigma}}_{uM} - \mathbf{\Sigma}_u\right\| \leq O_P\left(\left\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\right\|\right) + m_T$, so the tuning parameter should be set to go to 0 faster than the rate of convergence of $\widehat{\mathbf{\Sigma}}_u$, thus the modified version $\widehat{\mathbf{\Sigma}}_{uM}$

converges to $\mathbf{\Sigma}_u$ with the same rate as $\widehat{\mathbf{\Sigma}}_u$. Specifically, m_T should go to 0 faster than

$$\left\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\right\| = \begin{cases} O_P\left(c_1(N)\left(\frac{\log N}{T}\right)^{\frac{1}{2}} + c_0(N)\lambda^{1-q} + a_T\right), & \text{for thresholding,} \\ O_P\left(k\sqrt{\frac{\log N}{T}} + b_0(N)k^{-\alpha} + b_1(N)\right), & \text{for banding.} \end{cases}$$

4 Simulation

4.1 True Covariance Matrix

Similar to Cai and Liu (2011), we consider two types of sparse covariance matrices in the simulations to investigate the numerical properties of our proposed estimators.

- **Model 1 (banded matrix with ordering)**: $\mathbf{\Sigma} = \text{diag}\{A_1, A_2\}$, where $A_1 = (a_{ij})_{\frac{N}{2}\times\frac{N}{2}}$, $a_{ij} = \left(1 - \frac{|i-j|}{10}\right)^+$, $A_2 = 4\boldsymbol{J}_{\frac{N}{2}\times\frac{N}{2}}$. Here $\mathbf{\Sigma}$ is a two-block diagonal matrix, A_1 is a bandable sparse covariance matrix, and A_2 is the identity matrix multiplied by 4.

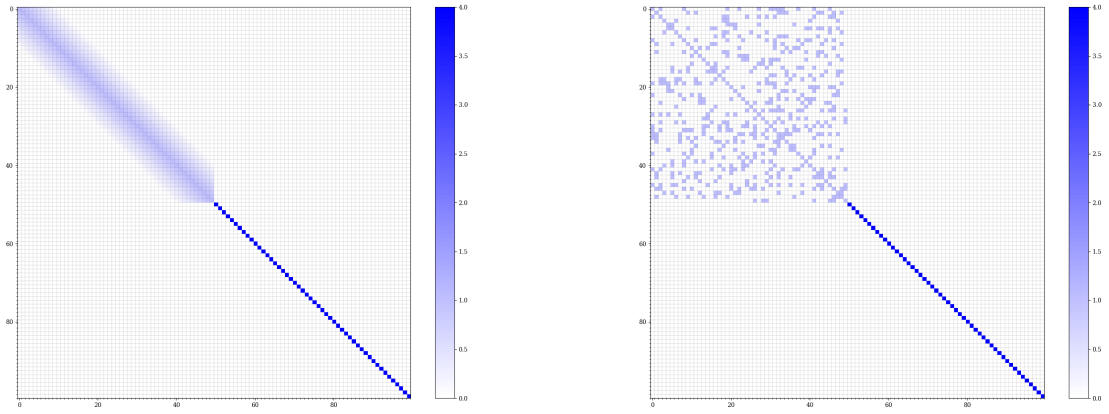
- **Model 2 (sparse matrix without ordering)**: $\mathbf{\Sigma} = \text{diag}\{A_1, A_2\}$, where $A_2 = 4\boldsymbol{J}_{\frac{N}{2}\times\frac{N}{2}}$, $A_1 = B + \epsilon\boldsymbol{J}_{\frac{N}{2}\times\frac{N}{2}}$, $B = (b_{ij})_{\frac{N}{2}\times\frac{N}{2}}$, whose elements independently follow:

$$b_{ij} = \begin{cases} \text{Ber}\left(\frac{20}{N}\right), & \text{for } i < j, \\ 1, & \text{for } i = j, \\ b_{ji}, & \text{for } i > j. \end{cases} \tag{13}$$

  $\text{Ber}(x)$ is a Bernoulli random variable that takes value 1 with probability x and value 0 with probability $1-x$, and $\epsilon = \max\{-\rho_{\min}(B), 0\} + 0.01$ to ensure that A_1 is positive definite.

4.2 Auxiliary Information

In the simulation, we directly generate the estimates of the Location Indicator Matrix L and the Relative Importance Indicator Matrix C, i.e., $\widehat{L}$ and $\widehat{C}$. The qualities of these estimates



(a) Model 1 (b) Model 2

Figure 1: Typical heatmaps of two banded and sparse models

are tuned by some hyper-parameters:

• **Observation Level l**: We set $l = 0.2$ which means $L_{ij} = 1$ if and only if $|r_{ij}| > 0.2$. Designed for the Network Guided Thresholding Estimator.

• **Type I error ζ**: Conditional on $L_{ij} = 0$, the probability of observing $\widehat{L}_{ij} = 1$. Designed for the Network Guided Thresholding Estimator.

• **Type II error $1 - p$**: Conditional on $L_{ij} = 1$, the probability of actually observing $\widehat{L}_{ij} = 1$. Designed for the Network Guided Thresholding Estimator.

• **Accuracy Rate η**: The probability of observing $j \in S_k^{\widehat{c}_i}$ conditional on $j \in S_k^{c_i}$. Designed for Network Guided Banding Estimator.

Table 1 lists the descriptions of these hyper-parameters and the ranges of values they take for the numerical experiment.

4.3 Numerical Results

The data is generated from a factor model and we focus on the covariance matrix of the residual. We draw i.i.d. samples $\boldsymbol{u}_t$ from $N(0, \sigma_u^2 \boldsymbol{\Sigma})$, where σ_u^2 is adjusted to match the daily variance

Table 1: Hyper-parameters Setup

Hyper-parameters	**Auxiliary Information Quality**			
	Very Bad	*Bad*	*Good*	*Perfect*
(p, ζ)	$(0.3, 0.1)$	$(0.6, 0.1)$	$(0.9, 0.1)$	$(1.0, 0.0)$
η	0.3	0.6	0.9	1.0

of the error term from the factor model of Liu et al. (2019). We use the CH-4 factor model[3] proposed by Liu et al. (2019), which consists of four factors: Market, Value-Minus-Growth, Small-Minus-Big and Pessimistic-Minus-Optimistic:

$$\boldsymbol{y}_t = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 f_{\text{MKT,t}} + \boldsymbol{\beta}_2 f_{\text{VMG,t}} + \boldsymbol{\beta}_3 f_{\text{SMB,t}} + \boldsymbol{\beta}_4 f_{\text{PMO,t}} + \boldsymbol{u}_t. \tag{14}$$

Regression results by using the weekly return of HS300 stocks from 2000 to 2021 show that the average level of σ_u is 4.85%, which can be adapted in our model. Besides, for generating purpose, we may need the true coefficients. One can simply run regression model for each HS300 component stock using 2000 - 2021 weekly data, then calculate the mean and standard deviation of the estimated coefficients. Based on the estimated coefficients, we dependently draw $\beta_{1i} \sim N(0.7013, 0.1961^2)$, $\beta_{2i} \sim N(-0.1582, 0.2055^2)$, $\beta_{3i} \sim N(-0.1200, 0.2182^2)$ and $\beta_{4i} \sim N(-0.0050, 0.2245^2)$, where the means and standard deviations are obtained in the last step. In addition, the true value of $\boldsymbol{\beta}_0$ is set to be $\mathbf{0}$.

For each model described in Subsection 4.1, a T-length sample of i.i.d. N-variate random vectors $\{\boldsymbol{u}_t\}$ is generated from the normal distribution with mean 0 and covariance matrix $\sigma_u^2 \boldsymbol{\Sigma}$, for $N = 100, 300, 500$. Then return vectors $\boldsymbol{y}_t$ are also obtained by Equation 14, where the weekly CH-4 returns are sampled from a normal distribution, whose mean and covariance matrix are set to be the same as the historical data from 2000 to 2021, shown in Table 2.

[3]The CH-4 factor model is found to suit China stock market well and outperform Fama-French 5 factor model.

Table 2: Descriptive Statistics of Factor Data

	Descriptive Statistics					**Correlation**			
	Count	Mean	Std.	Skew.	Kurt.	MKT	VMG	SMB	PMO
MKT	1119	0.1474%	3.3799%	-0.1019	2.5177	1.000	-0.237	0.159	-0.283
VMG	1119	0.2774%	1.7354%	1.0481	6.9085	-0.237	1.000	-0.637	0.215
SMB	1119	0.1189%	2.0311%	-0.5200	5.0999	0.159	-0.637	1.000	-0.137
PMO	1119	0.1887%	1.5882%	0.5986	8.1812	-0.283	0.215	-0.137	1.000

We gather estimates $\widehat{\boldsymbol{u}}_t$ and employ various methods to estimate $\sigma_u^2\widehat{\boldsymbol{\Sigma}}$. Each scenario, specified by either (N, T, p, ζ) or (N, T, η), is repeated 100 times to ensure robustness and consistency in our analysis. Our examination centers on the numerical efficacy[4] of both the Network Guided Thresholding Estimator and the Network Guided Banding Estimator, compared with a collection of purely statistical approaches: the Sample Covariance Estimator, Soft Thresholding Estimator, Hard Thresholding Estimator, Linear Shrinkage Estimator, and Nonlinear Shrinkage Estimator. The findings, delineated under both the Frobenius norm and the Matrix 2-norm, are catalogued in Table 3, illustrating the comparative performances of the diverse estimation techniques.

Panel A in Table 3 showcases the outcomes for our Model 1, where the true covariance matrix is banded with order. Here, we observe that both Network Guided estimators surpass their counterparts, provided the auxiliary network information is of reasonable quality. For the Network Guided Thresholding Estimator, barring the scenario where $N < T$ with $(p, \zeta) = (0.3, 0.1)$ indicating poor auxiliary information, it outperforms the sample covariance estimator, soft thresholding, hard thresholding estimator, and linear shrinkage estimator across all (p, ζ, N) combinations. Nonetheless, to eclipse the nonlinear shrinkage estimator, the quality

[4]Numerical performance is assessed through the comparison of $\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_\bullet$, incorporating both the Frobenius norm and the Matrix 2-norm.

of the information needs to be comparably higher. Specifically, the type I error ζ and type II error $1-p$ significantly impair the performance of the Network Guided Thresholding Estimator. In an ideal scenario where $(p, \zeta) = (1.0, 0.0)$, we achieve $\widehat{L} = L$, resulting in exceptionally low errors. When examining the Network Guided Banding Estimator, it exhibits smaller norms than all other purely statistical methods, assuming the accuracy rate parameter η is not excessively low. Particularly, with $\eta = 0.6$, the Network Guided Banding Estimator demonstrates superiority for most (N, T) combinations, notably when $N \geq T$. With comparable information quality, the Network Guided Banding Estimator typically outshines the Network Guided Thresholding Estimator, aligning with theoretical expectations. Our objective is not to pit the two Network Guided estimators against each other as it would not constitute a fair comparison; the Network Guided Banding Estimator necessitates auxiliary information that discloses the relative significance of neighbors for each node, rendering this method applicable. Panel B is devoted to the scenario where the true covariance matrix is a sparse matrix without order (Model 2). In this context, both Network Guided estimators continue to outshine the competition, assuming the auxiliary network information is of sufficient quality. Given that our Network Guided Banding Estimator is adaptable to a broader spectrum of "bandable" matrices, it proves to be effective in the settings of Model 2. Its performance remains outstanding, provided that the accuracy rate parameter η is not unduly low. Notably, with $\eta = 0.6$, the Network Guided Banding Estimator surpasses other methods across all (N, T) combinations. Similar to the previous model, the Network Guided Thresholding Estimator exhibits strong performance, particularly when the incidence of Type I errors is minimized.

In summary, our simulation exercise underscores the exceptional numerical properties of the proposed Network Guided estimators. Both estimators consistently outperform their counterparts, contingent upon the adequacy of the auxiliary network information.

Table 3: Simulation Results: Error Comparison of Different Estimators. We compare our methods with some benchmarks, including Sample Covariance matrix (Sample), Soft Thresholding (S-Thres.), Hard Thresholding (H-Thres.), Linear Shrinkage (L-Shrin.) and Non-linear Shrinkage (N-Shrin.). Note that Non-linear Shrinkage method only works when $T > N$. For every method, 100 times of simulations are made, thus we give the mean and standard deviation.

Setting		**Network Guided Thresholding**				**Network Guided Banding**				**Benchmarks**				
		$(0.3, 0.1)$	$(0.6, 0.1)$	$(0.9, 0.1)$	$(1.0, 0.0)$	$\eta = 0.3$	$\eta = 0.6$	$\eta = 0.9$	$\eta = 1.0$	Sample	S-Thres.	H-Thres.	L-Shrin.	N-Shrin.
• Panel A: Model 1, banded matrix with ordering														
$N = 100$	$\|\cdot\|_F$	14.77	11.46	7.14	3.59	13.97	10.74	5.91	3.04	14.47	16.56	16.43	12.25	7.56
		(0.07)	(0.10)	(0.17)	(0.25)	(0.26)	(0.39)	(0.48)	(0.32)	(0.32)	(0.03)	(0.76)	(0.19)	(0.30)
	$\|\cdot\|_2$	6.83	4.45	2.05	1.34	6.44	4.23	2.02	1.28	4.47	8.74	8.66	3.74	3.59
		(0.17)	(0.29)	(0.31)	(0.31)	(0.28)	(0.37)	(0.31)	(0.36)	(0.40)	(0.04)	(0.48)	(0.32)	(0.39)
$N = 300$	$\|\cdot\|_F$	28.47	23.00	17.14	6.44	24.65	18.94	10.59	5.40	43.25	29.26	28.86	29.11	
		(0.13)	(0.16)	(0.22)	(0.24)	(0.28)	(0.36)	(0.46)	(0.30)	(0.37)	(0.04)	(1.81)	(0.11)	
	$\|\cdot\|_2$	7.14	4.65	2.59	1.65	6.81	4.58	2.37	1.58	9.13	8.98	8.85	5.63	
		(0.10)	(0.16)	(0.21)	(0.25)	(0.17)	(0.24)	(0.26)	(0.25)	(0.41)	(0.02)	(0.60)	(0.16)	
$N = 500$	$\|\cdot\|_F$	39.34	33.66	27.11	8.33	31.96	24.63	13.68	6.96	71.88	37.91	41.10	41.44	
		(0.13)	(0.16)	(0.20)	(0.21)	(0.32)	(0.38)	(0.44)	(0.27)	(0.42)	(0.03)	(12.41)	(0.10)	
	$\|\cdot\|_2$	7.16	4.94	2.95	1.78	6.88	4.69	2.46	1.71	12.72	9.01	9.19	6.30	
		(0.09)	(0.13)	(0.13)	(0.23)	(0.17)	(0.23)	(0.26)	(0.27)	(0.38)	(0.02)	(1.63)	(0.10)	
• Panel B: Model 2, sparse matrix without ordering														
$N = 100$	$\|\cdot\|_F$	19.69	16.36	12.25	10.93	17.62	14.17	9.56	7.39	25.87	20.49	20.42	16.29	15.25
		(0.13)	(0.17)	(0.26)	(0.30)	(0.32)	(0.49)	(0.45)	(0.34)	(0.44)	(0.05)	(0.67)	(0.25)	(0.30)
	$\|\cdot\|_2$	7.58	5.14	3.13	2.73	7.26	4.87	2.88	2.26	7.02	9.79	9.72	6.83	5.73
		(0.24)	(0.30)	(0.19)	(0.23)	(0.30)	(0.39)	(0.22)	(0.24)	(0.50)	(0.07)	(0.59)	(0.58)	(0.78)
$N = 300$	$\|\cdot\|_F$	39.84	35.57	31.27	14.22	30.20	24.66	17.39	14.22	83.04	34.86	34.86	33.86	
		(0.20)	(0.26)	(0.30)	(0.33)	(0.29)	(0.40)	(0.46)	(0.33)	(0.50)	(0.06)	(0.06)	(0.10)	
	$\|\cdot\|_2$	7.86	5.73	4.29	2.71	7.39	5.06	3.22	2.71	15.27	9.86	9.86	9.09	
		(0.19)	(0.20)	(0.14)	(0.22)	(0.20)	(0.22)	(0.17)	(0.22)	(0.54)	(0.05)	(0.05)	(0.23)	
$N = 500$	$\|\cdot\|_F$	58.38	53.24	48.14	17.96	38.87	31.53	22.21	17.96	136.36	44.96	44.96	44.72	
		(0.27)	(0.28)	(0.33)	(0.37)	(0.29)	(0.40)	(0.46)	(0.37)	(0.63)	(0.05)	(0.05)	(0.07)	
	$\|\cdot\|_2$	8.25	6.02	4.91	2.67	7.30	4.95	3.19	2.67	21.20	9.73	9.73	9.43	
		(0.17)	(0.15)	(0.13)	(0.22)	(0.14)	(0.18)	(0.16)	(0.22)	(0.47)	(0.04)	(0.04)	(0.15)	

5 Empirical Study

5.1 Data

5.1.1 Assets Returns

Stocks in our sample are constituent stocks in 2021 of three famous indices in China , namely HS300 (000300.SH), CSI500 (000905.SH) and CSI800 (000906.SH), which consist of around 300, 500 and 800 stocks individually. The daily returns of the stocks are collected from the RESSET database, ranged from 2006 to 2021 with the ST stocks excluded. From RESSET database, we collect the daily return from 2006 - 2021 of these sample stocks.

5.1.2 News Co-mention Linkage Data

We analyzed over millions of articles from the Financial Text Intelligent Analysis Platform of RESSET and the Juyuan Database, spanning from 2006 to 2021. We selected articles that mentioned at least one publicly traded company in China's A-share market, totaling 1,138,247 news pieces left.

Following the approach of Ge et al. (2023), we define news-implied links based on shared mentions within the same news article. According to readers' reading habits, we proposed four methods to identify connectivity among firms, namely `one2one_passage`, `all_passage`, `one2one_setence` and `all_setence` approaches. In Table 4, we summarize the differences of these approaches:

Table 4: News Co-mention Types

	Firms Co-mentioned	
	in the same passage	**in the same sentence**
if more than two firms are co-mentioned	`all_passage`	`all_sentence`
if and only if two firms are co-mentioned	`one2one_passage`	`one2one_sentence`

At time t, we set the latest τ_0 days as the identification window[5]. For each stock pair (i, j), we count the number of co-mention M_{ij} under one co-mention type to construct the co-mention matrix $M = (M_{ij})$ for $i, j = 1, 2, \ldots, N$.

5.1.3 Analyst Coverage Linkage

In parallel, we explore linkages based on the analyst coverage, termed `Analyst`. This approach is supported by literature suggesting that shared analyst attention may indicate fundamental connections between companies, reflecting similarities over various dimensions (see Ali and Hirshleifer (2020), Israelsen (2016) and Kaustia and Rantala (2013)). We utilized the data from the Chinese Research Data Services Platform (CNRDS), covering analyst reports from January 2005 to December 2020. After the data cleaning, we identified 530,696 unique analyst reports to trace connections based on shared coverage. Starting from 2006, at time t. we use a one-year lag identification window for linkage construction. Similar to the news co-mention linkage, for each stock pair (i, j), we count the number of coverage M_{ij} during the identification window to build the analyst co-coverage connection matrix $M = (M_{ij})$ for $i, j = 1, 2, \ldots, N$.

5.1.4 Industry-based Linkage

Additionally, we examine linkages formed based on industry classifications, marked as `Industry`. This approach draws on the findings of Moskowitz and Grinblatt (1999) and Engelberg et al. (2018), who noted that stocks within the same sector often move together significantly. We analyzed three major industry classification systems in China: CSRC, CITIC, and Shenwan, updating annually according to the RESSET database. Our primary focus is the Shenwan primary classification, which is recognized as the leading system within China's financial industry.

[5]Empirically, for the purpose of testing linkage performance under short and long identification windows, τ_0 is chosen to be 21 (1 month) or 252 (12 months).

We report the summary statistics of these different networks in Table 5. Under `sentence_1`, each focal firm has 16 peer firms on average, fewer than 29 peers from `article_1`. This aligns with our expectations as the same sentence strategy removes the potential noise links from the same article strategy, resulting in fewer links identified. Furthermore, the number of peer firms identified naturally increases with the length of the identification window. For other linkage types, we generally observe a higher number of links, and each sample stock tends to have more peers on average.

Table 5: Networks Summary Statistics. The sample stocks include all listed stocks on the main board of the Shanghai Stock Exchange, Shenzhen Stock Exchange, and Growth Enterprise Market (GEM). ST shares are excluded.

Link Type	**Variables**	**Mean**	**Std.**	**Min.**	**Median**	**Max.**
`all_sentence_1`	# Stocks	1332	293	903	1234	2223
	# Peer firms	16	32	1	5	454
`all_sentence_12`	# Stocks	1750	233	1355	1742	2704
	# Peer firms	23	42	1	8	631
`all_passage_1`	# Stocks	1976	229	1478	1952	2816
	# Peer firms	29	51	1	10	757
`all_passage_12`	# Stocks	2122	278	1569	2121	2891
	# Peer firms	35	59	1	12	867
`analyst`	# Stocks	1326	348	476	1429	1872
	# Peer firms	98	84	1	75	609
`industry`	# Stocks	2336	795	1048	2313	3893
	# Peer firms	130	83	2	110	364

5.2 Methodolody

Our goal is to construct GMV portfolio with the help of auxiliary information. This subsection presents the procedures for the proposed model to be applied. We first de-factor the stock returns through observable factors. Then, we regard the covariance matrix of focal stocks' de-factored returns as static and use the proposed method and sample to estimate. In the training step, some tuning parameters are validated. Finally, we test the out-of-sample performance of the proposed models with benchmark models.

5.2.1 CH-4 Factor Model

We adopt CH-4 factors model as in Liu et al. (2019) to de-factor asset returns:

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 f_{\mathrm{MKT,t}} + \boldsymbol{\beta}_2 f_{\mathrm{VMG,t}} + \boldsymbol{\beta}_3 f_{\mathrm{SMB,t}} + \boldsymbol{\beta}_4 f_{\mathrm{PMO,t}} + \boldsymbol{u}_t \\ &= \boldsymbol{\beta}_0 + \boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{u}_t, \end{aligned} \tag{15}$$

where $\boldsymbol{f}_t$ is obtained from the authors' website. Thus the estimator of $\boldsymbol{\Sigma}_y = \mathrm{Cov}\left(\boldsymbol{y}_t, \boldsymbol{y}_t\right)$ is given by

$$\widehat{\boldsymbol{\Sigma}}_y = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^{\intercal} + \widehat{\boldsymbol{\Sigma}}_u, \tag{16}$$

where the factor loading matrix $\widehat{\boldsymbol{B}}$ is obtained by the OLS. Henceforth, our goal is to estimate the covariance matrix of residuals $\boldsymbol{\Sigma}_u$.

5.2.2 The Estimation of $[L_{ij}]_{N\times N}$ and $[C_{ij}]_{N\times N}$

News co-mention, as auxiliary information, can be used to facilitate both Thresholding and Banding estimation. For the Network Guided Thresholding, we also set a thresholding m for the news co-mention times. For a pair of stocks i and j, if we have $M_{ij} \geq m$ in the identification window, then $\widehat{L}_{ij} = 1$. Generally, we estimate the indicator matrix $\widehat{L} = \left(\widehat{L}_{ij}\right)$ as $\widehat{L}_{ij} = \mathbf{1}_{\{M_{ij}\geq m\}}$, where the tuning parameter m is chosen by the in-sample cross validation. Empirically, the objective function for the in-sample training is the GMV portfolio.

As for the Network Guided Banding Estimation, the news co-mention connection is ready to use since it provides integer counts with $0 \leq M_{ij} < \infty$. Therefore, we set $\widehat{C} = M$ and apply our network-guided procedure introduced in Section 2.

The analyst co-coverage data have exactly the same properties as the news co-mention data, and therefore, all the procedures are identical as described above. As a comparison, the industry-based linkage is a 0-1 indicator, where $M_{ij} = \mathbf{1}_{\{i \text{ and } j \text{ are in the same industry}\}}$. Therefore, we do not need to choose the tuning parameter m as we did for the news co-mention and analyst co-coverage linkages. For Network Guided Thresholding, we directly have $\widehat{L}_{ij} = M_{ij}$ while for Network Guided Banding, we have $\{i \in S_k^{c_j}, j \in S_k^{c_i}\}$ if and only if stock i, j are in the same industry.

5.2.3 Comparing the Out-of-sample Portfolios

As discussed in Engle et al. (2019) and Chen et al. (2019), constructing a global minimum variance (GMV) portfolio is a desirable way to assess the performance covariance matrix estimators. Compared to the optimal mean-variance (MV) portfolio, a global minimum variance (GMV) portfolio can avoid the estimation of asset mean returns, which contributes considerable noise. Therefore, we apply the proposed method to a portfolio management problem in this part. In particular, we compare the performance of GMV portfolios as in Ledoit and Wolf (2004). The theoretical weights for a GMV portfolio are given by

$$\boldsymbol{w}^{\text{GMV}} = \frac{\boldsymbol{\Sigma}_y^{-1}\mathbf{1}}{\mathbf{1}^\intercal \boldsymbol{\Sigma}_y^{-1}\mathbf{1}},$$

where $\boldsymbol{\Sigma}_y$ is the estimated covariance matrix of asset returns with $\mathbf{1}$ the conforming vector of ones. Given the factor structure asset returns, we have $\widehat{\boldsymbol{\Sigma}}_y = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^\intercal + \widehat{\boldsymbol{\Sigma}}_u$. The co-movement part can be straightforwardly estimated by the CH4 factor model, and our goal is to show that the proposed method can better estimate $\widehat{\boldsymbol{\Sigma}}_u$ and contribute to the GMV portfolio performance. Based on rolling window manners, which started in 2012, we take one year of data

for model training and substitute the in-sample results for a one-month test. We will continue this procedure until the end of 2021 and summarize the one-month out-of-sample performance. Namely, the portfolio is adjusted monthly for nine years in total.

Besides, for robustness test, we still consider the Maximal Return portfolio for any given variance level σ_0^2 and Minimal Variance portfolio for any given expected return level μ_0. Recall the construction of classical optimal portfolio, for example, given a return constraint μ_0, we have the minimization problem:

$$\min \boldsymbol{w}^\intercal \boldsymbol{\Sigma}_y \boldsymbol{w} \quad \text{s.t.} \quad \boldsymbol{w}^\intercal \boldsymbol{\mu} \geq \mu_0,$$

where $\boldsymbol{\mu} = E(\boldsymbol{y}_t)$, and the weight is given by

$$\boldsymbol{w}(\mu_0) = \frac{1}{|\boldsymbol{\Psi}|} \cdot \left[(\psi_{22} - \psi_{12}\mu_0) \boldsymbol{\Sigma}_y^{-1} \mathbf{1} + (\psi_{11}\mu_0 - \psi_{12}) \boldsymbol{\Sigma}_y^{-1} \boldsymbol{\mu} \right],$$

where the matrix $\boldsymbol{\Psi}$ is defined as

$$\boldsymbol{\Psi} = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{1}^\intercal \boldsymbol{\Sigma}_y^{-1} \mathbf{1} & \mathbf{1}^\intercal \boldsymbol{\Sigma}_y^{-1} \boldsymbol{\mu} \\ \boldsymbol{\mu}^\intercal \boldsymbol{\Sigma}_y^{-1} \mathbf{1} & \boldsymbol{\mu}^\intercal \boldsymbol{\Sigma}_y^{-1} \boldsymbol{\mu} \end{pmatrix},$$

details and proofs can be found in Chapter 1.6 of Linton (2019). Given the factor structure assets returns, we have $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{B}}\bar{\boldsymbol{f}}$ and $\widehat{\boldsymbol{\Sigma}}_y = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^\intercal + \widehat{\boldsymbol{\Sigma}}_u$ as discussed before. By the weight formula, the minimal variance given μ_0 is

$$\sigma_0^2 = \boldsymbol{w}(\mu_0)^\intercal \boldsymbol{\Sigma} \boldsymbol{w}(\mu_0) = \frac{1}{|\boldsymbol{\Psi}|} \left(\psi_{11}\mu_0^2 - 2\psi_{12}\mu_0 + \psi_{22} \right),$$

which also gives the mean-variance efficient frontier set $\{(\sigma_0, \mu_0), \mu_0 \geq 0\}$. Starting from maximization problem for any given σ_0^2 leads to the same efficient frontier. But note that the efficient frontier is in-sample, and when we set a fixed in-sample σ_0 or μ_0, the out-of-sample portfolio may give different values of standard deviation and mean return, which lead to out-of-sample efficient frontier. Similar to GMV portfolio, we choose tuning parameters via in-sample training and construct the out-of-sample efficient frontiers under different models.

Importantly, although most of the estimated covariance matrices are positive definite, we modify all non-positive definite covariance matrices $\widehat{\boldsymbol{\Sigma}}_u$ by the method given in Equation 12.

5.3 Empirical Results

5.3.1 Comparing GMV Portfolios

Table 6 reports the out-of-sample volatility (measured by standard deviation) of GMV portfolios constructed by different methods and stock samples, including constituent stocks of the HS300, CSI500, and CSI800 indices. We also summarize benchmark models as follows:

- **Sample**: Use the sample covariance matrix of $\widehat{\mathbf{\Sigma}}_u$ with positive definite correction if non-positive.
- **Linear Shrinkage**: Operate linear shrinkage of Ledoit and Wolf (2004) on $\widehat{\mathbf{\Sigma}}_u$ with positive definite correction if non-positive.
- **Factor Only**: Take $\widehat{\boldsymbol{B}}\widehat{\mathbf{\Sigma}}_f\widehat{\boldsymbol{B}}^\intercal + \text{diag}\{\widehat{\sigma}_1^2, \cdots, \widehat{\sigma}_N^2\}$ as $\widehat{\mathbf{\Sigma}}_y$.
- **Equal Weights**: Take equal weights $\frac{1}{N}$ of N assets as the out-of-sample GMV portfolios.

The overall results are shown in Panel A, where 'Best Thresholding' and 'Best Banding' are portfolios with the best performance from different auxiliary information, detailed in Panel B (Network Guided Thresholding) and Panel C (Network Guided Banding), respectively. The best connection information for Network Guided Thresholding is `one2one_passage_12` while for the Network Guided Banding is `one2one_passage_1`.

From the benchmarks presented in Panel A, it is evident that the 'Factors Only' approach consistently outperforms the 'Sample' method across all indices. This suggests that the co-movement part can excellently model the covariance of asset returns, with the sample covariance matrix of idiosyncratic risk being noisy. However, it is notable that the 'Linear Shrinkage' method offers a competitive, if not superior, reduction in standard deviation compared to 'Factors Only', especially for the CSI500 index, highlighting the potential of shrinkage methods in improving portfolio efficiency.

In the realm of Network Guided Estimation, the results exhibit a varied performance landscape. For the Network Guided Thresholding method (Panel B), incorporating `analyst` and `industry` network information leads to a noticeable improvement in the HS300 and CSI800 indices but shows a mixed effect on CSI500, suggesting that the effectiveness of network information varies over different groups of stocks. Furthermore, different network structures, such as `one2one_sentence` and `all_passage`, provide insights into how the granularity and context of network connections influence the estimation accuracy.

The Network Guided Banding approach (Panel C) generally shows an improvement over the traditional and network-thresholding methods for the HS300 index, mainly when using the `analyst` and `industry` networks. This underscores the importance of the quality and type of network information in enhancing covariance matrix estimation. The mixed results across different indices and network structures suggest that while network information is valuable, its application needs to be tailored to specific market conditions and characteristics.

Table 6: Out-of-sample Standard Deviation of GMV Portfolios. We compare different portfolios' out-of-sample standard deviations, most of which are GMV portfolios constructed based on certain covariance matrix estimators. 'Sample' refers to a simple sample estimator of $\mathbf{\Sigma}_u$, 'Factors Only' to setting $\widehat{\mathbf{\Sigma}}_u = \text{diag}\{\widehat{\sigma}_1^2, \cdots, \widehat{\sigma}_N^2\}$, and 'Equal Weights' to a simple equal-weights portfolio instead of GMV.

Index	**Out-of-sample Standard Deviation of GMV Portfolios Under Different Estimators**					
• Panel A: Overall						
	Sample	Linear Shrinkage	Factors Only	Equal Weights	Best Thresholding	Best Banding
HS300	0.0513	0.0480	0.0440	0.0717	**0.0426**	0.0445
CSI500	0.0739	0.0703	0.0732	0.0820	**0.0683**	0.0731
CSI800	0.0593	0.0575	0.0547	0.0769	**0.0499**	0.0532
• Panel B: Network Guided Thresholding						
	analyst	industry	all_passage_1	all_sentence_1	one2one_passage_1	one2one_sentence_12
HS300	0.0507	0.0472	0.0457	0.0457	0.0447	0.0470
CSI500	0.0722	0.0760	0.0685	**0.0683**	0.0686	0.0684
CSI800	0.0558	0.0604	0.0508	0.0503	0.0505	0.0510
	one2one_sentence_1	all_passage_12	all_sentence_12	one2one_passage_12		
HS300	0.0448	0.0508	0.0452	**0.0426**		
CSI500	0.0685	0.0756	0.0700	0.0687		
CSI800	0.0506	0.0582	**0.0499**	0.0500		
• Panel C: Network Guided Banding						
	analyst	industry	all_passage_1	all_sentence_1	one2one_passage_1	one2one_sentence_12
HS300	0.0460	0.0462	0.0483	0.0469	**0.0445**	0.0489
CSI500	0.0742	**0.0731**	0.0756	0.0733	0.0744	0.0768
CSI800	0.0598	0.0571	0.0558	0.0556	**0.0532**	0.0537
	one2one_sentence_1	all_passage_12	all_sentence_12	one2one_passage_12		
HS300	0.0467	0.0488	0.0504	0.0513		
CSI500	0.0737	0.0741	0.0735	0.0765		
CSI800	0.0538	0.0605	0.0588	0.0547		

5.3.2 Other Mean-Variance Portfolios

For robustness, we also test other optimal portfolios under different covariance matrix estimations. To avoid extreme situations (N is too small or too large), we choose constituent stocks of CSI500 index to delve into optimal portfolio tests. We calculate the out-of-sample efficient frontiers via different methods, which are shown in the Figure 2 with returns and volatility are all annualized.

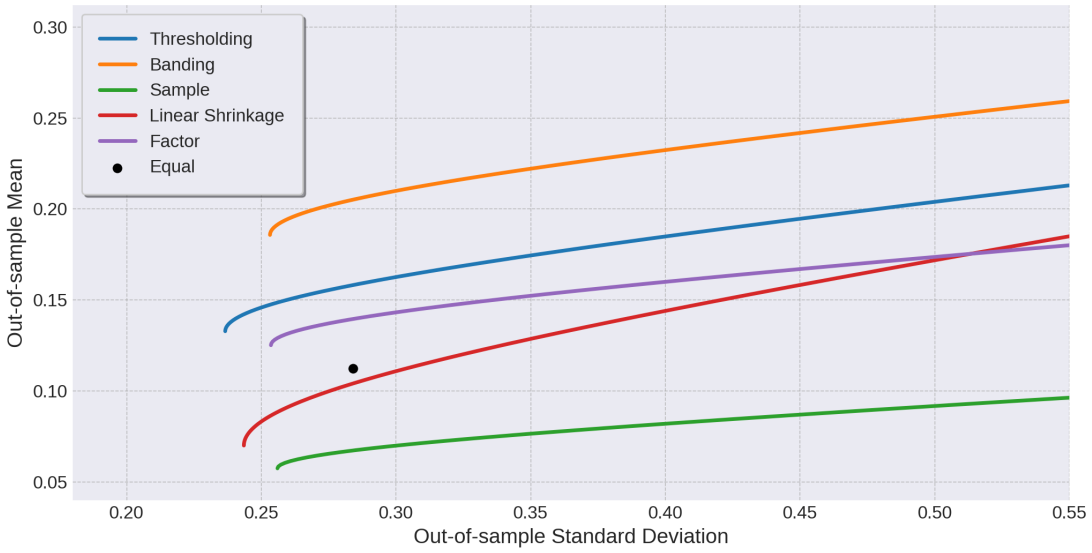


Figure 2: Out-of-sample Efficient Frontiers

In Figure 2, we can see that the Network Guided Thresholding method reaches the minimal variance, which matches the results in the Table 6. But when including out-of-sample performance, the portfolio constructed by Network Guided Banding dominates Network Guided Thresholding as well as all other baseline models. Besides Thresholding and Banding, only 'Factor' method performs significantly better than the 'Equal Weights' portfolio. Linear Shrinkage is not well performing when σ is low, but outperforms 'Factor' method when σ is relatively high. Finally, Sample covariance matrix is perfected dominated when constructing mean-variance portfolios in our study, which stresses the necessity of modifying of the estimation of large covariance matrix.

Table 7: Portfolios Performances Given Out-of-sample Standard Deviations.

Out-sample-sample Statistics		**Benchmarks**				**Network Guided**	
		Sample	Linear Shrinkage	Factors Only	Equal Weights	Best Thresholding	Best Banding
Std. = 26%	*Mean*	6.11%	9.13%	13.16%		15.02%	19.46%
	Sharpe	0.120	0.236	0.391		0.462	0.633
Std. = 27%	*Mean*	6.44%	9.72%	13.56%		15.38%	19.98%
	Sharpe	0.127	0.249	0.391		0.459	0.629
Std. = 28%	*Mean*	6.66%	10.23%	13.85%		15.70%	20.37%
	Sharpe	0.131	0.258	0.388		0.454	0.620
Std. = 28.41%	*Mean*	6.74%	10.42%	13.96%	11.23%	15.82%	20.51%
	Sharpe	0.131	0.261	0.386	0.290	0.451	0.616
Std. = 29%	*Mean*	6.84%	10.68%	14.10%		15.99%	20.70%
	Sharpe	0.132	0.265	0.383		0.448	0.610
Std. = 30%	*Mean*	7.00%	11.09%	14.32%		16.26%	21.00%
	Sharpe	0.133	0.270	0.377		0.442	0.600
Std. = 31%	*Mean*	7.15%	11.48%	14.52%		16.52%	21.27%
	Sharpe	0.134	0.273	0.372		0.436	0.589

Table 7 reports the portfolios performances for some given out-of-sample σ, results are consistent with the Figure 2. We also compared the Sharpe ratio under given portfolio volatility, where Network Guided portfolios have higher Sharpe ratios, with Network Guided Banding the best.

Furthermore, we analyse maximal Sharpe Ratio portfolios (or Mean-Variance optimal portfolios) under different models. We search the efficient frontier depicted in Figure 2 and find the mean-variance optimal results for each model to compare. Figure 3 plots the backtest performances over 10-year out-of-sample window of these portfolios, and the evaluation statistics are presented in Table 8. For simplicity, the tuning parameters are substituted by the results from the GMV portfolios. Unfortunately, due to the crash of the Chinese stock market during May and June 2015, no portfolio gets a decent Sharpe ratio higher than 1. But compared to other four benchmarks, Network Guided portfolios perform better in the whole period, especially the Banding one. For return and standard deviation, 'Factor' method is close to our Network Guided Thresholding, but 'Factor' method tends to produce higher maximum drawdown. Network Guided Banding portfolio provides the best performance in our backtest, with the highest return and the lowest maximum draw-down.

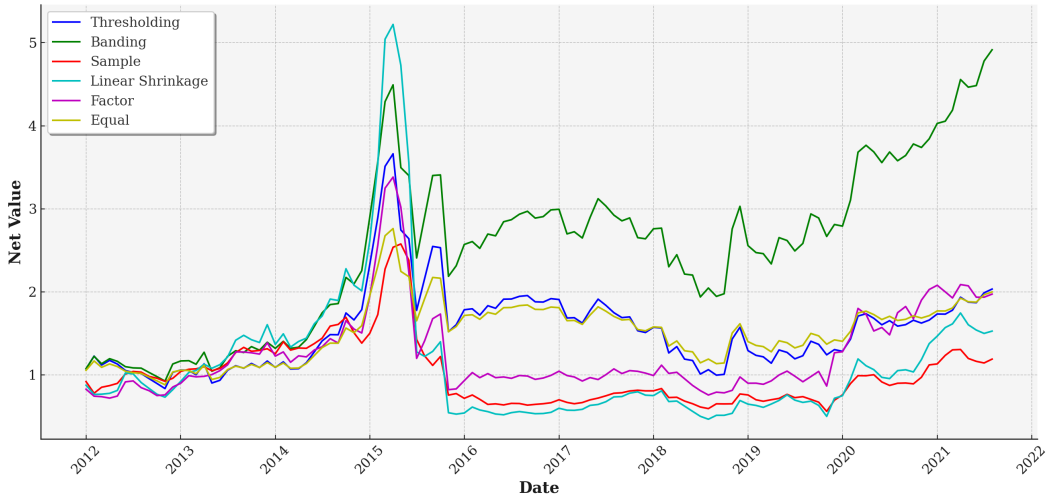


Figure 3: Out-of-sample Mean-Variance Optimal Portfolios

Table 8: Mean-Variance Optimal Portfolios Performances

	Sample	Linear Shrinkage	Factors Only	Equal Weights	Thresholding	Banding
Mean Return	7.25%	14.71%	13.42%	11.23%	14.68%	19.47%
Std. Dev.	31.77%	41.04%	26.59%	28.41%	25.20%	26.02%
Sharpe Ratio	0.134	0.285	0.392	0.290	0.464	0.633
Max Draw-down	78.24%	91.07%	77.60%	58.90%	72.79%	56.85%

In conclusion, these results affirm the utility of incorporating network information into covariance matrix estimation for portfolio optimization. Since the factor model can only capture the strong or global co-movement among asset returns while the auxiliary information dissects the weak or local effects among focal stock and its peers, as discussed in Ge et al. (2022). This is the main reason that auxiliary information helps the estimation of $\widehat{\mathbf{\Sigma}}_u$ However, they also highlight the complexity and contextual nature of financial markets, where the effectiveness of such information can vary across different environments and conditions. Future research could delve deeper into the mechanisms behind these variations and explore the integration of additional types of network data to refine the estimation process further.

6 Conclusion

In the era of big data, we are gaining access to more and more auxiliary information apart from the observations of $\{\boldsymbol{X}_t\}_{t=1}^T$, which can potentially help us to improve the performance of conventional statistical and econometric models. We give different revenues for how to incorporate information from other sources to enhance the estimation of large covariance matrix of asset returns. According to the types of auxiliary information, we tailor them to fit the conventional thresholding and banding estimators. We also provide theoretical results to show that the proposed methods have better properties with the help of extra information. Both simulation studies and empirical illustrations validate that the proposed estimators are outstanding

compared with many benchmark models.

In this paper, we mainly discuss the static covariance matrix. However, we suggest that a similar idea can be extended to many other settings, like the estimation of large dynamic covariance matrices. For example, dynamic network information could be well incorporated into the conditioning information set in Chen et al. (2019).

References

U. Ali and D. Hirshleifer. Shared analyst coverage: Unifying momentum spillover effects. *Journal of Financial Economics*, 136(3):649–675, 2020.

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, (6):2577–2604, 2008a.

P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008b.

T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

J. Chen, D. Li, and O. Linton. A new semiparametric estimation approach for large dynamic covariance matrices with multiple conditioning variables. *Journal of Econometrics*, 212(1): 155–176, 2019.

Z. Chen and C. Leng. Dynamic covariance models. *Journal of the American Statistical Association*, 111(515):1196–1207, 2016. doi: https://doi.org/10.1080/01621459.2015.1077712.

J. Engelberg, A. Ozoguz, and S. Wang. Know thy neighbor: Industry clusters, information spillovers, and market efficiency. *Journal of Financial and Quantitative Analysis*, 53(5): 1937–1961, 2018.

R. F. Engle, O. Ledoit, and M. Wolf. Large dynamic covariance matrices. *Journal of Business & Economic Statistics*, 37(2):363–375, 2019.

J. Fan, Y. Liao, and M. Mincheva. High-Dimensional Covariance Matrix Estimation in Approximate Factor Models. *The Annals of Statistics*, 39(6), Dec. 2011. ISSN 0090-5364. doi: 10.1214/11-AOS944.

J. Fan, Y. Liao, and M. Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 75(4), 2013.

J. Fan, A. Furger, and D. Xiu. Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high-frequency data. *Journal of Business & Economic Statistics*, 34(4):489–503, 2016.

S. Ge, S. Li, and O. Linton. News-implied linkages and local dependency in the equity market. *Journal of Econometrics*, 2022.

S. Ge, S. Li, and H. Zheng. Diamond cuts diamond: News co-mention momentum spillover prevails in china. *Available at SSRN 4489005*, 2023.

G. Hoberg and G. Phillips. Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5):1423–1465, 2016.

R. D. Israelsen. Does common analyst coverage explain excess comovement? *Journal of Financial and Quantitative Analysis*, 51(4):1193–1229, 2016.

M. Kaustia and V. Rantala. Common analyst-based method for defining peer firms. *Available at SSRN*, 2013.

O. Ledoit and M. Wolf. Honey, I Shrunk the Sample Covariance Matrix. (4):110–119, 2004. ISSN 0095-4918, 2168-8656. doi: 10.3905/jpm.2004.110.

O. Ledoit, M. Wolf, et al. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, 2012.

Z. M. Li and O. Linton. A remedi for microstructure noise. *Econometrica*, 90(1):367–389, 2022. doi: https://doi.org/10.3982/ECTA17505.

O. Linton. *Financial econometrics*. Cambridge University Press, 2019.

J. Liu, R. F. Stambaugh, and Y. Yuan. Size and value in china. *Journal of financial economics*, 134(1):48–69, 2019.

H. M. Markowitz et al. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

T. J. Moskowitz and M. Grinblatt. Do industries explain momentum? *The Journal of finance*, 54(4):1249–1290, 1999.

A. J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

Appendices

A Proof of Theorem 1

Proof. We rationally have the decomposition $\left\|\widehat{\mathbf{\Sigma}}_{\widehat{L}}^{\mathcal{T}}-\mathbf{\Sigma}\right\| \leq \left\|\widehat{\mathbf{\Sigma}}_{\widehat{L}}^{\mathcal{T}}-\widehat{\mathbf{\Sigma}}_{L}^{\mathcal{T}}\right\|+\left\|\widehat{\mathbf{\Sigma}}_{L}^{\mathcal{T}}-\widetilde{\mathbf{\Sigma}}_{L}^{\mathcal{T}}\right\|+\left\|\widetilde{\mathbf{\Sigma}}_{L}^{\mathcal{T}}-\mathbf{\Sigma}_{L}^{\mathcal{T}}\right\|+\left\|\mathbf{\Sigma}_{L}^{\mathcal{T}}-\mathbf{\Sigma}\right\|$, where $\widetilde{\mathbf{\Sigma}}=\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{u}_t\boldsymbol{u}_t^{\intercal}$ is the non-observable sample covariance matrix. However, please note our thresholding is set for R, thus, it is necessary to consider

$$\left\|\widehat{R}_{\widehat{L}}^{\mathcal{T}}-R\right\| \leq \left\|\widehat{R}_{\widehat{L}}^{\mathcal{T}}-\widehat{R}_{L}^{\mathcal{T}}\right\|+\left\|\widehat{R}_{L}^{\mathcal{T}}-\widetilde{R}_{L}^{\mathcal{T}}\right\|+\left\|\widetilde{R}_{L}^{\mathcal{T}}-R_{L}^{\mathcal{T}}\right\|+\left\|R_{L}^{\mathcal{T}}-R\right\|.$$

The first part is the distance between the "real thresholding" correlation matrix and "estimated thresholding" correlation matrix, which is different from the other parts, we leave it to be bounded finally.

For the second part, we have

$$\begin{aligned}
\left\|\widehat{R}_L^{\mathcal{T}} - \widetilde{R}_L^{\mathcal{T}}\right\| &\leq \max_{1\leq i\leq N} \sum_{j=1}^{N} \left|s_{L,\lambda}\left(\widehat{r}_{ij}\right) - s_{L,\lambda}\left(\widetilde{r}_{ij}\right)\right| \\
&\leq \max_{1\leq i\leq N} \sum_{j=1}^{N} \left\{\left|\widehat{r}_{ij} - \widetilde{r}_{ij}\right| \cdot I_{\{L_{ij}=1\}} + \left|s_{\lambda}\left(\widehat{r}_{ij}\right) - s_{\lambda}\left(\widetilde{r}_{ij}\right)\right| \cdot I_{\{L_{ij}=0\}}\right\} \\
&\leq c_1 \max_{i,j} \left|\widehat{r}_{ij} - \widetilde{r}_{ij}\right| + \max_{1\leq i\leq N} \sum_{j=1}^{N} \left|s_{\lambda}\left(\widehat{r}_{ij}\right) - s_{\lambda}\left(\widetilde{r}_{ij}\right)\right| \cdot I_{\{L_{ij}=0\}} \\
&\leq c_1 \max_{i,j} \left|\widehat{r}_{ij} - \widetilde{r}_{ij}\right| + \max_{1\leq i\leq N} \sum_{j=1}^{N} \left\{\left(\left|s_{\lambda}\left(\widehat{r}_{ij}\right) - \widehat{r}_{ij}\right| + \left|\widehat{r}_{ij} - \widetilde{r}_{ij}\right| + \left|s_{\lambda}\left(\widetilde{r}_{ij}\right) - \widetilde{r}_{ij}\right|\right) \cdot I_{\{L_{ij}=0\}}\right\} \\
&\leq c_1 \max_{i,j} \left|\widehat{r}_{ij} - \widetilde{r}_{ij}\right| + \max_{1\leq i\leq N} \sum_{j=1}^{N} \left\{\left(\left|\widehat{r}_{ij}\right| + \left|\widehat{r}_{ij} - \widetilde{r}_{ij}\right| + \left|\widetilde{r}_{ij}\right|\right) \cdot I_{\{L_{ij}=0\}}\right\} \\
&\leq c_1 \max_{i,j} \left|\widehat{r}_{ij} - \widetilde{r}_{ij}\right| + 2 \max_{1\leq i\leq N} \sum_{j=1}^{N} \left\{\left(\left|\widehat{r}_{ij} - \widetilde{r}_{ij}\right| + \left|\widetilde{r}_{ij}\right|\right) \cdot I_{\{L_{ij}=0\}}\right\}.
\end{aligned}$$

We consider an event $A_{2,1} = \left\{\max_{i,j} \left|\widehat{\sigma}_{ij} - \widetilde{\sigma}_{ij}\right| > Ma_T\right\}$ for some large M, Fan et al. (2011) proved that $P\left(A_{2,1}\right) = O\left(\frac{1}{N^2} + \kappa_1\left(N,T\right)\right)$ in their Lemma A.3. Besides, we define

$$\begin{aligned}
A_{2,2} &= \left\{\max_{1\leq i\leq N} \sum_{j=1}^{N} L_{ij}^0 \left|\widehat{\sigma}_{ij} - \widetilde{\sigma}_{ij}\right| > a_T\right\}, \\
A_{2,3} &= \left\{\max_{1\leq i\leq N} \sum_{j=1}^{N} L_{ij}^0 \left|\widetilde{\sigma}_{ij}\right| > \max_{1\leq i\leq N} \sum_{j=1}^{N} L_{ij}^0 \left|\sigma_{ij}\right| + a_T\right\},
\end{aligned}$$

which are both bounded by $O\left(\kappa_3\left(N,T\right)\right)$ from Assumption 2. Besides, we consider an event $A_{2,4} = \left\{\max_{i,j} \left|\widehat{\sigma}_{ij} - \sigma_{ij}\right| > A\sqrt{\frac{\log N}{T}}\right\}$ for some large A, Fan et al. (2011) proved that $P\left(A_{2,4}\right) = O\left(\frac{1}{N^2}\right)$ in their Lemma A.3. Then consider function $g\left(\sigma_{ij},\sigma_{ii},\sigma_{jj}\right) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$ who has bounded partial derivatives, which yields

$$\left|g\left(\widehat{\sigma}_{ij},\widehat{\sigma}_{ii},\widehat{\sigma}_{jj}\right) - g\left(\widetilde{\sigma}_{ij},\widetilde{\sigma}_{ii},\widetilde{\sigma}_{jj}\right)\right| \leq O\left(d_1\left|\widehat{\sigma}_{ij} - \widetilde{\sigma}_{ij}\right| + d_2\left|\widehat{\sigma}_{ii} - \widetilde{\sigma}_{ii}\right| + d_3\left|\widetilde{\sigma}_{ii} - \widetilde{\sigma}_{jj}\right|\right),$$

also bounded by $O\left(\max_{i,j}\left|\widehat{\sigma}_{ij} - \widetilde{\sigma}_{ij}\right|\right)$. Similarly, $\max_{i,j}\left|\widehat{r}_{ij} - \widetilde{r}_{ij}\right| \leq O\left(\max_{i,j}\left|\widehat{\sigma}_{ij} - \widetilde{\sigma}_{ij}\right|\right)$. Thus

in the set $A_2^c = \Omega - (A_{2,1} \cap A_{2,2} \cap A_{2,3} \cap A_{2,4})$, we have

$$\begin{aligned}
\left\| \widehat{R}_L^{\mathcal{T}} - \widetilde{R}_L^{\mathcal{T}} \right\| &\leq c_1 \max_{i,j} |\widehat{r}_{ij} - \widetilde{r}_{ij}| + 2 \max_{1 \leq i \leq N} \sum_{j=1}^{N} \left\{ (|\widehat{r}_{ij} - \widetilde{r}_{ij}| + |\widetilde{r}_{ij}|) \cdot I_{\{L_{ij}=0\}} \right\} \\
&\leq c_1 A a_T + 2 a_T + 2 \max_{1 \leq i \leq N} |\widetilde{r}_{ij}| \cdot I_{\{L_{ij}=0\}} \\
&\leq c_1 A a_T + 4 a_T + 2 \max_{1 \leq i \leq N} |r_{ij}| \cdot I_{\{L_{ij}=0\}} \\
&\leq c_1 A a_T + 4 a_T + 2 \max_{1 \leq i \leq N} |r_{ij}|^{1-q} |r_{ij}|^{q} \cdot I_{\{L_{ij}=0\}} \\
&\leq A' a_T + 2 \lambda^{1-q} c_0 (N) \leq A'' \left(\lambda^{1-q} c_0 (N) + a_T \right),
\end{aligned}$$

which yields for some large A,

$$P\left(\left\| \widehat{R}_L^{\mathcal{T}} - \widetilde{R}_L^{\mathcal{T}} \right\| > A \left(c_0 (N) \lambda^{1-q} + a_T \right) \right) = O\left(\frac{1}{N^2} + \kappa_1 (N,T) + \kappa_3 (N,T) \right). \tag{17}$$

Be careful it is necessary to let $A_{2,4}^c$ happen since we need to bounded the partial derivatives of g.

For the third part, we have

$$\begin{aligned}
\left\| \widetilde{R}_L^{\mathcal{T}} - R_L^{\mathcal{T}} \right\| &\leq \max_{1 \leq i \leq N} \sum_{j=1}^{N} |s_{L,\lambda} (\widetilde{r}_{ij}) - s_{L,\lambda} (r_{ij})| \\
&\leq \max_{1 \leq i \leq N} \sum_{j=1}^{N} \left\{ |\widetilde{r}_{ij} - r_{ij}| \cdot I_{\{L_{ij}=1\}} + |s_\lambda (\widetilde{r}_{ij}) - s_\lambda (r_{ij})| \cdot I_{\{L_{ij}=0\}} \right\} \\
&\leq \max_{1 \leq i \leq N} \sum_{j=1}^{N} \left\{ |\widetilde{r}_{ij} - r_{ij}| \cdot I_{\{L_{ij}=1\}} + (|s_\lambda (\widetilde{r}_{ij}) - \widetilde{r}_{ij}| + |\widetilde{r}_{ij} - r_{ij}| + |s_\lambda (r_{ij}) - r_{ij}|) \cdot I_{\{L_{ij}=0\}} \right\} \\
&\leq \max_{1 \leq i \leq N} \sum_{j=1}^{N} \left\{ |\widetilde{r}_{ij} - r_{ij}| \cdot I_{\{L_{ij}=1\}} + (|\widetilde{r}_{ij}| + |r_{ij}| + |\widetilde{r}_{ij} - r_{ij}|) \cdot I_{\{L_{ij}=0\}} \right\} \\
&\leq c_1 \cdot \max_{i,j} |\widetilde{r}_{ij} - r_{ij}| + \max_{1 \leq i \leq N} \sum_{j=1}^{N} \left\{ 2 (|\widetilde{r}_{ij}| + |r_{ij}|) \cdot I_{\{L_{ij}=0\}} \right\}.
\end{aligned}$$

Thus in the set $A_3^c = \Omega - (A_{2,3} \cap A_{2,4})$, we have

$$\begin{aligned}
\left\| \widetilde{R}_L^{\mathcal{T}} - R_L^{\mathcal{T}} \right\| &\leq c_1 \cdot \max_{i,j} |\widetilde{r}_{ij} - r_{ij}| + \max_{1 \leq i \leq N} \sum_{j=1}^{N} \left\{ 2 \left(|\widetilde{r}_{ij}| + |r_{ij}| \right) \cdot I_{\{L_{ij}=0\}} \right\} \\
&\leq c_1 (N) A \sqrt{\frac{\log N}{T}} + 4 \max_{1 \leq i \leq N} \sum_{j=1}^{N} |r_{ij}| \cdot I_{\{L_{ij}=0\}} + 2a_T \\
&\leq c_1 (N) A \sqrt{\frac{\log N}{T}} + 4\lambda^{1-q} \max_{1 \leq i \leq N} \sum_{j=1}^{N} |r_{ij}|^q \cdot I_{\{L_{ij}=0\}} + 2a_T \\
&\leq c_1 (N) A \sqrt{\frac{\log N}{T}} + 4\lambda^{1-q} c_0 (N) + 2a_T \\
&\leq A' \left(c_1 (N) \sqrt{\frac{\log N}{T}} + \lambda^{1-q} c_0 (N) + a_T \right),
\end{aligned}$$

which yields that

$$P\left(\left\| \widetilde{\mathbf{\Sigma}}_L^{\mathcal{T}} - \mathbf{\Sigma}_L^{\mathcal{T}} \right\| > A \left(c_1 (N) \sqrt{\frac{\log N}{T}} + c_0 (N) \lambda^{1-q} + a_T \right) \right) = O\left(\frac{1}{N^2} + \kappa_3 (N, T) \right), \quad (18)$$

for some large A.

For the fourth part, we have

$$\begin{aligned}
\left\| R_L^{\mathcal{T}} - R \right\| &\leq \max_{1 \leq i \leq N} \sum_{j=1}^{N} |s_{L,\lambda} (r_{ij}) - r_{ij}| = \max_{1 \leq i \leq N} \sum_{j=1}^{N} |s_\lambda (r_{ij}) - r_{ij}| \cdot I_{\{L_{ij}=0\}} \\
&\leq \max_{1 \leq i \leq N} \sum_{j=1}^{N} |r_{ij}| \cdot I_{\{L_{ij}=0\}} = \max_{1 \leq i \leq N} \sum_{j=1}^{N} |r_{ij}|^q |r_{ij}|^{1-q} \cdot I_{\{L_{ij}=0\}} \quad (19) \\
&\leq \lambda^{1-q} \max_{1 \leq i \leq N} \sum_{j=1}^{N} |r_{ij}|^q I_{\{L_{ij}=0\}} \leq \lambda^{1-q} c_0 (N).
\end{aligned}$$

Now turn back to the first part, which is the only one part including $\widehat{L}$, we have

$$\begin{aligned}
\left\| \widehat{R}_{\widehat{L}}^{\mathcal{T}} - \widehat{R}_L^{\mathcal{T}} \right\| &\leq \max_{1 \leq i \leq N} \sum_{j=1}^{N} \left| s_{\widehat{L},\lambda} (\widehat{r}_{ij}) - s_{L,\lambda} (\widehat{r}_{ij}) \right| \\
&\leq \max_{1 \leq i \leq N} \sum_{j=1}^{N} |s_\lambda (\widehat{r}_{ij}) - \widehat{r}_{ij}| \cdot I_{\left\{ \widehat{L}_{ij} \neq L_{ij} \right\}} \\
&\leq \max_{1 \leq i \leq N} \sum_{j=1}^{N} |\widehat{r}_{ij}| \cdot I_{\left\{ \widehat{L}_{ij} \neq L_{ij} \right\}}.
\end{aligned}$$

We define

$$A_{1,1} = \left\{ \max_{1 \leq i \leq N} \sum_{j=1}^{N} I_{\left\{ L_{ij}=1, \widehat{L}_{ij}=0 \right\}} > a_T c_1 \right\},$$

whose probability is bounded by $O(\kappa_2(N,T))$. Thus in the set $A_1^c = \Omega - (A_{1,1} \cap A_{2,2} \cap A_{2,3})$, we have

$$\begin{aligned}
\max_{1\le i\le N}\sum_{j=1}^{N}|\widehat{r}_{ij}|\cdot I_{\{\widehat{L}_{ij}\neq L_{ij}\}} &= \max_{1\le i\le N}\sum_{j=1}^{N}|\widehat{r}_{ij}|\cdot I_{\{\widehat{L}_{ij}=0,L_{ij}=1\}} + \max_{1\le i\le N}\sum_{j=1}^{N}|\widehat{r}_{ij}|\cdot I_{\{\widehat{L}_{ij}=1,L_{ij}=0\}} \\
&\le a_T c_1 \cdot \max_{i,j}|\widehat{r}_{ij}| + \max_{1\le i\le N}\sum_{j=1}^{N}|\widehat{r}_{ij}|\cdot I_{\{L_{ij}=0\}} \\
&\le a_T c_1 \cdot \max_{i,j}\left(|\widehat{r}_{ij}-\widetilde{r}_{ij}| + |\widetilde{r}_{ij}-r_{ij}| + |r_{ij}|\right) + \sum_{j=1}^{N}\left(|\widehat{r}_{ij}-\widetilde{r}_{ij}| + |\widetilde{r}_{ij}|\right)\cdot I_{\{L_{ij}=0\}} \\
&\le A\left(a_T + \lambda^{1-q}c_0(N)\right),
\end{aligned}$$

which yields that for large A,

$$P\left(\left\|\widehat{R}_{\widehat{L}}^{\mathcal{T}} - \widehat{R}_{L}^{\mathcal{T}}\right\| > A\left(a_T + c_0(N)\lambda^{1-q}\right)\right) = O\left(\kappa_2(N,T) + \kappa_3(N,T)\right). \tag{20}$$

Finally, collecting Equation (20), Equation (17), Equation (18) and Equation (19), we get

$$P\left(\left\|\widehat{R}_{\widehat{L}}^{\mathcal{T}} - R\right\| > A\left(c_1(N)\sqrt{\frac{\log N}{T}} + c_0(N))\lambda^{1-q} + a_T\right)\right) = O\left(\frac{1}{N^2} + \kappa_1(N,T) + \kappa_2(N,T) + \kappa_3(N,T)\right).$$

Now we look bakc to $\mathbf{\Sigma}$, importantly, when $A_{2,4}^c$ happens, we know $\left\|\widehat{D} - D\right\| = O\left(A\sqrt{\frac{\log N}{T}}\right)$ and since

$$\begin{aligned}
\left\|\widehat{\mathbf{\Sigma}}_{\widehat{L}}^{\mathcal{T}} - \mathbf{\Sigma}\right\| &= \left\|\widehat{D}\widehat{R}_{\widehat{L}}^{\mathcal{T}}\widehat{D} - DRD\right\| = \left\|\widehat{D}\left(\widehat{R}_{\widehat{L}}^{\mathcal{T}} - R\right)\widehat{D} + \widehat{D}R\widehat{D} - DRD\right\| \\
&\le \left\|\widehat{D}\left(\widehat{R}_{\widehat{L}}^{\mathcal{T}} - R\right)\widehat{D}\right\| + \left\|\widehat{D}R\widehat{D} - DRD\right\|,
\end{aligned}$$

and the first part is bounded by $O\left(\left\|\widehat{R}_{\widehat{L}}^{\mathcal{T}} - R\right\|\right)$ provided $\sigma_{ii} < M$ and the event $A_{2,4}^c$ happens, as well as the second part

$$\left\|\widehat{D}R\widehat{D} - DRD\right\| \le \left\|\widehat{D}R\left(\widehat{D} - D\right)\right\| + \left\|\left(\widehat{D} - D\right)RD\right\| \le O\left(A\sqrt{\frac{\log N}{T}}\right),$$

we have

$$P\left(\left\|\widehat{\mathbf{\Sigma}}_{\widehat{L}}^{\mathcal{T}} - \mathbf{\Sigma}\right\| > A\left(\left\|\widehat{R}_{\widehat{L}}^{\mathcal{T}} - R\right\| + \sqrt{\frac{\log N}{T}}\right)\right) = O\left(\frac{1}{N^2}\right).$$

In conclusion, since $c_1(N) \to \infty$, we get

$$P\left(\left\|\widehat{\mathbf{\Sigma}}_{\widehat{L}}^{\mathcal{T}} - \mathbf{\Sigma}\right\| > A\left(c_1(N)\sqrt{\frac{\log N}{T}} + c_0(N)\lambda^{1-q} + a_T\right)\right) = O\left(\frac{1}{N^2} + \kappa_1(N,T) + \kappa_2(N,T) + \kappa_3(N,T)\right),$$

which ends the proof. □

B Proof of Theorem 2

Proof. We have the decomposition

$$\left\|\widehat{R}_{\widehat{C}}^{\mathcal{B}}-R\right\| \leq \left\|\widehat{R}_{\widehat{C}}^{\mathcal{B}}-R_{\widehat{C}}^{\mathcal{B}}\right\|+\left\|R_{\widehat{C}}^{\mathcal{B}}-R_{C}^{\mathcal{B}}\right\|+\left\|R_{C}^{\mathcal{B}}-R\right\|.$$

The first part is

$$\begin{aligned}\left\|\widehat{R}_{\widehat{C}}^{\mathcal{B}}-R_{\widehat{C}}^{\mathcal{B}}\right\| &\leq \max_{1\leq i\leq N}\sum_{j=1}^{N}\left|b_{\widehat{C},k}\left(\widehat{r}_{ij}\right)-b_{\widehat{C},k}\left(r_{ij}\right)\right| \\ &= \max_{1\leq i\leq N}\sum_{j=1}^{N}\left|\widehat{r}_{ij}-r_{ij}\right| I_{\left\{i\in S_k^{\widehat{c}_j},j\in S_k^{\widehat{c}_i}\right\}} \\ &\leq k\max_{1\leq i\leq N}\left|\widehat{r}_{ij}-r_{ij}\right|,\end{aligned}$$

thus in the event $B_1^c=\Omega-\left(A_{2,1}\cap A_{2,4}\right)$, one may have $\left\|\widehat{R}_{\widehat{C}}^{\mathcal{B}}-R_{\widehat{C}}^{\mathcal{B}}\right\| \leq O\left(k\sqrt{\frac{\log N}{T}}\right)$, and $P\left(B_1\right)=O\left(\frac{1}{N^2}+\kappa_1\left(N,T\right)\right)$, which yieds

$$P\left(\left\|\widehat{R}_{\widehat{C}}^{\mathcal{B}}-R_{\widehat{C}}^{\mathcal{B}}\right\|>A\cdot k\sqrt{\frac{\log N}{T}}\right)=O\left(\frac{1}{N^2}+\kappa_1\left(N,T\right)\right). \tag{21}$$

We leave the second part, and for the third part, we have

$$\begin{aligned}\left\|R_{C}^{\mathcal{B}}-R\right\| &\leq \max_{1\leq i\leq N}\sum_{j=1}^{N}\left|b_{C,k}\left(r_{ij}\right)-r_{ij}\right| \\ &\leq \max_{1\leq i\leq N}\sum_{j=1}^{N}\left|r_{ij}\right|\left(I_{\left\{i\notin S_k^{c_j}\right\}}+I_{\left\{j\notin S_k^{c_i}\right\}}\right) \\ &\leq \max_{1\leq i\leq N}\sum_{j=1}^{N}\left|r_{ij}\right| I_{\left\{i\notin S_k^{c_j},j\in S_k^{c_i}\right\}}+\max_{1\leq i\leq N}\sum_{j=1}^{N}\left|r_{ij}\right| I_{\left\{j\notin S_k^{c_i}\right\}} \\ &\leq b_1\left(N\right)+b_0\left(N\right)k^{-\alpha}.\end{aligned} \tag{22}$$

For the second part, we have $\left\|R_{\widehat{C}}^{\mathcal{B}}-R_{C}^{\mathcal{B}}\right\| \leq \max_{1\leq i\leq N}\sum_{j=1}^{N}\left|b_{\widehat{C},k}\left(r_{ij}\right)-b_{C,k}\left(r_{ij}\right)\right|$. We know $\left|b_{\widehat{C},k}\left(r_{ij}\right)-b_{C,k}\left(r_{ij}\right)\right|$ can only be 0 or $|r_{ij}|$, thus we can focus on in which situations it becomes $|r_{ij}|$, i.e., when $\widehat{C}$ and C are different enough to make $b_{\widehat{C},k}\left(r_{ij}\right)$ has different quantity from $b_{C,k}\left(r_{ij}\right)$. Specifically, when $i\in S_k^{c_j}$ and $j\in S_k^{c_i}$ however $i\notin S_k^{\widehat{c}_j}$ or $j\in S_k^{\widehat{c}_i}$, or $i\notin S_k^{c_j}$ or

$j \notin S_k^{c_i}$ but $i \in S_k^{\widehat{c}_j}$ and $j \in S_k^{\widehat{c}_i}$, $\left|b_{\widehat{C},k}(r_{ij}) - b_{C,k}(r_{ij})\right|$ becomes $|r_{ij}|$. Thus, one may get

$$\begin{aligned}
\sum_{j=1}^{N}\left|b_{\widehat{C},k}(r_{ij}) - b_{C,k}(r_{ij})\right| &= \sum_{j=1}^{N}|r_{ij}|\, I_{\left\{(i,j)\in S_k^{c_j}\times S_k^{c_i},(i,j)\notin S_k^{\widehat{c}_j}\times S_k^{\widehat{c}_i}\right\}} + \sum_{j=1}^{N}|r_{ij}|\, I_{\left\{(i,j)\notin S_k^{c_j}\times S_k^{c_i},(i,j)\in S_k^{\widehat{c}_j}\times S_k^{\widehat{c}_i}\right\}} \\
&\leq \sum_{j=1}^{N}|r_{ij}|\, I_{\left\{(i,j)\in S_k^{c_j}\times S_k^{c_i},(i,j)\notin S_k^{\widehat{c}_j}\times S_k^{\widehat{c}_i}\right\}} + \sum_{j=1}^{N}|r_{ij}|\, I_{\left\{(i,j)\notin S_k^{c_j}\times S_k^{c_i}\right\}} \\
&\leq \sum_{j=1}^{N}|r_{ij}|\left(I_{\left\{i\in S_k^{c_j},i\notin S_k^{\widehat{c}_j}\right\}} + I_{\left\{j\in S_k^{c_i},j\notin S_k^{\widehat{c}_i}\right\}}\right) + c_1(N) + c_0(N)\,k^{-\alpha} \\
&\leq 2k\sqrt{\frac{\log N}{T}} + b_1(N) + b_0(N)\,k^{-\alpha}
\end{aligned} \tag{23}$$

with probability at least $1 - \kappa_4(N,T)$.

Combining Equation (21), Equation (23) and Equation (22) one may get

$$P\left(\left\|\widehat{R}_{\widehat{C}}^{\mathcal{B}} - R\right\| > A\left(k\sqrt{\frac{\log N}{T}} + b_0(N)\,k^{-\alpha} + b_1(N)\right)\right) = O\left(\frac{1}{N^2} + \kappa_1(N,T) + \kappa_4(N,T)\right)$$

holds for C large enough. Thus similar to the thresholding estimator, provided $\sigma_{ii} < M$, we have

$$P\left(\left\|\widehat{\mathbf{\Sigma}}_{\widehat{C}}^{\mathcal{B}} - \mathbf{\Sigma}\right\| > A\left(k\sqrt{\frac{\log N}{T}} + b_0(N)\,k^{-\alpha} + b_1(N)\right)\right) = O\left(\frac{1}{N^2} + \kappa_1(N,T) + \kappa_4(N,T)\right),$$

where one should note that $k = k_N \to \infty$ is a common setting (see the remark after Theorem 2). □