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

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

September 20, 2024

Abstract

We propose two novel frameworks that incorporate auxiliary network information into the estimation of large covariance matrices —Network Guided Thresholding and Network Guided Banding. Compared with existing methods which either ignores network information (e.g., the thresholding or shrinkage estimator) or imposes overly restrictive structure (e.g., the banding estimator), our proposed estimators take advantage of the auxiliary network information available in the era of big data. Our two estimators are designed to adapt to the specific features of the auxiliary network information at hand and to different structures of the covariance matrix. We show that both Network Guided estimators have great convergence rates over a larger class of sparse covariance matrices. Simulation studies indicate that these estimators generally outperform other purely statistical methods,

[*] The authors would like to thank Xiaohong Chen, Hashem Pesaran, Cheng Hsiao, Seok Young Hong, Jeroen Dalderop for their useful comments. The authors also thank Xiang Lu for his excellent research assistance.

[†] School of Finance, University of Nankai. Author email: sg751_shuyige@outlook.com
[‡] School of Economics, Peking University. Author email: lishaoran@pku.edu.cn
[§] Faculty of Economics, University of Cambridge. Author email: obl20@cam.ac.uk
[¶] Department of Economics, UCL. Author email: weiguang.liu@ucl.ac.uk
[‖] Mathematical Institute, University of Oxford. Author email: wen.su@maths.ox.ac.uk

particularly when the true covariance matrix is sparse and the auxiliary network provides reliable information. Empirically, we apply our methods to estimate the covariance matrix of asset returns using various forms of auxiliary network data to construct the Global Minimum Variance (GMV) and Mean-Variance Optimal (MVO) portfolios, which deliver better out-of-sample results compared to competitors.

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. Suppose that we have T observations $(\boldsymbol{X}_t : t = 1, \cdots, T)$ of an N-dimensional vectors $\boldsymbol{X}_t = (X_{1t}, \ldots, X_{Nt})^\intercal$ and we are interested in estimating the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{X}_t$. For many important empirical questions, the dimensionality N of the random vector under inspection is large and often larger than the sample size T, making the estimation of the covariance matrix challenging. For example, in portfolio management, we often need to estimate the covariance matrix of a large number of asset returns with a relatively short time series span. It is well known that the sample covariance matrix is ill-conditioned when the dimension exceeds the sample size. In that case, a consistent estimator for the covariance matrix can still be constructed using regularization techniques if the covariance matrix has additional structures. Two classes of covariance matrices with additional structures that have been extensively studied are bandable and sparse covariance matrices. This paper aims to generalize these two classes of covariance matrices and improve the associated estimators when auxiliary network information is available.

In the era of big data, we have easy access to auxiliary information beyond the observations of $\{\boldsymbol{X}_t\}_{t=1}^T$ that could potentially help us learn about the underlying structure of the covariance matrix (i.e., interconnectedness among entities).[1] Consider the case of equity return covariance. Israelsen (2016) found that stocks covered by similar sets of analysts co-move a lot. Ge et al. (2022) documented 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) constructed peer groups within which firms are fundamentally similar and, therefore, tend to co-move. All of the aforementioned auxiliary network information may help us understand the connectivity structure among stocks. However, the current literature either completely ignores this type of readily available auxiliary information (e.g., thresholding and

[1] Throughout this paper, we use the terms interconnectedness, network, connectivity, and linkages interchangeably.

shrinkage estimators) or utilizes some simple network structure under very restrictive settings (e.g., banding and tapering estimators).

In this paper, we propose a novel framework for incorporating auxiliary 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 methods for application and derive their theories accordingly.

The first method is *Network Guided Thresholding*. The method is applicable when auxiliary information identifies the location of large elements in the covariance matrix. Industry information is an example of such auxiliary information as it implies a block-diagonal network where every node is equally connected within an industry. The original series of thresholding methods (Bickel and Levina (2008a), Cai and Liu (2011), Fan et al. (2013)) retain the large elements in sample covariance and shrink the rest based on statistical information under the assumption of sparsity (or conditional sparsity). These thresholding estimators do not utilize any location information. In contrast, we use auxiliary network information to identify the location of large elements. We show that our proposed method improves efficiency by using additional information, and is applicable to a larger class of covariance matrices with a more refined notion of sparsity. With this additional information, we retain the large elements identified by the auxiliary network information in the sample covariance and then apply generalized thresholding to the remaining elements. The work closest to our method is Fan et al. (2016), where the authors utilize sector information and apply location-based thresholding by assuming that within-sector correlations are large and across-sector correlations are small and can therefore be ignored. However, the residual correlation structure of the factor model is not as simple as a block diagonal assumed by Fan et al. (2016), and our method accommodates more general 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 by using auxiliary network information to distinguish between large and small elements and to quantify their behaviors separately. We show the consistency of the estimator in the operator norm under certain conditions uniformly over the class of matrices that satisfy our sparsity condition. Next, we show that the Network Guided Thresholding estimator achieves a better convergence rate compared to Bickel and Levina (2008a), particularly when the auxiliary information is of high quality.

The second method we propose is called *Network Guided Banding*. Bickel and Levina (2008b) showed that uniformly over a class of the "approximately bandable" matrices, the banding estimator shows a superior convergence rate. From their definition, the elements of such a matrix are 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. However, 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. Compared with the original banding estimator, one key feature of this new Network Guided Banding method is that it is permutation-invariant. Unlike the Network Guided Thresholding, this method requires auxiliary network information to reveal the relative importance of neighbors for each node to be applicable. Again, in the case of the equity return covariance matrix, the auxiliary sector/industry network provides unweighted linkage information, meaning it simply indicates whether a pair of stocks are linked without quantifying its strength. In that case, we may apply the Network Guided Thresholding, but not the Network Guided Banding. In contrast, the analyst co-coverage network (Israelsen (2016)), news co-mentioning network (Ge et al. (2022)), and text-based product network (Hoberg and Phillips (2016)) are all weighted, meaning they assign different levels of strength to each connection. This allows us to use the degree of connectivity to rank the relative importance of neighbors for each node. For example, in the news co-mentioning network, firms mentioned together in the same piece of news are considered linked, and the frequency of these co-mentions can be used as a proxy for the strength of the linkages, thereby helping to rank their relative importance (Scherbina and Schlusche (2015), Schwenkler and Zheng (2019), Ge et al. (2022)). With these available auxiliary network information, we can apply both the Network Guided Thresholding and the Network Guided Banding. We show the consistency of the estimator in the operator norm uniformly over the class of matrices that satisfy a generalised bandable condition. We also show that the Network Guided Banding estimator achieves optimal rate as in Bickel and Levina (2008b) over a larger class of bandable matrices.

For both estimators, our theory allows for measurement errors when using the auxiliary information to identify the important elements, which are crucial in practice. For the Network Guided Thresholding estimator, we allow for errors in identifying the location of the relatively large elements. For the Network Guided Banding estimator, we also allow for asymmetry in the relative importance. Details and a discussion of measurement errors are given in Section 3.

In Monte Carlo experiments, we assume that asset returns are generated from a factor model where the true covariance matrix of idiosyncratic returns is sparse. For the application of the Network-Guided methods, we generate auxiliary information of varying quality. We then compare the performance of our two Network-Guided estimators with a set of competitors, including conventional thresholding and linear and nonlinear shrinkage approaches. The simulation results show that, as long as the auxiliary network information is of reasonable quality, our Network-Guided estimators consistently demonstrate superior finite sample performance compared to all benchmark methods. The relative performance of the two proposed estimators depends on the structure of the true covariance matrix and the characteristics of the auxiliary information.

Empirically, we apply our newly proposed Network-Guided estimators to construct Global Minimum Variance (GMV) and Mean-Variance Optimal (MVO) portfolios by incorporating auxiliary information. We utilize data from the Chinese stock market and adopt the Chinese CH-4 factor model from Liu et al. (2019) to analyze asset returns. We explore various sources of auxiliary information to identify linkages among listed stocks. Our first candidate is the news co-mention network. Similarly to Ge et al. (2023), we consider two types of news-implied linkages: co-mentions within the same passage and co-mentions within the same sentence. Since the frequency of co-mentions may serve as a proxy for the strength of linkage, these two types of news co-mention networks are weighted and can therefore be used as input for both the Network Guided Thresholding and Banding estimators. Additionally, we explore the analyst co-coverage network in China. This auxiliary information quantifies the strength of connectivity between two entities through a continuous measure (the number of analysts who co-cover the two entities), making it applicable to both of our proposed estimators. Finally, we consider the traditional industry classification. 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. We also compare the out-of-sample performance of several conventional statistical covariance matrix estimation methods. Our comparison spans different sets of constituent stocks, including HS300, CSI500, and CSI800. Our findings consistently indicate that incorporating auxiliary network information largely enhances the out-of-sample performance of GMV and MVO portfolios.

Literature Review: A growing number of methods have been proposed in the literature to study covariance matrix estimation when the dimensionality is large. Bickel and Levina (2008a) developed a theory for universal thresholding, which assumes the diagonal of the covariance matrix is uniformly bounded. Cai and Liu (2011) relaxed the uniform boundedness assumption and proposed an adaptive thresholding estimator with entry-adaptive thresholds. Fan et al. (2013) argued that common factors should be extracted first before applying thresholding when there are "extremely spiked" eigenvalues in the covariance matrix, making it conditionally sparse. Another strand of literature has attempted to correct the spectrum of the sample covariance matrix instead of imposing sparsity. For instance, Ledoit and Wolf (2004) and Ledoit and Wolf (2012) proposed linear and nonlinear shrinkage estimators that apply shrinkage to the eigenvalues of the sample covariance matrix. The linear shrinkage estimator combines the sample covariance matrix with a well-conditioned target matrix, such as the identity matrix. The nonlinear shrinkage estimator adjusts the eigenvalues using the asymptotic Marchenko–Pastur distribution. A key advantage of shrinkage estimators is that they are well-conditioned, while estimators based on sparsity often require selecting tuning parameters to ensure positive definiteness. However, shrinkage estimators may be less effective when the true covariance matrix is sparse. These aforementioned pure statistical methods rely solely on observations of $\{\boldsymbol{X}_t\}_{t=1}^T$ and completely disregard any network information that might be available from auxiliary data sources. There is also literature that embraces the use of very simple network information. Bickel and Levina (2008b) proposed banding and tapering estimators, where indices are ordered, and elements in the covariance matrix decrease in magnitude as one moves away from the diagonal. They demonstrated that the banding estimator achieves a superior convergence rate by leveraging the bandable structure. However, the underlying structure of these bandable matrices is quite restrictive, making the banding estimator inapplicable in most scenarios.

The novelty of this paper lies in augmenting the estimation of large covariance matrices with auxiliary network information. Depending on the features of the auxiliary network information at hand, we offer two distinct Network-Guided methods for application. We derive the corresponding theories and demonstrate that both Network-Guided estimators exhibit strong theoretical and numerical properties, as well as good empirical performance.

Although this paper focuses on applying network information to the estimation of large static covariance matrices, the same idea can be extended to the estimation of large dynamic covariance matrices. For instance, dynamic network information could be effectively incorporated into the conditioning information set, as suggested in Chen et al. (2019).

The remainder of this paper is structured as follows. Section 2 introduces the Network Guided Thresholding estimator and the Network Guided Banding estimator. In Section 3, we lay down the assumptions and derive the convergence results. Section 4 presents simulation studies that compare our proposed estimators with established baseline methods, while Section 5 provides an empirical application. Finally, Section 6 concludes. The Appendix contains proofs. The ready-to-use Python code can be found at `www.lishaoran.com`.

Notation: For vector $\boldsymbol{a} \in \mathbb{R}^d$, $\|\boldsymbol{a}\|$ stands for the Euclidean norm, i.e., $\|\boldsymbol{a}\| = (a_1^2 + \cdots, a_d^2)^{1/2}$. For matrix $A = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m) \in \mathbb{R}^{m\times d}$, $\|A\|_F$ denotes the matrix Frobenius norm, i.e., $\|A\|_F = (\|\boldsymbol{a}_1\|^2 + \cdots + \|\boldsymbol{a}_m\|^2)^{1/2}$; $\|A\| = \inf\left\{c > 0 : \|Ax\| \leq c\,\|x\|\,, \text{ for all } x \in \mathbb{R}^d\right\}$ is the operator norm. For two real-valued sequences $\{a_T\}$ and $\{b_T\}$, $a_T = o\left(b_T\right)$ implies $a_T/b_T \to 0$ when $T \to \infty$; $a_T = O\left(b_T\right)$ implies there exists some constant A, s.t. $a_T \leq Ab_T$ for all T; $a_T \asymp b_T$ means $0 < c < \frac{a_T}{b_T} < C < \infty$. We use $[a_{ij}]_{m\times n}$ to denote an $m \times n$ matrix whose (i, j)-th element is a_{ij} and $\boldsymbol{J}_{N\times N}$ to denote a $N \times N$ unit matrix.

2 Model Setup

Suppose that we have observations $\boldsymbol{X}_t = (X_{1t}, \ldots, X_{Nt})^\intercal$, $t = 1, \ldots, T$ of a N-dimensional random vector $\boldsymbol{X}_t$ with mean $E(\boldsymbol{X}_t) = \boldsymbol{\mu}$ and variance $E((\boldsymbol{X}_t - \boldsymbol{\mu})(\boldsymbol{X}_t - \boldsymbol{\mu})^\intercal) = \boldsymbol{\Sigma} = [\sigma_{ij}]_{N\times N}$. 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. In the following, we propose two theoretical frameworks for augmenting large covariance matrix estimation with auxiliary network information. The choice of framework depends on the specific features of the available auxiliary network information.

2.1 Network Guided Thresholding

When the auxiliary information identifies the location of large elements in the covariance matrix, we propose the following *Network Guided Thresholding* method. Recall the following definition of uniformity class of sparse covariance matrices given by Bickel and Levina (2008a),

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

Here, the sparsity pattern parameter q and sparsity magnitude parameter $c_0(N)$ jointly characterize the requirements on the off-diagonal elements for the covariance matrix to be sparse. The reason why we call q the sparsity pattern parameter and $c_0(N)$ the sparsity magnitude parameter can be seen from the following two types of sparse covariance matrices, (1) one with only a small number of non-zero off-diagonal elements with large magnitude, and (2) one with off-diagonal elements that are all non-zero but small in magnitude. For $q = 0$, the expression $\sum_{j=1}^{N} |\sigma_{ij}|^q$ counts the number of non-zero off-diagonal elements in each row and ignores the magnitude of the non-zero elements. Hence, for the first type of covariance matrix, $c_0(N)$ can be small as there are only a small number of non-zero elements, but $c_0(N)$ would have to be N for the second type. On the other hand, if we fix a q that is close to 1, the second type of sparse covariance matrix potentially has a much smaller magnitude $c_0(N)$. Intuitively, while q indicates which type or pattern of sparsity is in place, $c_0(N)$ tells us how sparse the covariance matrix is with respect to this pattern.

With these two examples in mind, it is now clear that one inconvenience of $\mathcal{U}_\tau$ is that it tries to capture two types of sparsity patterns with one set of parameters $(q, c_0(N))$. Cai and Zhou (2012) showed that the minimax optimal convergence rate is faster for estimating the sparse covariance matrix of the first type and if the sparsity magnitude parameter $c_0(N)$ is smaller. By incorporating auxiliary information about the sparsity pattern, particularly regarding the location of the few large elements, we can treat the two types of sparsity separately with our Network Guided Thresholding estimator, which is shown to have superior performance both in theory and in numerical experiments. We consider the following extension to the uniformity class in Equation 2 by treating the two sparsity patterns separately via the following *Location Indicator Matrix*,

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

This matrix is defined through the correlation coefficients matrix $R = [r_{ij}]_{N \times N}$ and an observation level parameter $l > 0$. r_{ij} represents the correlation coefficient between assets i and j. For $s \in \{0, 1\}$, we use the shorthand notation $L^s_{ij} = I_{\{L_{ij}=s\}}$. L^1 is the Location Indicator Matrix of large elements that exceed the observation level l in absolute value in the correlation matrix, and L^0 is the Location Indicator Matrix of small elements that are below l in the correlation matrix. For example, $l = 0.1$ means that the Location Indicator Matrix helps identify those pairs with $|r_{ij}| > 0.1$. It is obvious that $L^1 = L$ and $L^0 = \boldsymbol{J}_{N,N} - L^1$, where $\boldsymbol{J}_{N,N}$ is a unit matrix. We then define the following uniformity class:

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

where $D = \text{diag}\{\sqrt{\sigma_{11}}, \cdots, \sqrt{\sigma_{NN}}\}$, R is the correlation coefficient matrix and L is the location indicator matrix corresponding to R. By treating large elements ((i, j) pairs such that $L^1_{ij} = 1$) and small elements ((i, j) pairs such that $L^0_{ij} = 1$) differently, this uniformity class allows for separate restrictions on the number of large elements and the growth rate of the remaining small elements. In the special case of $l = 1$, the class $\mathcal{U}_1$ reduces to the traditional uniformity class $\mathcal{U}_\tau$.

Although the Location Indicator Matrix $L = L(R, l)$ depends on the observation level l, we can choose a $l^* = l^*(R)$ such that

$$\max_{1 \leq i \leq N} \sum_j L^1_{ij} \asymp \max_{1 \leq i \leq N} \sum_j L^0_{ij} |r_{ij}|^q, \tag{5}$$

and $L^* = L(R, l^*)$. Then the class $\mathcal{U}_1(q, c_0, c_1, M)$ can be defined without specifying the observation level l by replacing L with L^* in Equation 4.

We propose the following *Network Guided Thresholding Estimator* for $\boldsymbol{\Sigma} \in \mathcal{U}_1$,

$$T_{L,\lambda}\left(\widehat{R}\right) = \left[s_{L,\lambda}\left(\widehat{\sigma}_{ij}/\sqrt{\widehat{\sigma}_{ii}\widehat{\sigma}_{jj}}\right)\right]_{N\times N} \quad \text{with} \quad s_{L,\lambda}(r_{ij}) = r_{ij} I_{\{L_{ij}=1\}} + s_\lambda(r_{ij}) I_{\{L_{ij}=0\}}, \tag{6}$$

where $s_\lambda(x)$ is the generalized thresholding operator.[2] The corresponding estimator for the covariance matrix is

$$\widehat{\boldsymbol{\Sigma}}_L^{\mathcal{T}} := \widehat{D} T_{L,\lambda}\left(\widehat{R}\right)\widehat{D}.$$

Notice that the Location Indicator Matrix is unobserved, and we need to use auxiliary information to estimate L and, therefore, L^s for $s \in \{0,1\}$. Denote the estimator of the Location Indicator Matrix as $\widehat{L}$. Then we have the feasible Network Guided Thresholding Estimator as follows:

$$\widehat{\boldsymbol{\Sigma}}_{\widehat{L}}^{\mathcal{T}} := \widehat{D} T_{\widehat{L},\lambda}\left(\widehat{R}\right)\widehat{D}, \tag{7}$$

where we use the estimated Location Indication Matrix $\widehat{L}$. It is restrictive to assume that the auxiliary information precisely identifies the location of the few large elements, and therefore, we allow for estimation errors in $\widehat{L}$ as discussed later in Assumption 3.

In this framework, we have two thresholding parameters, λ and l. λ is an empirical tuning parameter used to smooth out small estimates, commonly seen in traditional thresholding methods. In contrast, l is an additional parameter that links the Location Indicator Matrix L with the population covariance matrix and does not directly enter the estimator. In addition, to obtain a better convergence rate, it is ideal to set $\lambda \geq l$, which can further shrink the estimates of those pairs with $L_{ij} = 0$.

2.2 Network Guided Banding

When the auxiliary data reveal the relative importance of neighbors for each node, additional information can be extracted, potentially improving the convergence rate. In such cases, we propose the *Network Guided Banding* method. Recall that the original Banding and Tapering methods are effective when there is a natural "order" or "distance" among variables; Bickel and

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

Levina (2008b) considered the following uniformity class of covariance matrices:

$$\mathcal{U}_b(\varepsilon, \alpha, c) = \left\{ \mathbf{\Sigma} : \max_j \sum_{i:|i-j|>k} |\sigma_{ij}| \leq c k^{-\alpha} \text{ for all } k, \text{ and } 0 < \varepsilon \leq \rho_{\min}(\mathbf{\Sigma}) \leq \rho_{\max}(\mathbf{\Sigma}) \leq \frac{1}{\varepsilon} \right\}, \tag{8}$$

where $\rho_{\min}(\cdot)$ and $\rho_{\max}(\cdot)$ give the minimal and maximal eigenvalues of a matrix; ε is a positive constant independent of N; $\alpha > 0$ captures the dependence structure in the class and $0 \leq k < N$. Bickel and Levina (2008b) showed that when this banding condition is satisfied, a better convergence rate can be achieved by taking advantage of the underlying structure.

The original Banding and Tapering methods are primarily applicable to time series, which have a natural ordering.[3] In many practical applications, however, entities are not ordered. Therefore, we extend their method by allowing for a more general underlying connectivity (network) structure, making these methods applicable to a wider range of covariance matrices. To begin, we define a new order $\langle \{1, \cdots, N\}, \succ \rangle$ for an N-dimensional vector $\boldsymbol{a} = (a_1, \ldots, a_N)^\intercal$ with distinct elements as follows:

$$i \succ j \iff a_i > a_j.$$

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 location 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 8) by directly comparing the relative magnitudes (not a real "distance") of entries for each row of a matrix. We use the correlation counterpart of Equation 8 for a fair comparison under heteroskedasticity.

[3] For a permuted matrix, there are also methodologies for estimating the banding structure. For example, see Giraud et al. (2023).

Specifically, we consider a generalized uniformity class of covariance matrices:

$$\mathcal{U}_2(\varepsilon,\alpha,b_0,M)=\left\{\boldsymbol{\Sigma}=DRD:\max_i\sigma_{ii}<M,\ \sum_{j\notin S_k^{\operatorname{abs}(r_i)}}|r_{ij}|<b_0\left(N\right)k^{-\alpha}\text{ for all } i,k,\ \text{and}\ \rho_{\max}\left(R\right)\leq\frac{1}{\varepsilon}\right\},\quad(9)$$

where r_i is the i-th column (row) of correlation matrix R, and $\operatorname{abs}\left(r_i\right)=\left(\left|r_{i1}\right|,\cdots,\left|r_{iN}\right|\right)$ gives the absolute values of the correlation coefficients. $S_k^{\operatorname{abs}(r_i)}$ gives the set of indexes of the k-biggest elements. Notice that when $k=1$, $S_k^{\operatorname{abs}(r_i)}=\{i\}$ as the self-correlation is always the largest. When $k>1$, $S_k^{\operatorname{abs}(r_i)}$ includes i itself and the set of $k-1$ nearest neighbours. Essentially, the correlations between non-neighboring pairs need to be small under Equation 9. Compared with the original banding, this method is permutation-invariant and accommodates a more general connectivity (network) structure.

A *Relative Importance Matrix* $C=[C_{ij}]_{N\times N}$ can be defined as follows: For each row r_i of a correlation matrix $R=[r_{ij}]_{N\times N}$, we define c_{ij} to be the rank of $|r_{ij}|$ in the set $\operatorname{abs}\left(r_i\right)$. For example, if $|r_{ij}|$ is the smallest in $\operatorname{abs}\left(r_i\right)$, then $c_{ij}=1$; if $|r_{ij}|$ is the second smallest then $c_{ij}=2$; and c_{ii} is always N. In this way, we can ensure that $S_k^{c_i}=S_k^{\operatorname{abs}(r_i)}$ for all k, and C is unique for a given R. For a correlation matrix $R=[r_{ij}]_{N\times N}$, we define a Relative Importance Matrix $C=[C_{ij}]_{N\times N}$ with non-negative elements. For each row i (or column), the elements of $C_i=(C_{i1},\ldots,C_{iN})$ provide rankings based on the magnitude from $\operatorname{abs}(r_i)$. Given the Relative Importance Matrix C, we define the Network Guided Banding Estimator as follows:

$$\begin{aligned}&\widehat{\boldsymbol{\Sigma}}_C^{\mathcal{B}}=\widehat{D}B_{C,k}\left(\widehat{R}\right)\widehat{D}\quad\text{with}\quad B_{C,k}\left(\widehat{R}\right)=\left[b_{C,k}\left(\widehat{r}_{ij}\right)\right]_{N\times N},\\&b_{C,k}\left(r_{ij}\right)=r_{ij}I_{\left\{i\in S_k^{c_j},j\in S_k^{c_i}\right\}}=\begin{cases}r_{ij}, & i\in S_k^{c_j}\text{ and } j\in S_k^{c_i},\\0, & \text{otherwise}.\end{cases}\end{aligned}\qquad(10)$$

We do not observe the Relative Importance Matrix C. We need to form an estimator $\widehat{C}$ utilizing an auxiliary dataset. The feasible estimator is $\widehat{\boldsymbol{\Sigma}}_{\widehat{C}}^{\mathcal{B}}$.

It is noteworthy that $\widehat{\boldsymbol{\Sigma}}_{\widehat{C}}^{\mathcal{B}}$ is not strictly a banding or tapering estimator because the k-

neighbour relationship could be asymmetric, i.e., $i \in S_k^{c_j} \not\Leftrightarrow j \in S_k^{c_i}$ for a symmetric matrix R. The asymmetry is important for modeling the correlation structure of financial assets. For example, in a scale-free network, which is commonly observed in financial markets, a central asset may be connected to many other assets, making it one of the k-nearest neighbors for many nodes. However, the reverse is not necessarily true. Additionally, to implement the Network Guided Banding estimation, we need to determine k, i.e., the number of neighbors each node has. In practice, the optimal k can be selected via cross-validation.

2.3 Conditional Sparsity

Asset returns are exposed to common factor risks, leading to high co-movements in their returns, making it inappropriate to assume sparsity for the return covariance matrix. Therefore, we make a conditional sparsity assumption and adopt the following factor structure for asset returns:

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{f}_{1,t} + \boldsymbol{\beta}_2 \boldsymbol{f}_{2,t} + \cdots + \boldsymbol{\beta}_K \boldsymbol{f}_{K,t} + \boldsymbol{u}_t \\ &= \boldsymbol{\beta}_0 + \boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{u}_t, \end{aligned} \tag{11}$$

$t = 1, 2, \cdots, T$, where $\boldsymbol{y}_t$ is the $N \times 1$ assets return at time t, $\boldsymbol{f}_t$ is the $K \times 1$ vector of observable factor returns, $\boldsymbol{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_K)$ is the $N \times K$ factor loading matrix, $\boldsymbol{\beta}_0$ is the mean vector, and $\boldsymbol{u}_t$ is the zero-mean idiosyncratic term, which may contain cross-sectional dependencies. Factor models have long been employed in modeling asset returns; see, for example, Ross (1976), Chamberlain and Rothschild (1982), Fama and French (1993), etc.

Under the factor structure and the assumption that factors $\boldsymbol{f}_t$ and idiosyncratic returns $\boldsymbol{u}_t$ are independent, the covariance matrix of the returns can be decomposed into

$$\mathrm{Cov}(\boldsymbol{y}_t, \boldsymbol{y}_t) = \boldsymbol{\Sigma}_y = \boldsymbol{B}\boldsymbol{\Sigma}_f \boldsymbol{B}^\intercal + \boldsymbol{\Sigma}_u. \tag{12}$$

We follow Fan et al. (2011) and assume $\boldsymbol{\Sigma}_u$ to be sparse, i.e., the covariance matrix of returns $\boldsymbol{\Sigma}_y$ is conditionally sparse. Given that the factors are observable, the coefficients from Equation 11 can be easily estimated using ordinary least squares (OLS). After obtaining $\widehat{\boldsymbol{B}}$, the covariance from common factor component is $\widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^\intercal$ where

$$\widehat{\boldsymbol{\Sigma}}_f = \frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{f}_t - \overline{\boldsymbol{f}}\right)\left(\boldsymbol{f}_t - \overline{\boldsymbol{f}}\right)^\intercal, \quad \overline{\boldsymbol{f}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_t. \tag{13}$$

The challenge of estimating the return covariance matrix lies in the estimation of $\boldsymbol{\Sigma}_u$. With the OLS estimates, we can calculate the 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 $\widehat{\boldsymbol{\Sigma}}_u = \frac{1}{T}\sum_{t=1}^{T}\widehat{\boldsymbol{u}}_t\widehat{\boldsymbol{u}}_t^\intercal = (\widehat{\sigma}_{ij})_{N\times N}$. Depending on whether the auxiliary dataset reveals weighted or unweighted network information, we can obtain estimates of the Location Indicator Matrix or the Relative Importance Matrix. Then, we can apply the appropriate network-guided approach to obtain the feasible Network Guided Thresholding Estimator $\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_{u,\widehat{L}}$ or feasible Network Guided Banding Estimator $\widehat{\boldsymbol{\Sigma}}^{\mathcal{B}}_{u,\widehat{C}}$.

3 Main Results

3.1 Theoretical Results

In this subsection, we introduce 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 to denote any sufficiently large constant that does not depend on N and T.

Assumption 1. *(a) Sequence $\{\boldsymbol{u}_t, \boldsymbol{f}_t\}$ is strong stationary, α-mixing and ergodic, with $\boldsymbol{u}_t$ having zero means and covariance matrix $\boldsymbol{\Sigma}_u$. 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 that do not depend on N (thus uniformly mixing over N). Additionally, there are constants $\underline{c}, \overline{c}$, s.t., $0 < \underline{c} < \inf_{i,j} Var(u_{it}u_{jt}) < \sup_{i,j} Var(u_{it}u_{jt}) < \overline{c}$, $\underline{c} < \rho_{\min}\left(\boldsymbol{\Sigma}_u\right) < \rho_{\max}\left(\boldsymbol{\Sigma}_u\right) < \overline{c}$.*

(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 for any $x > 0$, we have $\sup_i P\left(|u_{it}| > x\right) \leq \exp\left\{-\phi_3 x^{\phi_4}\right\}$.

(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, $P\left(\max_i \frac{1}{T}\sum_{t=1}^{T}|u_{it} - \widehat{u}_{it}|^2 > Aa_T^2\right) \leq O\left(\kappa_1(N,T)\right)$ and $P\left(\max_{i,t}|u_{it} - \widehat{u}_{it}| > A\right) = o(1)$.

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

Remark: The first part of condition (a) allows the idiosyncratic components to be weakly dependent, while the second part requires the well-posedness of $\mathbf{\Sigma}^{-1}$. Condition (b) ensures that the distributions of u_{it} have exponential-type tails, which allows us to apply the large deviation theory. Condition (c) facilitates the study of the estimated error covariance matrix when direct observations are not available. Conditions (a), (b), and (c) correspond to Assumptions 2.1, 2.2 in Fan et al. (2011). Condition (d) is an additional assumption to ensure good asymptotic properties, as proposed in Theorem 2.1 of Fan et al. (2011). Given these assumptions, one can easily show that

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

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

In addition, for the factor model, we borrow the following assumptions from Fan et al. (2011).

Assumption 2. *(a) There exists a constant $A > 0$, s.t., $E\left(y_{it}^2\right) < A$, $E\left(f_{it}^2\right) < A$, and $\beta_{ij} < A$ for all i, j, t. Besides, there exists a constant ϕ_5 which satisfies $3\phi_5^{-1} + \phi_2^{-1} > 1$ and $\phi_6 > 0$, s.t.,*

$$\sup_i P\left(|f_{it}| > x\right) \leq \exp\left\{-\left(x/\phi_6\right)^{\phi_5}\right\}.$$

(b) $\rho_{\min}\left(\mathbf{\Sigma}_f\right) > 0$ uniformly. In addition, there exists a postive definite matrix $\mathbf{\Omega}$, s.t., $\left\|\frac{1}{N}\boldsymbol{B}^\intercal\boldsymbol{B} - \mathbf{\Omega}\right\| = o(1)$.

(c) $K = o(N)$*,* $K^4 (\log N)^2 = o(T)$*, and* $(\log N)^{2/\phi_2 - 1} = o(T)$*.*

Remark: Condition (a) ensures that the factors have finite variance and that the factor loadings are bounded. The exponential-type tail condition allows us to apply the Bernstein-type inequality. The first part of condition (b) ensures that $\rho_{\min}(\mathbf{\Sigma}_y)$ is bounded away from zero, while the second part implies that the factors should be pervasive. These conditions for factors and loadings are easily satisfied when K is fixed and finite.

Note that Assumption 1 and Assumption 2 are common assumptions that we need to impose for both types of network-guided estimators. In the following subsections, we outline the assumptions required to establish the asymptotic properties of each network-guided estimator and present the corresponding theoretical results.

3.1.1 Network Guided Thresholding Estimator

We assume that $\mathbf{\Sigma}_u \in \mathcal{U}_1(q, c_0, c_1, M)$ defined in Equation 4, which extends the sparsity condition from Bickel and Levina (2008a). To derive the asymptotic results for our Network Guided Thresholding estimator $\widehat{\mathbf{\Sigma}}_{\widehat{L}}^{\mathcal{T}} = \widehat{D} T_{\widehat{L},\lambda}\left(\widehat{R}\right)\widehat{D}$, apart from the common Assumption 1 and Assumption 2, the following additional assumptions need to be imposed.

Assumption 3. *(a) The function* s_λ *satisfies (i)* $|s_\lambda(t) - t| \leq \lambda$*, (ii)* $|s_\lambda(t)| \leq t$*, and (iii)* $s_\lambda(t) = 0$ *for* $|t| \leq \lambda$*;*

(b) Suppose $\widehat{L}$ *is the estimator of* L*, we assume*

$$P\left(\max_{1\leq i\leq N}\sum_{j=1}^{N} I_{\left\{L_{ij}=1,\widehat{L}_{ij}=0\right\}} > \varrho_T c_1(N)\right) = O(\kappa_2(N,T)),$$
$$P\left(\max_{1\leq i\leq N}\sum_{j=1}^{N} I_{\left\{L_{ij}=0,\widehat{L}_{ij}=1\right\}} > \varrho_T c_1(N)\right) = O(\kappa_2(N,T)),$$
$$P\left(\max_{1\leq i\leq N}\sum_{j=1}^{N} I_{\left\{L_{ij}=0,\widehat{L}_{ij}=0,|\widehat{r}_{ij}|>l\right\}} > \varrho_T c_1(N)\right) = O(\kappa_2(N,T)),$$

for some $\kappa_2(N,T) = o(1)$ *and* $\varrho_T \to 0$*.*

Remark: Condition (a) is condition (iii) in Rothman et al. (2009), which is a common assumption for thresholding estimation. Condition (b) restricts the number of misclassified elements, ensuring that the auxiliary network information is of relatively good quality. In other words, the false-positive (FP) and false-negative (FN) probabilities, defined in Equation 14 below, cannot be too large.

$$\mathrm{FP} = P\left(\widehat{L}_{ij} = 1 \middle| L_{ij} = 0\right), \quad \mathrm{FN} = P\left(\widehat{L}_{ij} = 0 \middle| L_{ij} = 1\right). \tag{14}$$

Both FP and FN probabilities represent the quality of the auxiliary information. $P\left(\max_{1\le i\le N}\sum_{j=1}^{N} I_{\left\{\widehat{L}_{ij}=0, L_{ij}=0, |\widehat{r}_{ij}|>l\right\}} > \varrho_T c_1(N)\right)$ restricts the joint distribution of $\widehat{r}_{ij}$ and $\widehat{L}_{ij}$. This suggests that, for a small element r_{ij}, given $\widehat{L}_{ij} = 0$, the sample correlation coefficient $\widehat{r}_{ij}$ is small with high probability.

We present the asymptotic properties of the Network Guided Thresholding estimator in Theorem 1.

Theorem 1. *Suppose that Assumption 1, Assumption 2 and Assumption 3 hold with $l \le \lambda$, then for some constant A which does not depend on (N, T), we have:*

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

for some constant A which does not depend on (N, T), where $\|\cdot\|$ represents the operator norm, and κ_1, κ_2 are introduced in Assumption 1 and Assumption 3.

Remark: The error term due to the estimation of large elements is $c_1(N)\sqrt{\frac{\log N}{T}}$, and the effect of small elements appears in $c_0(N)\lambda^{1-q}$ and the error ϱ_T is due to the use of $\widehat{L}$ rather than the true L. From Equation 5, when $c_0(N)$ and $c_1(N)$ are both $O(1)$ or $c_0(N) \asymp c_1(N)$, the optimal choice of λ is $\lambda_T \asymp \left(\frac{\log N}{T}\right)^{1/2(1-q)}$, which then gives

$$\left\|\widehat{\boldsymbol{\Sigma}}_{\widehat{L}}^{\mathcal{T}} - \boldsymbol{\Sigma}\right\| = O_P\left(c_0(N)\left(\sqrt{\frac{\log N}{T}} + \varrho_T\right)\right), \tag{15}$$

and provided $c_0(N)\sqrt{\frac{\log N}{T}} \to 0$, we get $\left\|\widehat{\boldsymbol{\Sigma}}_{\widehat{L}}^{\mathcal{T}} - \boldsymbol{\Sigma}\right\| = o_P(1)$. Compared to the standard thresholding estimator, for example, as in Bickel and Levina (2008a) and Rothman et al. (2009), which has a convergence rate $c_0(N)\left(\frac{\log N}{T}\right)^{\frac{1-q}{2}}$, our estimator achieves a faster convergence rate when the auxiliary information is of good quality. For example, when the estimated indicator matrix is perfect, i.e., $\varrho_T \equiv 0$, we have

$$\left\|\widehat{\boldsymbol{\Sigma}}_{\widehat{L}}^{\mathcal{T}} - \boldsymbol{\Sigma}\right\| = O_P\left(c_0(N)\sqrt{\frac{\log N}{T}}\right),$$

which approaches zero faster than $c_0(N)\left(\frac{\log N}{T}\right)^{\frac{1-q}{2}}$.

3.1.2 Network Guided Banding Estimator

We assume that $\boldsymbol{\Sigma}_u \in \mathcal{U}_2(\varepsilon, \alpha, b_0, M)$ defined in Equation 9, which extends the class of bandable covariance matrix in Bickel and Levina (2008b). Again, we need to make further assumptions on the auxiliary network information to derive the asymptotic results for our Network Guided Banding Estimator $\widehat{\boldsymbol{\Sigma}}_{\widehat{C}}^{\mathcal{B}} = \widehat{D} B_{\widehat{C},k}\left(\widehat{R}\right)\widehat{D}$.

Assumption 4. *(a) For R and C, there exists b_1, s.t. $\max_{1\le i\le N} \sum_{j=1}^{N} |r_{ij}|\, I_{\left\{i\notin S_k^{c_j}, j\in S_k^{c_i}\right\}} < b_1(N)$, when $k = k_T \to \infty$;*[4]

(b) Suppose that $\widehat{C}$ is the estimator for C, and there exists a sequence $\kappa_3(N,T) \to 0$ when $T \to \infty$, for some constant A which does not depend on (N,T),

$$P\left(\frac{1}{k}\sum_{j=1}^{N} I_{\left\{j\in S_k^{c_i}, j\notin S_k^{\widehat{c}_i}\right\}} > A\sqrt{\frac{\log N}{T}}\right) = O\left(\kappa_3(N,T)\right), \quad P\left(\frac{1}{k}\sum_{j=1}^{N} I_{\left\{i\in S_k^{c_j}, i\notin S_k^{\widehat{c}_j}\right\}} > A\sqrt{\frac{\log N}{T}}\right) = O\left(\kappa_3(N,T)\right).$$

Remark: Our Relative Importance Matrix C may not imply a symmetric neighbor network, i.e., $i \in S_k^{c_j}$ does not indicate $j \in S_k^{c_i}$. In our procedure, we retain those r_{ij} when $i \in S_k^{\widehat{c}_j}$ and $j \in S_k^{\widehat{c}_i}$ while smoothing out other r_{ij}. Then asymmetry cases like $|r_{ij}|\, I_{\left\{i\notin S_k^{c_j}, j\in S_k^{c_i}\right\}}$ will contribute to the error, and we need the sum of these cases to be bounded by $b_1(N)$. Condition (a) restricts the degree of asymmetry; alternatively speaking, most of the asymmetric terms need to be small.

[4] Note that if $k_T = N$, we have $\sum_{j=1}^{N} |r_{ij}|\, I_{\left\{i\notin S_k^{c_j}, j\in S_k^{c_i}\right\}} \equiv 0$.

Condition (b) assumes that the number of misclassified cases for large elements is bounded by $O\left(k\sqrt{\frac{\log N}{T}}\right)$, which puts restrictions on the false negative error, i.e., $P\left(j \notin S_k^{\widehat{c}_i} \middle| j \in S_k^{c_i}\right)$. It is noteworthy that, unlike the additional restriction on errors caused by small elements in Assumption 3, the errors of non-neighboring elements for banding decay exponentially by well-chosen k in class $\mathcal{U}_2$. Therefore, the false positive error $P\left(j \in S_k^{\widehat{c}_i} \middle| j \notin S_k^{c_i}\right)$ is controlled implicitly.

Theorem 2. *Suppose that Assumption 1, Assumption 2, Assumption 4 hold and $k = k_T \to \infty$. Then,*

$$P\left(\left\|\widehat{\mathbf{\Sigma}}_{\widehat{C}}^{\mathcal{B}} - \mathbf{\Sigma}\right\| > A\left(k\sqrt{\frac{\log N}{T}} + b_0(N)\, k^{-\alpha} + b_1(N)\right)\right) = O\left(\frac{1}{N^2} + \kappa_1(N,T) + \kappa_3(N,T)\right),$$

for some constant A which does not depend on (N,T), where $\|\cdot\|$ represents the operator norm, and κ_1, κ_3 are introduced in Assumption 1 and Assumption 4.

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 due to the "asymmetry" introduced in Assumption 4. Additionally, the error caused by using the estimated Relative Importance Matrix $\widehat{C}$ is bounded by $O\left(\sqrt{\frac{\log N}{T}}\right)$ (details can be found in the proof of Theorem 2), thus dominated by the first component. Bickel and Levina (2008a) suggests an optimal choice of k, which is $k \asymp \left(\frac{\log N}{T}\right)^{-1/2(\alpha+1)}$, then we get

$$\left\|\widehat{\mathbf{\Sigma}}_{\widehat{C}}^{\mathcal{B}} - \mathbf{\Sigma}\right\| = O_P\left((1 + b_0(N))\left(\frac{\log N}{T}\right)^{\frac{\alpha}{2(\alpha+1)}} + b_1(N)\right). \tag{16}$$

If matrix C implies a network of symmetric neighbors ($i \in S_k^{c_j} \Leftrightarrow j \in S_k^{c_i}$), then our bound in Equation 16 becomes $O_P\left((1 + b_0(N))\left(\frac{\log N}{T}\right)^{\frac{\alpha}{2(\alpha+1)}}\right)$, which matches the result 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 $\left\|\widehat{\mathbf{\Sigma}}_{\widehat{C}}^{\mathcal{B}} - \mathbf{\Sigma}\right\| = o_P(1)$.

3.1.3 Convergence of $\widehat{\mathbf{\Sigma}}_y$

Given the factor structure, once we obtain $\widehat{\mathbf{\Sigma}}_u$, the feasible estimator for $\mathbf{\Sigma}_y$ is

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

where $\widehat{\boldsymbol{B}}$ is obtained by OLS estimation. We then follow the framework of Fan et al. (2011) to derive the asymptotic results for $\widehat{\mathbf{\Sigma}}_y$. To this end, they consider the entropy loss norm,[5] defined as

$$\left\|\widehat{\mathbf{\Sigma}}_y - \mathbf{\Sigma}_y\right\|_E = \left(\frac{1}{N}\mathrm{tr}\left\{\left(\widehat{\mathbf{\Sigma}}_y\mathbf{\Sigma}_y^{-1} - \boldsymbol{J}_{N\times N}\right)^2\right\}\right)^{1/2}, \tag{17}$$

which also equals $N^{-\frac{1}{2}}\left\|\mathbf{\Sigma}_y^{-\frac{1}{2}}\left(\widehat{\mathbf{\Sigma}}_y - \mathbf{\Sigma}_y\right)\mathbf{\Sigma}_y^{-\frac{1}{2}}\right\|_F$.

Corollary 1. *Under Assumption 1, Assumption 2, then (i) When Assumption 3 holds and our Network Guided Thresholding estimator $\widehat{\mathbf{\Sigma}}_{\widehat{L}}^{\mathcal{T}}$ attains the best convergence rate $c_0(N)\left(\sqrt{\frac{\log N}{T}} + \varrho_T\right)$, we have*

$$P\left(\left\|\widehat{\mathbf{\Sigma}}_y - \mathbf{\Sigma}_y\right\|_E > A\left(K\frac{\sqrt{N}\log N}{T} + \sqrt{K}\sqrt{\frac{\log N}{T}} + \frac{c_0(N)}{\sqrt{N}}\left(\sqrt{\frac{\log N}{T}} + \varrho_T\right)\right)\right) = O\left(\frac{1}{N^2} + \kappa_{1,2}\right),$$

where $\kappa_{1,2} := \kappa_1(N,T) + \kappa_2(N,T)$;

(ii) When Assumption 4 holds and our Network Guided Banding estimator $\widehat{\mathbf{\Sigma}}_{\widehat{C}}^{\mathcal{B}}$ attains the best convergence rate $(1 + b_0(N))\left(\frac{\log N}{T}\right)^{\frac{\alpha}{2(\alpha+1)}} + b_1(N)$, we have

$$P\left(\left\|\widehat{\mathbf{\Sigma}}_y - \mathbf{\Sigma}_y\right\|_E > A\left(K\frac{\sqrt{N}\log N}{T} + \sqrt{K}\sqrt{\frac{\log N}{T}} + \frac{(1 + b_0(N))}{\sqrt{N}}\left(\frac{\log N}{T}\right)^{\frac{\alpha}{2(\alpha+1)}} + \frac{b_1(N)}{\sqrt{N}}\right)\right) = O\left(\frac{1}{N^2} + \kappa_{1,3}\right),$$

where $\kappa_{1,3} := \kappa_1(N,T) + \kappa_3(N,T)$.

For both estimators, when $K\sqrt{N}\frac{\log N}{T} \to 0$, we have $\left\|\widehat{\mathbf{\Sigma}}_y - \mathbf{\Sigma}_y\right\|_E = o_P(1)$. This condition also reduces to $\sqrt{N}\log N = o(T)$ in the case where K is finite.

[5] Fan et al. (2012) provided an upper bound for $\left\|\widehat{\mathbf{\Sigma}}_y - \mathbf{\Sigma}_y\right\|_F$, but for this upper bound to go to zero, $N^2 < T$ is required, making $\left\|\widehat{\mathbf{\Sigma}}_y - \mathbf{\Sigma}_y\right\|_F$ or $\left\|\widehat{\mathbf{\Sigma}}_y - \mathbf{\Sigma}_y\right\|$ unsuitable as a criterion here.

3.2 Global Minimum Variance Portfolios

To evaluate the performance of the covariance matrix estimation, we consider constructing the Global Minimum Variance (GMV) portfolio. It is notoriously hard to estimate both the first and second moments of asset returns through past observations, which motivates us to use the GMV portfolio as a playground to test our methodology. Compared with the mean-variance optimal portfolio as in Markowitz (1952), the GMV portfolio avoids the estimation error of the expectation of asset returns, which can better reflect the performance of covariance matrix estimators.

Under mild assumptions, the portfolio weights for the GMV portfolio are:

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

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 11, we have $\mathrm{Cov}(\boldsymbol{y}_t, \boldsymbol{y}_t) = \boldsymbol{\Sigma}_y = \boldsymbol{B}\boldsymbol{\Sigma}_f\boldsymbol{B}^\intercal + \boldsymbol{\Sigma}_u$. The covariance from common factors are $\widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^\intercal$ as in Equation 13. Then, with the help of auxiliary network information, we can estimate $\boldsymbol{\Sigma}_u$ using our Network-Guided approach. After obtaining $\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_{u,\widehat{L}}$ or $\widehat{\boldsymbol{\Sigma}}^{u\mathcal{B}}_{u,\widehat{C}}$, we can construct $\widehat{\boldsymbol{\Sigma}}_y = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^\intercal + \widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_{u,\widehat{L}}$ or $\widehat{\boldsymbol{\Sigma}}_y = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^\intercal + \widehat{\boldsymbol{\Sigma}}^{\mathcal{B}}_{u,\widehat{C}}$ and derive the estimated weights $\widehat{\boldsymbol{\omega}} = \frac{\widehat{\boldsymbol{\Sigma}}_y^{-1}\mathbf{1}}{\mathbf{1}^\intercal\widehat{\boldsymbol{\Sigma}}_y^{-1}\mathbf{1}}$.

3.3 Positive Definiteness of $\widehat{\boldsymbol{\Sigma}}_y$

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

$$\widehat{\boldsymbol{\Sigma}}_{M_0} = \widehat{\boldsymbol{\Sigma}} + (m_T - \widehat{\rho}_N) \cdot \boldsymbol{J}_{N\times N}, \tag{18}$$

where $\boldsymbol{J}_{N\times N}$ is the $N \times N$ identity matrix and $m_T > 0$ is a tuning parameter. This modification ensures that the smallest eigenvalue is positive, making $\widehat{\boldsymbol{\Sigma}}_{M_0}$ invertible. Chen et al. (2019) further augment Equation 18 by defining

$$\widehat{\boldsymbol{\Sigma}}_M = \widehat{\boldsymbol{\Sigma}} \cdot \mathbf{1}_{\{\widehat{\rho}_N > 0\}} + \widehat{\boldsymbol{\Sigma}}_{M_0} \cdot \mathbf{1}_{\{\widehat{\rho}_N \leq 0\}} = \widehat{\boldsymbol{\Sigma}} + (m_T - \widehat{\rho}_N) \cdot \boldsymbol{J}_{N\times N} \cdot \mathbf{1}_{\{\widehat{\rho}_N \leq 0\}}, \tag{19}$$

which ensures that $\widehat{\boldsymbol{\Sigma}}$ is retained when it is already positive definite, while $\widehat{\boldsymbol{\Sigma}}_{M_0}$ is used when non-positive eigenvalues appear.

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

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

where $\widehat{\boldsymbol{\Sigma}}_f$ is naturally invertible in a (finite) factor structure while $\widehat{\boldsymbol{\Sigma}}_u^{-1}$ may not be well-defined. We modify $\widehat{\boldsymbol{\Sigma}}_u$ using Equation 19. However, since

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

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

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

leading to $\left\|\widehat{\boldsymbol{\Sigma}}_{uM} - \boldsymbol{\Sigma}_u\right\| \leq O_P\left(\left\|\widehat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u\right\|\right) + m_T$. Thus, the tuning parameter should approach zero faster than the convergence rate of $\widehat{\boldsymbol{\Sigma}}_u$, ensuring that the modified version $\widehat{\boldsymbol{\Sigma}}_{uM}$ converges to $\boldsymbol{\Sigma}_u$ at the same rate as $\widehat{\boldsymbol{\Sigma}}_u$. Specifically, m_T should go to 0 faster than

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

4 Simulation

4.1 True Covariance Matrix

Similarly 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.

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

- **Setup 2 (sparse matrix without ordering)**: $\boldsymbol{\Sigma} = \mathrm{diag}\{A_1, A_2\}$, 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:

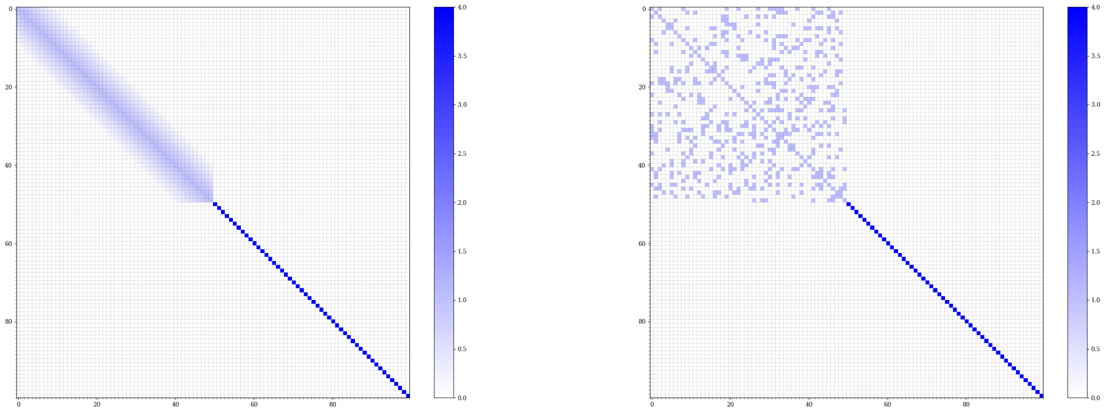
$$b_{ij} = \begin{cases} \mathrm{Ber}\left(\frac{20}{N}\right), & \text{for } i < j, \\ 1, & \text{for } i = j, \\ b_{ji}, & \text{for } i > j. \end{cases} \tag{20}$$

  $\mathrm{Ber}\,(x)$ is a Bernoulli random variable that takes the value 1 with probability x and the 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 Network Information

In the simulation, we fix the true Location Indicator Matrix L and the Relative Importance Matrix C, and generate their estimates, i.e., $\widehat{L}$ and $\widehat{C}$. The qualities of these estimates are tuned by some hyperparameters:

1. **Network Guided Thresholding**. We fix the Observation Level parameter $l = 0.2$, which means $L_{ij} = 1$ if and only if $|r_{ij}| > 0.2$.



(a) Setup 1

(b) Setup 2

Figure 1: Typical heatmaps of two banded and sparse models

False Positive error $\boldsymbol{\zeta}$: Conditional on $L_{ij} = 0$, the probability of $\widehat{L}_{ij} = 1$.

False Negative error $\mathbf{1} - \boldsymbol{p}$: Conditional on $L_{ij} = 1$, the probability of $\widehat{L}_{ij} = 0$.

2. **Network Guided Banding**.

Accuracy Rate $\boldsymbol{\eta}$: Conditional on $j \in S_k^{c_i}$, the probability of $j \in S_k^{\widehat{c}_i}$.

Table 1 lists the descriptions of these hyperparameters and the ranges of values that they take in the numerical experiment. The estimate $\widehat{L}$, derived from auxiliary information in the real world, is typically sparse, meaning that the number of elements with $\widehat{L}_{ij} = 1$ is limited. Therefore, it is reasonable to set small values for ζ (e.g., 0.1 and 0) in our simulation setup.

Table 1: Hyperparameters and Auxiliary Network Information Quality.

Hyperparameters	**Auxiliary Information Quality**			
	Very Low	*Low*	*High*	*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

4.3 Numerical Results

To better match the real asset return data in our empirical work, we generate asset returns using the CH-4 factor model proposed by Liu et al. (2019), which consists of four factors: Market, Value-Minus-Growth, Small-Minus-Big and Pessimistic-Minus-Optimistic.[6] We first use weekly return data for all listed stocks and CH-4 factor data from 2000 to 2021 to fit the CH-4 model for each stock and then collect the factor loading coefficients. Doing so for all stocks provides the mean and standard deviation for each loading parameter, as well as the covariance matrix for de-factored idiosyncratic returns.

To be precise, from the estimates, we have $\beta_{1i} \sim N\left(0.7013, 0.1961^2\right)$, $\beta_{2i} \sim N\left(-0.1582, 0.2055^2\right)$, $\beta_{3i} \sim N\left(-0.1200, 0.2182^2\right)$, $\beta_{4i} \sim N\left(-0.0050, 0.2245^2\right)$, and $\boldsymbol{u}_t \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_u\right)$. To match the data used in the empirical study, we independently draw $\beta_{1i}, \beta_{2i}, \beta_{3i}, \beta_{4i}$, and $\boldsymbol{u}_t$ from the aforementioned distributions. We sample factor returns from normal distributions, with means and covariance matrices set to match the historical data from 2000 to 2021, as shown in Table 2. Setting $\boldsymbol{\beta}_0$ to $\mathbf{0}$, we generate the asset return data for $N = 100, 300, 500$ and $T = 300$ using the following factor model:

$$\boldsymbol{y}_t = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 f_{\text{MKT,t}} + \boldsymbol{\beta}_2 f_{\text{VMG,t}} + \boldsymbol{\beta}_3 f_{\text{SMB,t}} + \boldsymbol{\beta}_4 f_{\text{PMO,t}} + \boldsymbol{u}_t. \tag{21}$$

Table 2: Descriptive Statistics of Weekly CH-4 Factor Data from 2000 to 2021

	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

Using the simulated data, we estimate the CH-4 factor model and collect residuals $\widehat{\boldsymbol{u}}_t$,

[6] The CH-4 factor model is found to suit the Chinese stock market well and outperforms the Fama-French 5-factor model.

and then we employ various methods to estimate its covariance. Each scenario, specified by (N, T, p, ζ) or (N, T, η), is repeated 100 times. Our examination centers on the finite sample performance 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.[7] The results, evaluated using both the Frobenius norm and the operator norm, are presented in Table 3, illustrating the comparative performances of the various estimation methods.

Panel A in Table 3 shows the results for our setup 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, except in the scenario where $N < T$ with $(p, \zeta) = (0.3, 0.1)$, indicating poor auxiliary information, it outperforms the sample covariance estimator, soft thresholding estimator, hard thresholding estimator, and linear shrinkage estimator in all combinations (p, ζ, N). Nonetheless, to outperform the nonlinear shrinkage estimator, the quality of the information needs to be comparably higher. As expected, the FP error ζ and FN error $1 - p$ affect the performance of the Network Guided Thresholding Estimator. In an ideal scenario where $(p, \zeta) = (1.0, 0.0)$, we have $\widehat{L} = L$.

When examining the Network Guided Banding Estimator, it exhibits smaller norms than all other purely statistical methods, as long as the accuracy rate parameter η is not excessively low. For example, with $\eta = 0.6$, a moderate accuracy rate, the Network Guided Banding Estimator demonstrates superiority for most (N, T) combinations, particularly when $N \geq T$. With comparable information quality, the Network Guided Banding Estimator typically outshines the Network Guided Thresholding Estimator, aligning with theoretical expectations. However, our objective is not to compare the two Network-Guided estimators against each

[7] Numerical performance is assessed through the comparison of both the Frobenius norm and the operator norm, i.e., $\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_F$ and $\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|$.

other, as this would not be a fair comparison given that each approach is designed for different types of auxiliary network information.

Panel B is devoted to the scenario where the true covariance matrix is a sparse matrix without order (setup 2). In this context, both Network-Guided estimators continue to outperform in the competition, as long as the auxiliary network information is of decent quality. Note that the traditional banding method cannot be applied to estimate covariance matrices from setup 2. However, our Network Guided Banding Estimator is adaptable to a broader spectrum of bandable matrices and shows good performance in this setting, provided that the accuracy rate parameter η is not unduly low. For example, with $\eta = 0.6$, the Network Guided Banding Estimator surpasses other methods across all (N, T) combinations. Similarly to the previous setup, the Network Guided Thresholding Estimator exhibits strong performance, particularly when the FP errors are small.

In summary, our simulation exercise highlights the promising numerical properties of the proposed Network-Guided estimators. Both estimators generally outperform traditional purely statistical approaches, provided that the auxiliary information is of decent quality.

Table 3: Simulation Results. 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.), in terms of Frobenius norm and the operator norm. Note that Python package `nonlinshrink` only works when $T > N$. For each method, we perform 100 simulations and report the mean and standard deviation. Results are displayed for different values of $N = 100, 300, 500$, with T fixed at 300.

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: Setup 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)
	$\|\cdot\|$	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)	
	$\|\cdot\|$	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)	
	$\|\cdot\|$	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: Setup 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)
	$\|\cdot\|$	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)	
	$\|\cdot\|$	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)	
	$\|\cdot\|$	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 of three well-known indices in China in 2021, namely HS300 (000300.SH), CSI500 (000905.SH), and CSI800 (000906.SH), which consist of approximately 300, 500, and 800 stocks, respectively. The daily returns of these stocks were collected from the RESSET database, covering the period from 2006 to 2021, with ST stocks excluded.[8]

5.1.2 News Co-mention Linkage Data

We analyzed 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.

Following the approach of Ge et al. (2023), we identify news-implied links based on co-mentions within the same news article. We propose four methods to identify linkages based on different types of co-mentions, namely `one2one_passage`, `all_passage`, `one2one_sentence` and `all_sentence` approaches. In Table 4, we summarize the differences between these link identification strategies:

At time t, we use the latest τ_0 days as the identification window.[9] For each of the four link identification strategies, we count the number of co-mentions M_{ij} for each stock pair (i, j), and

[8] In this article, we assume that 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.

[9] Empirically, we choose τ_0 to be 21 (1 month) or 252 (12 months) to examine the performance of the link identification strategy under short and long identification windows.

Table 4: News Co-mention Types and Link Identification

	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`

then we 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 stock linkages based on analyst coverage, denoted as `Analyst`. This approach is supported by the literature suggesting that shared analyst coverage can indicate fundamental connections between companies, reflecting similarities across various dimensions (see Ali and Hirshleifer (2020), Israelsen (2016), and Kaustia and Rantala (2013)). We used data from the Chinese Research Data Services Platform (CNRDS), covering analyst reports from January 2005 to December 2020. After data cleaning, we identified 530,696 unique analyst reports to trace connections based on shared coverage. Starting from 2006, at time t, we use the most recent one-year window for link identification. For each pair of stocks (i, j), we count the number of co-coverages M_{ij} during the identification window to build the analyst co-coverage linkage matrix $M = (M_{ij})$ for $i, j = 1, 2, \ldots, N$.

5.1.4 Industry-based Linkage

Stocks within the same sector or industry often co-move beyond exposure to common risk factors. Based on this, we examine linkages formed based on industry classifications, denoted as `Industry`. We analyzed three major industry classification systems in China: CSRC, CITIC, and Shenwan, updating annually using the RESSET database. Our primary focus is the Shenwan primary classification, which is recognized as the leading system within China's financial industry.

5.1.5 Summary Statistics of All Types of Auxiliary Network

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 potential noise links present in the same article strategy, resulting in fewer identified links. Additionally, 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, with each sample stock having more linked stocks 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 the 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
	# Linked Stocks	16	32	1	5	454
`all_sentence_12`	# Stocks	1750	233	1355	1742	2704
	# Linked Stocks	23	42	1	8	631
`all_passage_1`	# Stocks	1976	229	1478	1952	2816
	# Linked Stocks	29	51	1	10	757
`all_passage_12`	# Stocks	2122	278	1569	2121	2891
	# Linked Stocks	35	59	1	12	867
`analyst`	# Stocks	1326	348	476	1429	1872
	# Linked Stocks	98	84	1	75	609
`industry`	# Stocks	2336	795	1048	2313	3893
	# Linked Stocks	130	83	2	110	364

5.2 Methodology

We use the Global Minimum Variance (GMV) portfolio as a testing ground to evaluate different covariance matrix estimation techniques. We are particularly interested in whether the

GMV portfolio, constructed with the help of auxiliary network information, outperforms other methods. This subsection presents the procedures for applying our proposed Network-Guided estimators to the stock return covariance matrix, followed by an out-of-sample comparison.

5.2.1 CH-4 Factor Model

We first de-factor the stock returns using observable factors, adopting the CH-4 factors model from Liu et al. (2019):

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 f_{\mathrm{MKT,t}} + \boldsymbol{\beta}_2 f_{\mathrm{VMG,t}} + \boldsymbol{\beta}_3 f_{\mathrm{SMB,t}} + \boldsymbol{\beta}_4 f_{\mathrm{PMO,t}} + \boldsymbol{u}_t \\ &= \boldsymbol{\beta}_0 + \boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{u}_t, \end{aligned} \tag{22}$$

where the time series of the four factors can be obtained from the author's website. The estimator of $\boldsymbol{\Sigma}_y = \mathrm{Cov}\left(\boldsymbol{y}_t, \boldsymbol{y}_t\right)$ is given by

$$\widehat{\boldsymbol{\Sigma}}_y = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^{\intercal} + \widehat{\boldsymbol{\Sigma}}_u, \tag{23}$$

where the factor loading matrix $\widehat{\boldsymbol{B}}$ is obtained using OLS. 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 $[C_{ij}]_{N\times N}$

Depending on the nature of the auxiliary network information, we can choose which network-guided method to apply. In general, if the auxiliary network dataset provides unweighted linkage information, that is, simply indicating whether a pair of stocks is linked without quantifying the strength of the connection, then we can apply the Network Guided Thresholding method but not the Network Guided Banding method. On the other hand, if the auxiliary network dataset provides weighted linkage information, revealing the relative importance of neighbors for each node, then we can apply both the Network Guided Thresholding and the Network Guided Banding methods.

The news co-mention auxiliary network implies weighted linkage information and can therefore be applied to both Network Guided Thresholding and Network Guided Banding estimation. To apply the Network Guided Thresholding, we first estimate the Location Indicator Matrix L. Since L is a zero-one matrix, we tune a threshold parameter m, setting $\widehat{L}_{ij} = 1$ only if i and j are co-mentioned more than m times in a given identification window. That is, $\widehat{L}_{ij} = \mathbf{1}\{M_{ij} \geq m\}$, where the tuning parameter m is chosen by the in-sample cross-validation. To apply the Network Guided Banding Estimation, we construct an estimate of the Relative Importance Matrix C from the news co-mention matrix M, where a typical entry $0 \leq M_{ij} < \infty$ provides integer counts of the times of co-mentions. For each row M_i of M, we set $\widehat{c}_{ij}$ to be the rank of M_{ij} within that row, and the number of neighbors k of each asset is also determined by the in-sample cross-validation. Given the estimates of L and C, the network-guided procedure introduced in Section 2 can be applied. In our empirical work, we select the value of m and k that minimize the variance of the portfolio in the training sample.

The analyst co-coverage linkage network has the same properties as the news co-mention auxiliary network. Therefore, all procedures are identical to those described above. In contrast, the industry-based linkage network is unweighted, where $M_{ij} = \mathbf{1}\{i \text{ and } j \text{ are in the same industry}\}$. Given the nature of the industry network, we cannot learn relative importance information from it, and thus cannot apply our Network Guided Banding method. However, by setting $\widehat{L}_{ij} = M_{ij}$, we can still apply the Network Guided Thresholding method.

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 an effective method to evaluate the performance of covariance matrix estimators. Unlike the optimal mean-variance (MV) portfolio, the GMV portfolio avoids the need to estimate asset mean returns, which can introduce considerable noise.

In this part, we apply the proposed method to a portfolio management problem. Specifically, we compare the performance of GMV portfolios as outlined in Ledoit and Wolf (2004). The theoretical weights for a GMV portfolio are given by

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

where $\boldsymbol{\Sigma}_y$ is the covariance matrix of asset returns, and $\mathbf{1}$ is the conforming vector of ones.

The estimator of the return covariance matrix can be decomposed as $\widehat{\boldsymbol{\Sigma}}_y=\widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^{\intercal}+\widehat{\boldsymbol{\Sigma}}_u$ under the factor structure. The factor component can be easily estimated using the CH-4 factor model. Our goal is to demonstrate that the proposed method can more accurately estimate $\boldsymbol{\Sigma}_u$ and thus improve the performance of the GMV portfolio. Using a rolling window approach starting in 2012, we train the model with one year of data and use the in-sample results for a one-month test. This procedure is repeated until the end of 2021.

For robustness check, we also consider the maximum return portfolio for any given variance level σ_0^2 and the minimal variance portfolio for any given expected return level μ_0. Recall the construction of the classical optimal portfolio. For example, given a return constraint μ_0, we have the minimization problem:

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

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

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

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 of asset returns, we have $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{B}}\bar{\boldsymbol{f}}$ and $\widehat{\boldsymbol{\Sigma}}_y = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^\intercal + \widehat{\boldsymbol{\Sigma}}_u$. Plug in the optimal weight, the minimal variance given μ_0 is

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

which also gives the mean-variance efficient frontier set $\{(\sigma_0, \mu_0), \mu_0 \geq 0\}$. Starting from the maximization problem for any given σ_0^2 leads to the same efficient frontier. However, note that the efficient frontier is in-sample. When we set a fixed in-sample σ_0 or μ_0, the out-of-sample portfolio may yield different values for standard deviation and mean return, resulting in an out-of-sample efficient frontier. Similarly to the GMV portfolio, we select tuning parameters through in-sample training and then construct the out-of-sample efficient frontiers under different models.

Importantly, although most of the estimated covariance matrices are positive definite, we modify any non-positive definite covariance matrices $\widehat{\boldsymbol{\Sigma}}_u$ using the method outlined in Equation 19.

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 using different methods and stock samples, including the constituent stocks of the HS300, CSI500, and CSI800 indices. We consider the following benchmark models:

1. **Sample**: Use the sample covariance matrix of $\widehat{\boldsymbol{\Sigma}}_u$ with a positive definite correction if necessary, and compute $\widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^\intercal + \widehat{\boldsymbol{\Sigma}}_u$.

2. **Linear Shrinkage**: Operate linear shrinkage of Ledoit and Wolf (2004) on $\widehat{\boldsymbol{\Sigma}}_u$ with positive definite correction if necessary, and compute $\widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^\intercal + \widehat{\boldsymbol{\Sigma}}_u$

3. **Factor Only**: Use $\widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^\intercal + \mathrm{diag}(\widehat{\sigma}_1^2, \cdots, \widehat{\sigma}_N^2)$ as $\widehat{\boldsymbol{\Sigma}}_y$.

4. **Equal Weights**: Assign equal weights $\frac{1}{N}$ to each of the N assets for the out-of-sample GMV portfolios.

The overall results are shown in Panel A, where "Best Thresholding" and "Best Banding" represent the best-performing portfolios utilizing different types of auxiliary information. Further details are provided in Panel B (Network Guided Thresholding) and Panel C (Network Guided Banding), respectively.

Results for the benchmark models are presented in Panel A. The "Factors Only" approach consistently outperforms the "Sample" method across all indices. Estimating the covariance matrix using a factor model reduces the estimation error inherent in the sample covariance matrix, thereby mitigating the impact of individual asset noise. The "Sample" method can exhibit high estimation error due to the large number of parameters, with noise potentially leading to poor out-of-sample performance. Additionally, the "Linear Shrinkage" method offers a competitive, and in some cases superior, reduction in standard deviation compared to "Factors Only", particularly for the CSI500 index. This underscores the potential of shrinkage methods in enhancing portfolio allocation.

For the Network Guided Thresholding (Panel B), the results exhibit a varied performance landscape. Incorporating `analyst` and `industry` network information, the Network Guided Thresholding method does not demonstrate superior performance compared to the "Factors Only" method. However, news-implied linkages generally prove more effective in enhancing covariance matrix estimation, with reduced out-of-sample volatility in most cases. Turning to the results of the guided Banding approach (Panel C), news-implied networks again show greater effectiveness than other auxiliary network information in improving covariance matrix estimation. Note that `industry` provides only unweighted linkage information, making it unsuitable for Banding.

Combining the results from both Network-Guided estimators, we find that news-based auxiliary network information is more effective in identifying linked pairs compared to other sources.

This finding aligns with the results of Ge et al. (2023). However, we recognize that the best-performing strategy for identifying news-implied linkages varies across different indices. This suggests that while auxiliary news-implied network information is valuable, its application should be tailored to specific market conditions and characteristics due to the complex nature of financial markets. Future research could explore the mechanisms behind these variations to further refine the estimation process.

Table 6: Out-of-sample Standard Deviation of GMV Portfolios. We compare the out-of-sample standard deviations of GMV portfolios constructed using different covariance matrix estimators, while adopting the factor structure for asset returns. The covariance due to common factors remains the same across all methods, with variations arising in the estimation of $\boldsymbol{\Sigma}_u$. "Sample" refers to a simple sample estimator of $\boldsymbol{\Sigma}_u$; "Factors Only" refers to setting $\widehat{\boldsymbol{\Sigma}}_u = \mathrm{diag}\{\widehat{\sigma}_1^2, \cdots, \widehat{\sigma}_N^2\}$; and "Equal Weights" refers to a portfolio with equal weights for all assets. The out-of-sample standard deviation of the best-performing portfolio for each index is highlighted in bold.

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**	0.0733
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		0.0483	0.0469	**0.0445**	0.0489
CSI500	0.0742		0.0756	**0.0733**	0.0744	0.0768
CSI800	0.0598		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

We also compare the performance of various optimal portfolios under different covariance matrix estimations, focusing on the CSI500 index for illustrative purposes. We calculate the out-of-sample efficient frontiers using different methods, as shown in Figure 2, with all returns and volatilities annualized.

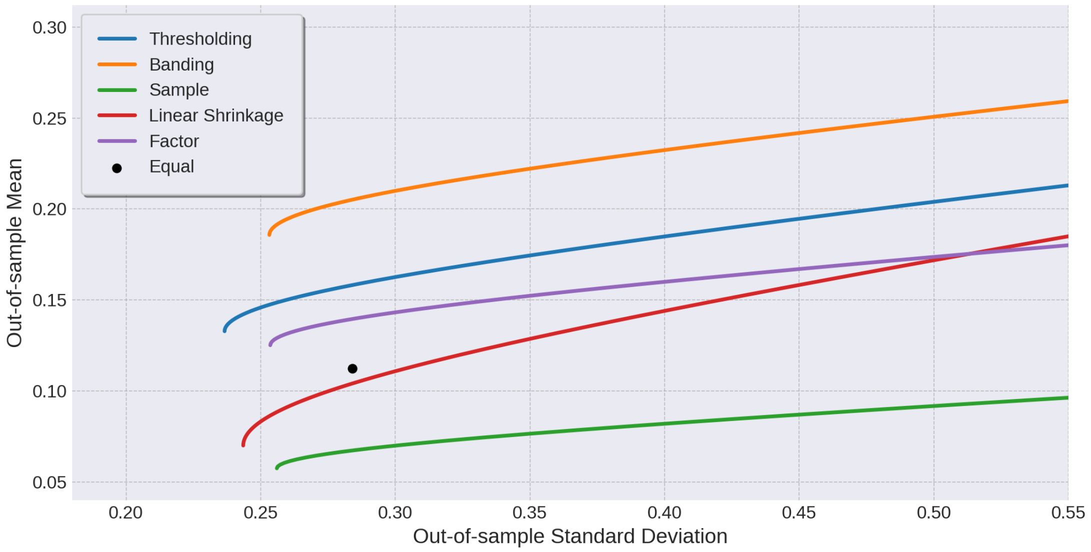


Figure 2: Out-of-sample Efficient Frontiers. For the two Network-Guided methods, we show the efficient frontiers using the best auxiliary information. "Thresholding" refers to Network Guided Thresholding, while "Banding" refers to Network Guided Banding.

From Figure 2, we see that the Network Guided Thresholding method achieves minimal variance, which aligns with the results in Table 6. Considering out-of-sample mean returns, the portfolio constructed using Network Guided Banding outperforms both Network Guided Thresholding and all other baseline models. Apart from Thresholding and Banding, only the "Factor" method generates higher average returns than the "Equal Weights" portfolio for all volatility levels. Linear Shrinkage generates low mean returns compared with others when the volatility is small, but produces larger mean returns than the "Factor" method when the volatility is relatively high. Finally, the sample covariance matrix consistently underperforms in constructing mean-variance portfolios in our study, highlighting the necessity of improving the estimation of large covariance matrices.

Table 7: Portfolios Performance Given Out-of-Sample Standard Deviations. We compare the out-of-sample mean return and Sharpe ratio of the best portfolio under a given volatility constraint, constructed using different covariance matrix estimators.

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 out-of-sample portfolio performances for various values of the volatility. The results are consistent with those in Figure 2. We compare the Sharpe ratios under a given portfolio volatility, with Network-Guided portfolios exhibiting higher Sharpe ratios than all benchmark models, and the Network Guided Banding performing the best.

Furthermore, we analyze the maximal Sharpe Ratio portfolios (or Mean-Variance optimal portfolios) under different models. We search the efficient frontier depicted in Figure 2 to find the mean-variance optimal results for each model for comparison. Figure 3 plots the back-test performance over a 10-year out-of-sample window for these portfolios, and the evaluation statistics are presented in Table 8. Due to the crash of the Chinese stock market in May and June 2015, no portfolio achieves a Sharpe ratio greater than 1. However, compared to the other four benchmarks, the Network Guided portfolios perform better over the entire period, especially the Banding method. In terms of return and standard deviation, the "Factor" method is close to our Network Guided Thresholding, but it tends to produce a higher maximum draw-down. The Network Guided Banding portfolio provides the best performance in our backtest, with the highest return and the lowest maximum drawdown. It is worth noting that a simple equal-weight portfolio actually performs better than many of the other benchmark methods. This is not surprising. DeMiguel et al. (2009) documented that the gains from mean-variance optimal diversification are often more than offset by estimation errors in practice, making a naive equal-weight strategy more efficient than previously thought.

Table 8: Mean-Variance Optimal Portfolios Performances. This table reports the mean, standard deviation, Sharpe ratio, and maximum drawdown of mean-variance portfolios constructed using different covariance matrix estimation methods. "Thresholding" refers to Network Guided Thresholding, while "Banding" refers to Network Guided Banding.

	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 empirical results validate the utility of incorporating network infor-

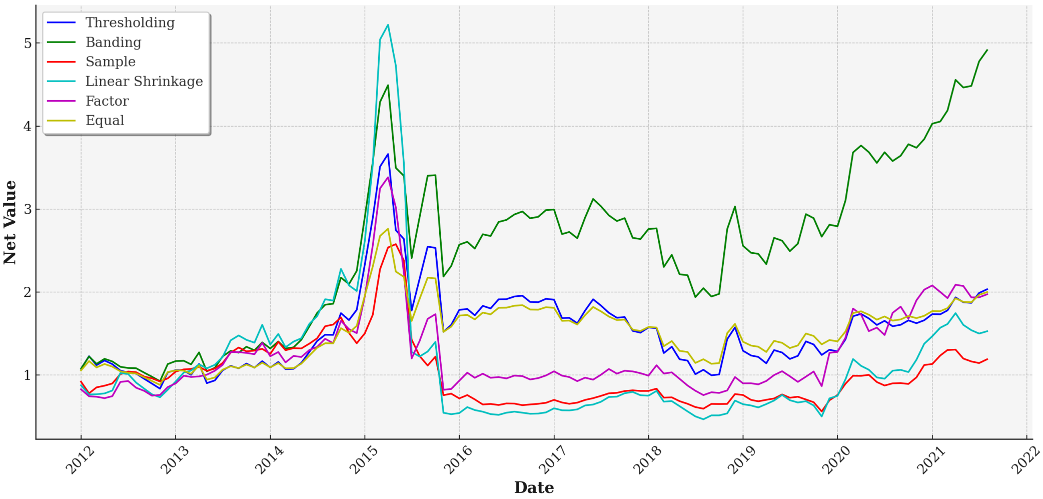


Figure 3: Out-of-sample Mean-Variance Optimal Portfolios. This figure tracks the net value of mean-variance portfolios constructed using different covariance matrix estimation methods. "Thresholding" refers to Network Guided Thresholding, while "Banding" refers to Network Guided Banding.

mation into covariance matrix estimation for portfolio optimization. While the factor model primarily captures the strong cross-sectional dependence among asset returns, auxiliary information such as news, as discussed in Ge et al. (2022), helps identify local or weak cross-sectional dependencies. This is why auxiliary information improves the estimation of $\mathbf{\Sigma}_u$, the covariance matrix of de-factored returns. However, these findings also highlight the complex 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 to further refine the estimation process.

6 Conclusion

In the era of big data, the availability of auxiliary information beyond the observations of $\{\boldsymbol{X}_t\}_{t=1}^T$ offers valuable opportunities to enhance the performance of conventional statistical and econometric models. Our study provides theoretical results demonstrating that integrating auxiliary data, when tailored to fit conventional thresholding and banding methods, exhibits improved properties. Both simulation studies and empirical illustrations validate that the proposed estimators outperform many benchmark models, provided the auxiliary network information is of reasonable quality. Therefore, the answer to "should we augment large covariance matrix estimation with auxiliary network information?" is a yes; integrating auxiliary network information of decent quality into conventional covariance matrix estimation methods is recommended.

In this paper, we focus primarily on the estimation of static covariance matrices. However, we suggest that a similar approach can be extended to other settings, such as the estimation of large dynamic covariance matrices. For instance, dynamic network information could be effectively incorporated into the conditioning information set, as discussed 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. doi: 10.1016/j.jfineco.2019.10.007.

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, (6):2577–2604, 2008a. doi: 10.1214/08-AOS600.

P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008b. doi: 10.1214/009053607000000758.

T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011. doi: http://www.jstor.org/stable/41416401.

T. T. Cai and H. H. Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389–2420, 2012. doi: 10.1214/12-AOS998.

G. Chamberlain and M. Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304, 1982. doi: doi.org/10.2307/1912275.

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. doi: 10.1016/j.jeconom.2019.04.025.

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

V. DeMiguel, L. Garlappi, and R. Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953, 2009. doi: 10.1093/rfs/hhm075.

R. F. Engle, O. Ledoit, and M. Wolf. Large dynamic covariance matrices. *Journal of Business & Economic Statistics*, 37(2):363–375, 2019. doi: 10.1080/07350015.2017.1345683.

E. F. Fama and K. R. French. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56, 1993. doi: 0.1016/0304-405X(93)90023-5.

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

J. Fan, J. Zhang, and K. Yu. Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606, 2012. doi: 10.1080/01621459.2012.682825.

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. doi: 10.1111/rssb.12016.

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. doi: 10.1080/07350015.2015.1052458.

S. Ge, S. Li, and O. Linton. News-implied linkages and local dependency in the equity market. *Journal of Econometrics*, 2022. doi: 10.1016/j.jeconom.2022.07.004.

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

C. Giraud, Y. Issartel, and N. Verzelen. Localization in 1d non-parametric latent space models from pairwise affinities. *Electronic Journal of Statistics*, 17(1):1587–1662, 2023. doi: 10.1214/23-EJS2134.

G. Hoberg and G. Phillips. Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5):1423–1465, 2016. doi: 10.1086/688176.

R. D. Israelsen. Does common analyst coverage explain excess comovement? *Journal of Financial and Quantitative Analysis*, 51(4):1193–1229, 2016. doi: www.jstor.org/stable/44157611.

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. *The Journal of Portfolio Management*, 30(4):110–119, 2004. doi: 10.3905/jpm.2004.110.

O. Ledoit and M. Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, 2012. doi: 10.1214/12-AOS989.

Z. M. Li and O. Linton. A remedi for microstructure noise. *Econometrica*, 90(1):367–389, 2022. doi: 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. doi: 10.1016/j.jfineco.2019.03.008.

H. M. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952. doi: 10.1111/j.1540-6261.1952.tb01525.x.

S. Ross. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3): 341–360, 1976. doi: 10.1016/0022-0531(76)90046-6.

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. doi: www.jstor.org/stable/40591909.

A. Scherbina and B. Schlusche. Economic linkages inferred from news stories and the predictability of stock returns. *Available at SSRN 2363436*, 2015.

G. Schwenkler and H. Zheng. The network of firms implied by the news. *Available at SSRN 3320859*, 2019.

Appendices

A Proofs

A.1 Proof of Theorem 1

Proof. By triangle inequality, we have $\left\|\widehat{R}^{\mathcal{T}}_{\widehat{L}} - R\right\| \le \left\|\widehat{R}^{\mathcal{T}}_{\widehat{L}} - R^{\mathcal{T}}_{\widehat{L}}\right\| + \left\|R^{\mathcal{T}}_{\widehat{L}} - R^{\mathcal{T}}_{L}\right\| + \left\|R^{\mathcal{T}}_{L} - R\right\|$.

For the first part, under Assumption 3,

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

For some constant $A > 0$, we define

$$A_{1,1} = \left\{ \max_{i,j} |\widehat{\sigma}_{ij} - \sigma_{ij}| > A\sqrt{\frac{\log N}{T}} \right\} \cup \left\{ \max_{i,j} |\widehat{\sigma}_{ij} - \widetilde{\sigma}_{ij}| > A\sqrt{\frac{\log N}{T}} \right\},$$

whose probability is shown to be bounded by $O\left(\frac{1}{N^2} + \kappa_1(N, T)\right)$ in Lemma A.3 of Fan et al. (2011), with

$$\widetilde{\sigma}_{ij} = \frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt}, \quad \widehat{\sigma}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \widehat{u}_{it}\widehat{u}_{jt}.$$

For function $g(\sigma_{ij}, \sigma_{ii}, \sigma_{jj}) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$, it is straightforward to show that all the three following first-order derivatives,

$$g_1 = \frac{1}{\sqrt{\sigma_{ii}\sigma_{jj}}}, \quad g_2 = \frac{-\sigma_{ij}}{2\sigma_{ii}^{3/2}\sqrt{\sigma_{jj}}}, \quad g_3 = \frac{-\sigma_{ij}}{2\sqrt{\sigma_{ii}}\sigma_{jj}^{3/2}},$$

are bounded. Firstly, Equation 4 implies $\sigma_{ii} \leq M$ and $\sigma_{jj} \leq M$. Consequently, $|\sigma_{ij}| \leq M$ is directly from the Cauchy-Schwarz inequality. Furthermore, condition (a) in Assumption 1, which states $\rho_{\min}(\boldsymbol{\Sigma}_u) > \underline{c} > 0$, suggests that σ_{ii} and σ_{jj} are bounded from below. When $A_{1,1}^c$ occurs and T is large enough, $\widehat{\sigma}_{ij}$ and σ_{ij} can be close sufficiently. Then we have for all i, j, $|\widehat{\sigma}_{ij}|$ is bounded, and $\widehat{\sigma}_{ii}$, $\widehat{\sigma}_{jj}$ are also bounded from below. Thus, due to the bounded partial derivatives of the function g, $\max_{i,j} |\widehat{r}_{ij} - r_{ij}| \leq O(\max_{i,j} |\widehat{\sigma}_{ij} - \sigma_{ij}|)$. Therefore, under the event $A_1^c := \Omega - (A_{1,1} \cap A_{1,2} \cap A_{1,3} \cap A_{1,4})$ where

$$A_{1,2} = \left\{ \max_{1\leq i\leq N} \sum_{j=1}^{N} I_{\left\{L_{ij}=1, \widehat{L}_{ij}=0\right\}} > \varrho_T c_1(N) \right\},$$

$$A_{1,3} = \left\{ \max_{1\leq i\leq N} \sum_{j=1}^{N} I_{\left\{L_{ij}=0, \widehat{L}_{ij}=1\right\}} > \varrho_T c_1(N) \right\},$$

$$A_{1,4} = \left\{ \max_{1\leq i\leq N} \sum_{j=1}^{N} I_{\left\{\widehat{L}_{ij}=0, L_{ij}=0, |\widehat{r}_{ij}|>l\right\}} > \varrho_T c_1(N) \right\},$$

we have, because $l \le \lambda$,

$$\begin{aligned}\left\|\widehat{R}_{\widehat{L}}^{\mathcal{T}} - R_{\widehat{L}}^{\mathcal{T}}\right\| &\le c_1\left(N\right)\max_{i,j}\left|\widehat{r}_{ij} - r_{ij}\right| + 2\max_{1\le i\le N}\sum_{j=1}^{N}\left(I_{\left\{\widehat{L}_{ij}=1,L_{ij}=0\right\}} + I_{\left\{\widehat{L}_{ij}=0,L_{ij}=1\right\}}\right) + \max_{1\le i\le N}\sum_{j=1}^{N} I_{\left\{\widehat{L}_{ij}=0,L_{ij}=0,\left|\widehat{r}_{ij}\right|>\lambda\right\}}\\ &\le Ac_1\left(N\right)\sqrt{\frac{\log N}{T}} + 2\varrho_T c_1\left(N\right) + \max_{1\le i\le N}\sum_{j=1}^{N} I_{\left\{\widehat{L}_{ij}=0,L_{ij}=0,\left|\widehat{r}_{ij}\right|>l\right\}}\\ &\le A\left(c_1\left(N\right)\sqrt{\frac{\log N}{T}} + c_1\left(N\right)\varrho_T\right),\end{aligned}$$

which yields

$$P\left(\left\|\widehat{R}_{\widehat{L}}^{\mathcal{T}} - R_{\widehat{L}}^{\mathcal{T}}\right\| > A\left(c_1\left(N\right)\sqrt{\frac{\log N}{T}} + c_1\left(N\right)\varrho_T\right)\right) = O\left(\frac{1}{N^2} + \kappa_1 + \kappa_2\right) \tag{24}$$

by condition (b) in Assumption 3.

For the second part, we have

$$\begin{aligned}\left\|R_{\widehat{L}}^{\mathcal{T}} - R_{L}^{\mathcal{T}}\right\| &\le \max_{1\le i\le N}\sum_{j=1}^{N}\left(\left|r_{ij} - s_\lambda\left(r_{ij}\right)\right| I_{\left\{\widehat{L}_{ij}=1,L_{ij}=0\right\}} + \left|s_\lambda\left(r_{ij}\right) - r_{ij}\right| I_{\left\{\widehat{L}_{ij}=0,L_{ij}=1\right\}}\right)\\ &\le \max_{1\le i\le N}\sum_{j=1}^{N}\left(2\left|r_{ij}\right| I_{\left\{\widehat{L}_{ij}=1,L_{ij}=0\right\}} + \lambda I_{\left\{\widehat{L}_{ij}=0,L_{ij}=1\right\}}\right)\\ &\le 2l\max_{1\le i\le N}\sum_{j=1}^{N} I_{\left\{\widehat{L}_{ij}=1,L_{ij}=0\right\}} + \lambda\max_{1\le i\le N}\sum_{j=1}^{N} I_{\left\{\widehat{L}_{ij}=0,L_{ij}=1\right\}}.\end{aligned}$$

Next, under the set $A_{1,2}^c \cap A_{1,3}^c$,

$$\left\|R_{\widehat{L}}^{\mathcal{T}} - R_{L}^{\mathcal{T}}\right\| \le 2l\max_{1\le i\le N}\sum_{j=1}^{N} I_{\left\{\widehat{L}_{ij}=1,L_{ij}=0\right\}} + \lambda\max_{1\le i\le N}\sum_{j=1}^{N} I_{\left\{\widehat{L}_{ij}=0,L_{ij}=1\right\}} \le 3\cdot\varrho_T c_1\left(N\right).$$

From condition (b) in Assumption 3, we get

$$P\left(\left\|R_{\widehat{L}}^{\mathcal{T}} - R_{L}^{\mathcal{T}}\right\| > A\varrho_T c_1\left(N\right)\right) = O\left(\kappa_2\left(N,T\right)\right). \tag{25}$$

For the third part, with the property of function s_λ and class $\mathcal{U}_1$, we have

$$\begin{aligned} \left\| R_L^{\mathcal{T}} - R \right\| &\leq \max_{1\leq i\leq N} \sum_{j=1}^{N} \left(\left| r_{ij} - r_{ij} \right| I_{\{L_{ij}=1\}} + \left| s_\lambda \left(r_{ij} \right) - r_{ij} \right| I_{\{L_{ij}=0\}} \right) \\ &= \max_{1\leq i\leq N} \sum_{j=1}^{N} \left| s_\lambda \left(r_{ij} \right) - r_{ij} \right| I_{\{L_{ij}=0\}} \leq 2 \max_{1\leq i\leq N} \sum_{j=1}^{N} \left| r_{ij} \right| I_{\{L_{ij}=0\}} \\ &= 2 \max_{1\leq i\leq N} \sum_{j=1}^{N} \left| r_{ij} \right|^{1-q} \left| r_{ij} \right|^{q} \cdot I_{\{L_{ij}=0\}} \leq \lambda^{1-q} c_0 \left(N \right). \end{aligned} \tag{26}$$

Finally, collecting Equation 24, Equation 25 and Equation 26, we get

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

which gives the consistency of $\widehat{R}_{\widehat{L}}^{\mathcal{T}}$.

Now we return to $\mathbf{\Sigma}$. For the operator norm, $\left\| \widehat{D} - D \right\| = O\left(A\sqrt{\frac{\log N}{T}} \right)$ holds under $A_{1,1}^c$. Triangle inequality gives

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

The first term is bounded by $O\left(\left\| \widehat{R}_{\widehat{L}}^{\mathcal{T}} - R \right\| \right)$ provided $\sigma_{ii} < M$ in Equation 4 and the event $A_{1,1}^c$. For the second part, under $A_{1,1}^c$, we have

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

Hence, we obtain $P\left(\left\| \widehat{\mathbf{\Sigma}}_{\widehat{L}}^{\mathcal{T}} - \mathbf{\Sigma} \right\| > A \left(\left\| \widehat{R}_{\widehat{L}}^{\mathcal{T}} - R \right\| + \sqrt{\frac{\log N}{T}} \right) \right) = O\left(\frac{1}{N^2} + \kappa_1 \left(N, T \right) \right)$. In conclusion, since $c_1 \left(N \right) \to \infty$, we get

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

which ends the proof. □

A.2 Proof of Theorem 2

***Proof*.** By triangle inequality, we have $\left\|\widehat{R}_{\widehat{C}}^{\mathcal{B}} - R\right\| \leq \left\|\widehat{R}_{\widehat{C}}^{\mathcal{B}} - R_{\widehat{C}}^{\mathcal{B}}\right\| + \left\|R_{\widehat{C}}^{\mathcal{B}} - R_{C}^{\mathcal{B}}\right\| + \left\|R_{C}^{\mathcal{B}} - R\right\|$. For operator norm, 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 under the event $A_{1,1}^c$, one has $\left\|\widehat{R}_{\widehat{C}}^{\mathcal{B}} - R_{\widehat{C}}^{\mathcal{B}}\right\| \leq O\left(k\sqrt{\frac{\log N}{T}}\right)$, which yields

$$P\left(\left\|\widehat{R}_{\widehat{C}}^{\mathcal{B}} - R_{\widehat{C}}^{\mathcal{B}}\right\| > A \cdot k\sqrt{\frac{\log N}{T}}\right) = O\left(\frac{1}{N^2} + \kappa_1\left(N, T\right)\right). \tag{27}$$

For the second part, we have $\left\|R_{\widehat{C}}^{\mathcal{B}} - R_{C}^{\mathcal{B}}\right\| \leq \max_{1\leq i\leq N} \sum_{j=1}^{N} \left| b_{\widehat{C},k}\left(r_{ij}\right) - b_{C,k}\left(r_{ij}\right)\right|$. Here, note that $\left| b_{\widehat{C},k}\left(r_{ij}\right) - b_{C,k}\left(r_{ij}\right)\right|$ can only be 0 or $|r_{ij}|$. Specifically, there are two cases where the error is non-zero (and equals $|r_{ij}|$):

1. $(i,j) \in S_k^{c_j} \times S_k^{c_i}$, but $(i,j) \notin S_k^{\widehat{c}_j} \times S_k^{\widehat{c}_i}$;

2. $(i,j) \notin S_k^{c_j} \times S_k^{c_i}$, but $(i,j) \in S_k^{\widehat{c}_j} \times S_k^{\widehat{c}_i}$.

Therefore, we have

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

under event B_2^c where

$$B_2=\left\{\frac{1}{k}\sum_{j=1}^{N} I_{\left\{j\in S_k^{c_i},j\notin S_k^{\widehat{c_i}}\right\}}>A\sqrt{\frac{\log N}{T}}\right\}\cup\left\{\frac{1}{k}\sum_{j=1}^{N} I_{\left\{i\in S_k^{c_j},i\notin S_k^{\widehat{c_j}}\right\}}>A\sqrt{\frac{\log N}{T}}\right\}.$$

Thus with condition (b) in Assumption 4 we get

$$P\left(\left\|R^{\mathcal{B}}_{\widehat{C}}-R^{\mathcal{B}}_{C}\right\|>A\cdot\left(k\sqrt{\frac{\log N}{T}}+b_1\left(N\right)+b_0k^{-\alpha}\right)\right)=O\left(\kappa_3\left(N,T\right)\right). \tag{28}$$

For the third part, we have

$$\begin{aligned}\left\|R^{\mathcal{B}}_{C}-R\right\|&\leq\max_{1\leq i\leq N}\sum_{j=1}^{N}\left|b_{C,k}\left(r_{ij}\right)-r_{ij}\right|\\&\leq\max_{1\leq i\leq N}\sum_{j=1}^{N}\left|r_{ij}\right|\left(I_{\left\{i\notin S_k^{c_j}\right\}}+I_{\left\{j\notin S_k^{c_i}\right\}}\right)\\&\leq\max_{1\leq i\leq N}\sum_{j=1}^{N}\left|r_{ij}\right|I_{\left\{i\notin S_k^{c_j},j\in S_k^{c_i}\right\}}+\max_{1\leq i\leq N}\sum_{j=1}^{N}\left|r_{ij}\right|I_{\left\{j\notin S_k^{c_i}\right\}}\\&\leq b_1\left(N\right)+b_0\left(N\right)k^{-\alpha}.\end{aligned} \tag{29}$$

Combining Equation 27, Equation 28 and Equation 29 one may get

$$P\left(\left\|\widehat{R}^{\mathcal{B}}_{\widehat{C}}-R\right\|>A\left(k\sqrt{\frac{\log N}{T}}+b_0\left(N\right)k^{-\alpha}+b_1\left(N\right)\right)\right)=O\left(\frac{1}{N^2}+\kappa_1\left(N,T\right)+\kappa_3\left(N,T\right)\right).$$

Then similar to the threshold estimator, the consistency of $\widehat{\mathbf{\Sigma}}^{\mathcal{B}}_{\widehat{C}}$ can be derived under event $A_{1,1}^c$, which is $\left\|\widehat{\mathbf{\Sigma}}^{\mathcal{B}}_{\widehat{C}}-\mathbf{\Sigma}\right\|\leq A\left(\left\|\widehat{R}^{\mathcal{B}}_{\widehat{C}}-R\right\|+\sqrt{\frac{\log N}{T}}\right)$. Finally, since $k=k_T\to\infty$, we obtain

$$P\left(\left\|\widehat{\mathbf{\Sigma}}^{\mathcal{B}}_{\widehat{C}}-\mathbf{\Sigma}\right\|>A\left(k\sqrt{\frac{\log N}{T}}+b_0\left(N\right)k^{-\alpha}+b_1\left(N\right)\right)\right)=O\left(\frac{1}{N^2}+\kappa_1\left(N,T\right)+\kappa_3\left(N,T\right)\right),$$

which ends the proof. □

A.3 Proof of Corollary 1

***Proof*.** Let $\mathbf{D}_T = \widehat{\mathbf{\Sigma}}_f - \mathbf{\Sigma}_f$, $\mathbf{C}_T = \widehat{\boldsymbol{B}} - \boldsymbol{B}$, we then have

$$\left\|\widehat{\mathbf{\Sigma}}_y - \mathbf{\Sigma}_y\right\|_E^2 \leq A\left(\left\|\boldsymbol{B}\mathbf{D}_T\boldsymbol{B}^{\intercal}\right\|_E^2 + \left\|\boldsymbol{B}\widehat{\mathbf{\Sigma}}_f\mathbf{C}_T^{\intercal}\right\|_E^2 + \left\|\mathbf{C}_T\mathbf{\Sigma}_f\mathbf{C}_T^{\intercal}\right\|_E^2 + \left\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\right\|_E^2\right), \tag{30}$$

for some constant A. Under our Assumptions 1 and 2, one can show

$$\begin{aligned} &P\left(\left\|\boldsymbol{B}\mathbf{D}_T\boldsymbol{B}^{\intercal}\right\|_E^2 + \left\|\boldsymbol{B}\widehat{\mathbf{\Sigma}}_f\mathbf{C}_T^{\intercal}\right\|_E^2 > A\cdot\left(\frac{K\log N}{T} + \frac{K^2\log T}{NT}\right)\right) = O\left(\frac{1}{N^2}\right), \\ &P\left(\left\|\mathbf{C}_T\mathbf{\Sigma}_f\mathbf{C}_T^{\intercal}\right\|_E^2 > A\cdot\frac{K^2N\left(\log N\right)^2}{T^2}\right) = O\left(\frac{1}{N^2}\right). \end{aligned}$$

The proof can be found in Lemma B.3 of Fan et al. (2011). And for the part $\left\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\right\|_E^2$, we have

$$\left\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\right\|_E^2 = \frac{1}{N}\left\|\mathbf{\Sigma}_u^{-\frac{1}{2}}\left(\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\right)\mathbf{\Sigma}_u^{-\frac{1}{2}}\right\|_F^2 \leq \frac{\rho_{\max}\left(\mathbf{\Sigma}_u^{-1}\right)^2}{N}\left\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\right\|^2,$$

which is discussed before for both thresholding and banding estimators.

For the **Network Guided Thresholding** estimator, if $\widehat{\mathbf{\Sigma}}_{\widehat{L}}^{\mathcal{T}}$ attains the best convergence rate $c_0\left(N\right)\left(\sqrt{\frac{\log N}{T}} + \varrho_T\right)$, then we have

$$P\left(\left\|\widehat{\mathbf{\Sigma}}_y - \mathbf{\Sigma}_y\right\|_E > A\left(K\frac{\sqrt{N}\log N}{T} + \sqrt{K}\sqrt{\frac{\log N}{T}} + \frac{c_0\left(N\right)}{\sqrt{N}}\left(\sqrt{\frac{\log N}{T}} + \varrho_T\right)\right)\right) = O\left(\frac{1}{N^2} + \kappa_{1,2}\right),$$

where $\kappa_{1,2} := \kappa_1\left(N,T\right) + \kappa_2\left(N,T\right)$.

For the **Network Guided Banding** estimator, if $\widehat{\mathbf{\Sigma}}_{\widehat{C}}^{\mathcal{B}}$ attains the best convergence rate $\left(1 + b_0\left(N\right)\right)\left(\frac{\log N}{T}\right)^{\frac{\alpha}{2(\alpha+1)}} + b_1\left(N\right)$, then we have

$$P\left(\left\|\widehat{\mathbf{\Sigma}}_y - \mathbf{\Sigma}_y\right\|_E > A\left(K\frac{\sqrt{N}\log N}{T} + \sqrt{K}\sqrt{\frac{\log N}{T}} + \frac{\left(1 + b_0\left(N\right)\right)}{\sqrt{N}}\left(\frac{\log N}{T}\right)^{\frac{\alpha}{2(\alpha+1)}} + \frac{b_1\left(N\right)}{\sqrt{N}}\right)\right) = O\left(\frac{1}{N^2} + \kappa_{1,3}\right),$$

where $\kappa_{1,3} := \kappa_1\left(N,T\right) + \kappa_3\left(N,T\right)$. □