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

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

February 16, 2026

Abstract: This paper uses the *auxiliary network information*, observed in addition to the original sample, to infer latent network structures in the population correlation matrix and thus improve high-dimensional covariance matrix estimation. Building on estimated Location Indicator and Relative Importance matrices, we propose two *Network-Guided* estimators. Network-Guided Thresholding uses auxiliary network data to regularize the large and small elements in the sample covariance matrices differentially, delivering a faster convergence rate over a more general class of sparse covariance matrices when auxiliary information is informative. Network-Guided Banding extends the banding estimators to allow for data without a natural ordering, using the relative importance of elements indicated by the auxiliary datasets to construct a neighbor ordering, which can achieve the optimal convergence rate that would be infeasible without the auxiliary network information. An extensive simulation studies show robust finite-sample gains of the proposed Network-Guided estimators over existing benchmark methods. The proposed methods also deliver superior out-of-sample performance relative to the established baseline models in the empirical application of constructing Global Minimum Variance (GMV) and Mean-Variance Optimal (MVO) portfolios in the Chinese stock market with various sources of auxiliary network information, including analyst co-coverage, news co-mentions, and industry classifications.

Keywords: Auxiliary network information; high-dimensional covariance matrix; thresholding; banding; portfolio management

JEL Classification: C13, C58, G11

*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. Wen Su is supported by the EPSRC CDT in Mathematics of Random Systems (EP/S023925/1).

[†]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

1 Introduction

Covariance matrix estimation is a central problem in statistics and econometrics. Given a dataset $\mathcal{X}(N,T) = \{\boldsymbol{X}_t : t = 1,\dots,T\}$ of stationary N-dimensional vectors, such as N asset returns observed over T periods, the target is to estimate the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{X}_t$. The sample covariance estimator is ill-conditioned in the high dimensionality regime when N is large relative to T. This motivates a large literature on regularized estimation under additional structural assumptions such as $\boldsymbol{\Sigma}$ being sparse or bandable. Most existing studies focus on estimating $\boldsymbol{\Sigma}$ using $\mathcal{X}(N,T)$ alone. However, in the big-data era, we often have access to auxiliary datasets $\mathcal{I}(N,T)$ in addition to $\mathcal{X}(N,T)$ that also contain information about $\boldsymbol{\Sigma}$. For example, stocks that are co-mentioned in business news exhibit excess co-movement beyond standard risk factors (Ge et al., 2022), which can help identify large entries in the covariance matrix. This paper therefore asks a natural question: can we improve upon existing high-dimensional covariance estimators by using *both* the original dataset $\mathcal{X}(N,T)$ and the auxiliary information $\mathcal{I}(N,T)$, and if so, how? We find a positive answer to the question and propose estimators that exploit the auxiliary network information to help tackle the challenges posed by the high-dimensionality of the covariance estimation problem.

In particular, this paper proposes generalizations of the bandable and sparse covariance classes and improves the corresponding estimators when auxiliary network information is available. We develop *network-guided* covariance estimators that integrate the auxiliary data $\mathcal{I}(N,T)$ with the original dataset $\mathcal{X}(N,T)$ to enhance classical thresholding and facilitate banding.

We build on the literature showing that the equity return covariance matrix of financial assets encodes economically meaningful network structure (see, e.g., Tumminello et al. (2010), Billio et al. (2012), Fan et al. (2016a), Zhang et al. (2025)). In this paper, we define two latent network structures implied by the population correlation matrix that parallel the sparse and bandable covariance classes of Bickel and Levina (2008a) and Bickel and Levina (2008b): one links asset pairs whose correlations are large in magnitude, while the other links each asset to its most strongly correlated neighbors.

In practice, the latent network structures are unobserved, but many datasets contain economic-linkage information that is informative about these latent networks.[1] Beyond the news co-mention example, Israelsen (2016) found that stocks covered by similar sets of analysts co-move strongly, and Hoberg and Phillips (2016) used textual analysis of firms' 10-K reports

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

to construct peer groups of fundamentally similar firms that also tend to co-move. We refer to these economically motivated linkages as auxiliary network information: a data set $\mathcal{I}(N,T)$ that is distinct from the return sample and can be used to estimate the latent network structure.

Our estimation framework is built precisely to exploit this type of auxiliary information in large covariance matrix estimation. We explicitly model the probability that the auxiliary network correctly identifies (i) *absolutely important* links—via entrywise false positives and false negatives—and (ii) *relatively important* links—via errors in the neighbor-ranking structure used by banding. These probabilities are allowed to depend on both the quality of the auxiliary network and the underlying true correlation structure of asset returns. Combining the auxiliary information $\mathcal{I}(N,T)$ and the original sample $\mathcal{X}(N,T)$, the paper proposes two complementary Network-Guided estimators; we establish their theoretical properties and demonstrate strong finite-sample performance in simulations and empirical applications.

The first method is *Network-Guided Thresholding*, which is applicable when auxiliary information identifies the location of large elements in the covariance matrix. 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), relying only on $\mathcal{X}(N,T)$. In contrast, we use auxiliary network information to guide the thresholding: we first locate a set of entries that are likely to be large according to the auxiliary network and retain them, and then apply a generalized thresholding rule to the remaining entries. We establish theoretical results for the proposed Network-Guided Thresholding estimator. Relative to the universal thresholding framework of Bickel and Levina (2008a), we accommodate a more general class of sparse covariance matrices by using auxiliary network information to distinguish between large and small elements and to quantify their behaviors separately. When the auxiliary network is sufficiently informative, this separation yields improved convergence results compared to purely data-driven thresholding. Industry information is an example of such auxiliary information, as it implies a block-diagonal network where every node is connected within an industry. Fan et al. (2016a) used industry sector information to implement a location-based thresholding that treats within-sector correlations as large and across-sector correlations as negligible. Our approach allows for general auxiliary networks beyond industry blocks and does not require the idiosyncratic covariance under factor models to be block diagonal, so Network-Guided Thresholding can exploit richer information and accommodate more general dependence structures.

The second method we propose is called *Network-Guided Banding*. Bickel and Levina (2008b) studied a class of bandable matrices, whose elements get 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 directly 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 our 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. In our equity return covariance setting, sector/industry networks provide only unweighted linkage information—they indicate whether two stocks are linked but not how strong the link is—so they can be used for Network-Guided Thresholding but not for Network-Guided Banding. By contrast, analyst co-coverage networks (Israelsen, 2016), news co-mention networks (Ge et al., 2022), and text-based product networks (Hoberg and Phillips, 2016) are weighted: they assign different strengths to each connection and thus allow us to rank neighbor importance (e.g., by the frequency of analyst co-coverage or news co-mentions; see also Scherbina and Schlusche (2015); Schwenkler and Zheng (2019)). When such weighted auxiliary networks are available, we can apply both Network-Guided Thresholding and Network-Guided Banding via ranking the strength of connection of each entity's neighbors. We show that the Network-Guided Banding estimator attains the optimal convergence rate subject to errors in the ranking over a larger class of covariance matrices satisfying a generalized bandable condition.

In Monte Carlo experiments, we generate returns 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 an extensive set of established benchmarks. 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 use Chinese equity data and the CH-4 factor model (Liu et al., 2019) to construct Global Minimum Variance (GMV) and Mean–Variance Optimal (MVO) portfolios with our Network-Guided estimators. We consider three auxiliary networks: weighted news co-mentions (passage- and sentence-level, following Ge et al., 2025), weighted analyst co-coverage (by the number of shared analysts), and standard industry classifications. To guard against

low-quality auxiliary information, we implement a cross-validation safeguard that automatically turns off network guidance and reverts to the classical benchmark when the auxiliary network is uninformative. We evaluate out-of-sample portfolio performance across HS300, CSI500, and CSI800, and find that incorporating auxiliary network information improves the out-of-sample performance of both GMV and MVO portfolios relative to conventional covariance estimators.

Our contributions are threefold. First, we formalize a general framework for incorporating auxiliary information as an additional source that enhances conventional statistical procedures while remaining compatible with classical high-dimensional asymptotic analysis. A key feature is that we explicitly model the quality of the auxiliary information and give a set of mild conditions under which it improves estimation, allowing us to combine the original dataset $\mathcal{X}(N,T)$ and the auxiliary dataset $\mathcal{I}(N,T)$ in a way that outperforms traditional methods. This perspective is conceptually distinct from traditional "transfer" approaches that primarily exploit information extracted from the same data source—such as representation-based transfer learning for vector autoregressions (Li et al., 2020; Lin et al., 2025). In many financial applications, the big-data environment also means that *additional* datasets—often much larger and richer than the original dataset—have become widely available and economically informative. Second, we accordingly develop *Network-Guided* estimators for high-dimensional covariance matrices that augment classical thresholding and banding with auxiliary network information, providing a practical, theoretically grounded framework in which network guidance strengthens thresholding when it reliably highlights strong links and makes banding-type regularization feasible even without a natural ordering. Third, we demonstrate the empirical value of the proposed estimators in portfolio allocation, and we provide an implementable diagnostic/selection rule that allows practitioners to downweight or disregard auxiliary network information when it does not improve performance, ensuring that network guidance is beneficial rather than binding. The out-of-sample performance gains from the proposed methods are economically significant.

Related literature. A growing body of work studies high-dimensional covariance estimation using only observations $\{\boldsymbol{X}_t\}_{t=1}^T$. For sparsity-based approaches, Bickel and Levina (2008a) developed a theory for universal thresholding under sparsity conditions, and subsequent work refined thresholding by allowing entry-adaptive thresholds and by incorporating factor structures with observed or latent factors (see Cai and Liu (2011), Fan et al. (2011), and Fan et al. (2013)). Another strand of the literature regularizes the sample covariance via spectral correc-

tions rather than sparsity. For example, Ledoit and Wolf (2004) and Ledoit and Wolf (2012) proposed linear and nonlinear shrinkage estimators that shrink the sample eigenvalues. As noted by Lam (2019), shrinkage estimators are typically well-conditioned and automatically positive definite, but they can be less effective when dependence is genuinely sparse or highly structured.

Related work imposes structure beyond entrywise sparsity. For example, Bickel and Levina (2008b) studied banding and tapering under a bandable assumption, where covariances decay away from the diagonal and a meaningful ordering (e.g., time or space) is available. However, such restrictions are often unavailable in cross-sectional finance, and although seriation and permutation-recovery methods can infer latent orderings (see, e.g., Flammarion et al., 2019; Giraud et al., 2023), they remain fully data-driven and are not yet standard in covariance estimation. We argue that when N is large relative to T, relying only on $\{\boldsymbol{X}_t\}$ to learn dependence can be suboptimal when informative additional data are available.We therefore exploit structural information about the underlying covariance matrix using auxiliary network information. We quantify the quality of these auxiliary networks within a network-formation framework, in which the probability of observing an edge in the auxiliary network depends on the magnitude of the correlation between the corresponding entities. Research on network formation dates back to the seminal Erdős-Rényi model, (Erdős et al., 1960). A growing literature studies network formation with degree heterogeneity; see, for example, Rinaldo et al. (2013) and Chen et al. (2021) on the estimation of β and sparse β model, Bickel and Chen (2009) on stochastic block models, and Handcock et al. (2007) on latent position cluster models.

Our paper bridges these strands of literature by treating auxiliary network information as an additional input to covariance regularization. In particular, our network guidance strengthens thresholding when auxiliary links identify globally strong dependence entries, and it makes banding-type regularization feasible even without a natural ordering.

The remainder of this paper is structured as follows. In Section 2, we introduce two latent network structures implied by the population correlation matrix and present a general framework that links these structures to auxiliary network information. Building on this framework, Section 3 proposes two new estimators: the Network-Guided Thresholding estimator and the Network-Guided Banding estimator. In Section 4, we lay out the assumptions and present the theoretical results. Section 5 presents simulation studies that compare our proposed estimators with established baseline methods, while Section 6 provides an empirical application. Finally, Section 7 concludes. The Appendix contains proofs and supplementary analysis. 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 $\boldsymbol{A} = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m) \in \mathbb{R}^{m\times d}$, $\|\boldsymbol{A}\|_F$ denotes the matrix Frobenius norm, i.e., $\|\boldsymbol{A}\|_F = (\|\boldsymbol{a}_1\|^2 + \cdots + \|\boldsymbol{a}_m\|^2)^{1/2}$; $\|\boldsymbol{A}\| = \inf\left\{c > 0 : \|\boldsymbol{A}\boldsymbol{x}\| \le c\,\|\boldsymbol{x}\|,\ \text{for all } \boldsymbol{x} \in \mathbb{R}^d\right\}$ is the operator norm. For two real-valued sequences $\{a_T\}$ and $\{b_T\}$, $a_T = o(b_T)$ implies $a_T/b_T \to 0$ when $T \to \infty$; $a_T = O(b_T)$ implies there exists some constant A, s.t. $a_T \le Ab_T$ for all T; $a_T \asymp b_T$ means $0 < c < a_T/b_T < C < \infty$ for some constants c and C. We use $[a_{ij}]_{m\times n}$ to denote an $m \times n$ matrix whose (i, j)-th entry is a_{ij}. Let $\boldsymbol{I}_{N\times N}$ be the $N \times N$ identity matrix, and let $\boldsymbol{1}_{N\times N}$ and $\boldsymbol{1}_{N\times 1}$ denote the $N \times N$ all-ones matrix and the $N \times 1$ all-ones vector, respectively.

2 Covariance Matrix and Network Information

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 covariance matrix $E[(\boldsymbol{X}_t - \boldsymbol{\mu})(\boldsymbol{X}_t - \boldsymbol{\mu})^\intercal] = \boldsymbol{\Sigma} = [\sigma_{ij}]_{N\times N}$. The corresponding correlation matrix is $\boldsymbol{R} = [r_{ij}]_{N\times N}$, with $\boldsymbol{\Sigma} = \boldsymbol{DRD}$, where $\boldsymbol{D} = \text{diag}\{\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{NN}}\}$. The sample covariance estimator is

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

where $\bar{\boldsymbol{X}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{X}_t$. Methods in the literature have focused on the regularization of $\hat{\boldsymbol{\Sigma}}$, which only depends on the samples $\mathcal{X}(N, T) = \{\boldsymbol{X}_t : t = 1, \ldots, T\}$ in the high-dimensional settings.

We argue that when there exists auxiliary information $\mathcal{I}(N, T)$ that complements $\mathcal{X}(N, T)$ and is informative about the true correlation matrix $\boldsymbol{R}$, such additional information should be exploited when estimating $\boldsymbol{\Sigma}$. In this paper, we take the financial market as our main setting to present the methods, theory, and applications: $\boldsymbol{X}_t$ denotes observations of asset returns, and the auxiliary information $\mathcal{I}(N, T)$ includes industry classifications, analyst co-coverage, news co-mentions, and other economic linkages that are informative about return co-movement.

In this section, we first define two types of latent network structures implied by the population correlation matrix and then introduce their estimated counterparts constructed from auxiliary network information. We then model the edge formation process via an observation probability function and specify the corresponding regularity conditions.

2.1 Latent Network Structures

The literature has documented that the correlation matrix $\boldsymbol{R} = [r_{ij}]_{N\times N}$ of financial assets encodes latent network structures (Tumminello et al. (2010), Billio et al. (2012), Fan et al. (2016a), Zhang et al. (2025)). In this subsection, we define two latent network structures implied by the true correlation matrix that are consistent with the two classes of covariance matrices considered in Bickel and Levina (2008a) and Bickel and Levina (2008b).

Definition 1 (Latent Network Structure from Population Correlation Matrix)**.** *Let $G = (V, E)$ denote an undirected graph with vertex set $V = \{1, \ldots, N\}$. Given the population correlation matrix $\boldsymbol{R}$, we define:*
(i) The edge set E is determined by the magnitude of correlations,

$$E^L = \{(i, j) : |r_{ij}| > \lambda\}$$

for some threshold $\lambda \geq 0$. In this case, two assets are connected if their absolute correlation is above a certain threshold.
(ii) The edge set E is determined by the rank ordering of correlations,

$$E^C = \{(i, j) : c_{ij} > N - k \text{ and } c_{ji} > N - k\},$$

for some integer $k > 0$, where c_{ij} denotes the rank of $|r_{ij}|$ among $\{|r_{i1}|, \ldots, |r_{iN}|\}$.[2] *In other words, $(i, j) \in E^C$ if and only if i and j are among each other's k most strongly correlated assets.*

The first network links pairs whose correlations exceed a fixed threshold λ, while the second links each asset to its k most strongly correlated neighbors. These two constructions mirror the sparse and bandable covariance classes (Bickel and Levina (2008a,b)), which are driven by absolute thresholds and relative ranks individually.

To model and estimate these two types of latent graphs using auxiliary information, we associate them with matrices that encode absolute and relative correlation strength, respectively. For the first type, which is based on correlation magnitudes, we define the *Location Indicator*

[2]We assign mean ranks in the case of tied values, following the standard approach used in nonparametric tests such as the Wilcoxon and Mann–Whitney tests (see Lehmann (2006)). For example, the values 5, 7, 7, 8, 9 receive the ranks 1, 2.5, 2.5, 4, and 5, respectively.

Matrix:

$$\boldsymbol{L} = [L_{ij}]_{N\times N}, \quad L_{ij} = I_{\{|r_{ij}|>\lambda\}} = \begin{cases} 1, & |r_{ij}| > \lambda, \\ 0, & |r_{ij}| \leq \lambda, \end{cases} \tag{2}$$

for a threshold parameter λ. Each entry L_{ij} thus takes value 0 or 1, indicating whether there is an edge between assets i and j, that is, $L_{ij} = I_{\{(i,j)\in E^L\}}$ as defined in Definition 1. We also define the complement matrix $\boldsymbol{L}^0 = [L^0_{ij}]_{N\times N} = \mathbf{1}_{N\times N} - \boldsymbol{L}$, where $\mathbf{1}_{N\times N}$ is the $N \times N$ matrix with all entries equal to 1.

As for the second type, we define the *Relative Importance Matrix* $\boldsymbol{C} = [c_{ij}]_{N\times N}$. A typical entry c_{ij} is the ascending rank of $|r_{ij}|$ in the vector $\text{abs}(\boldsymbol{r}_i) := (|r_{i1}|, \ldots, |r_{iN}|)^\intercal$, which captures the relative importance of j for asset i. We use $S_k^{\boldsymbol{c}_i}$ to denote the set of the k most strongly correlated neighbors of asset i. Hence,

$$\{(i,j) \in E^C\} \quad \Longleftrightarrow \quad \{i \in S_k^{\boldsymbol{c}_j}\} \cap \{j \in S_k^{\boldsymbol{c}_i}\},$$

where E^C is defined in Definition 1.

Both network structures, encoded in the Location Indication Matrix $\boldsymbol{L}$ and the Relative Importance Matrix $\boldsymbol{C}$, are *latent* since they are implied by the unknown correlation matrix $\boldsymbol{R}$. Fortunately, we have datasets other than asset returns that contain information about the latent network structures. In the next subsection, we introduce *observable* auxiliary network information and show how it can be used to estimate these latent network structures.

2.2 Auxiliary Network Information

After removing common factors, stocks with economic linkages tend to display larger residual correlations; we discuss this evidence in more detail in the empirical section. Importantly, these economic linkages are not obtained from the asset return sample $\mathcal{X}(N,T) = \{\boldsymbol{X}_t,\ t = 1, \ldots, T\}$ but from additional sources $\mathcal{I}(N,T)$, such as industry classifications (Fan et al., 2016a), analyst co-coverage (Israelsen, 2016), and news co-mentions (Ge et al., 2022). We therefore refer to these linkages as *auxiliary network information*. Since $\mathcal{I}(N,T)$ can take many forms, we first introduce the abstract concept of auxiliary network information and then provide several natural examples.

Definition 2 (Auxiliary Network Information)**.** *The auxiliary network information $\mathcal{I}(N,T)$ is the dataset observed in addition to the asset returns $\mathcal{X}(N,T)$, which can be used to infer about the latent network structure and construct estimators $\widehat{\boldsymbol{L}}$ and $\widehat{\boldsymbol{C}}$.*

For example, industry affiliation provides a simple form of auxiliary network information. We set $M_{ij} = 1$ if firms i and j belong to the same industry, and $M_{ij} = 0$ otherwise. Such a binary network is naturally suited for estimating the latent adjacency matrix $\boldsymbol{L}$: one may simply define $\widehat{L}_{ij} = 1$ whenever $M_{ij} = 1$. However, since $\boldsymbol{M}$ records only the presence or absence of a link, it does not carry direct information about the strength of the connection. Consequently, it may be more suitable use such information to construct the Location Indicator $\boldsymbol{L}$ than the Relative Importance Matrix $\boldsymbol{C}$.

By contrast, some auxiliary network datasets (such as analyst co-coverage or news co-mentions) are informative about the strength of the connection between i and j. Let M_{ij} denote the number of analysts (or news items) that cover (or mention) both stocks i and j within a given time window. In this case, $\boldsymbol{M}$ can be used to estimate both $\boldsymbol{L}$ and $\boldsymbol{C}$, since a larger M_{ij} implies a stronger connection. Specifically, given a threshold m, we may construct the adjacency estimator $\widehat{L}_{ij} = I_{\{M_{ij}>m\}}$. To estimate $\boldsymbol{C}$, one convenient choice is to set $\widehat{c}_{ij}$ as the row rank of M_{ij} among $\{M_{i1}, \ldots, M_{iN}\}$.

Remark: First, our definition of auxiliary network information ensures that it contributes knowledge beyond what is contained in the return sample. It also rules out using the sample correlation matrix itself to estimate $\widehat{\boldsymbol{L}}$ or $\widehat{\boldsymbol{C}}$, since it is fully determined by $\{\boldsymbol{X}_t\}$ and therefore adds no information beyond what is already in the data.[3] In particular, when $\mathcal{I}(N,T)$ is of high quality, the random $\widehat{\boldsymbol{L}}$ will be close to $\boldsymbol{L}$ with higher probability. We will further discuss these observation probabilities in the next subsection. For simplicity, we use the notation $\mathcal{I}(N,T)$ to indicate that auxiliary information is available for the N assets over the sampling period $1, ..., T$. In practice, however, the auxiliary dataset can have a much larger sample size than $\mathcal{X}(N,T)$.

2.3 Observation Probability

2.3.1 Estimated Location Indicator Matrix

We use the observable auxiliary network data to estimate the location indicator matrix $\boldsymbol{L}$. We model the probability that the estimated location indicator matrix records a link between assets i and j, $P\left(\widehat{L}_{ij} = 1\right)$, via an observation probability function $g_T : [0,1] \to [0,1]$,

$$P\left(\widehat{L}_{ij} = 1\right) = g_T(|r_{ij}|)\,, \qquad P\left(\widehat{L}_{ij} = 0\right) = 1 - g_T(|r_{ij}|)\,.$$

[3] In our simulation and empirical analyses, we nonetheless include the sample correlation matrix as an auxiliary network, referred to as self-banding, as a benchmark for comparison.

Intuitively, firms with stronger correlations tend to have a higher probability of having an edge in the auxiliary network. The explicit form of g_T need not be specified; it suffices that g_T satisfies certain regularity conditions. g_T may be modeled by, for example, the β-model, latent space models, and other edge formation models.[4] For illustration, we present some representative specifications in Figure 1, where we show how observation probability changes as parameter changes in a piecewise power law model (left) and a latent space model (middle).

Given threshold λ, the optimal auxiliary network has observation probability $g_T^*(x) = 0$ for $x \leq \lambda$ and $g_T^*(x) = 1$ for $x > \lambda$, i.e., the optimal auxiliary network perfectly separates the small and large entries in the correlation matrix with certainty, see the right panel of Figure 1.

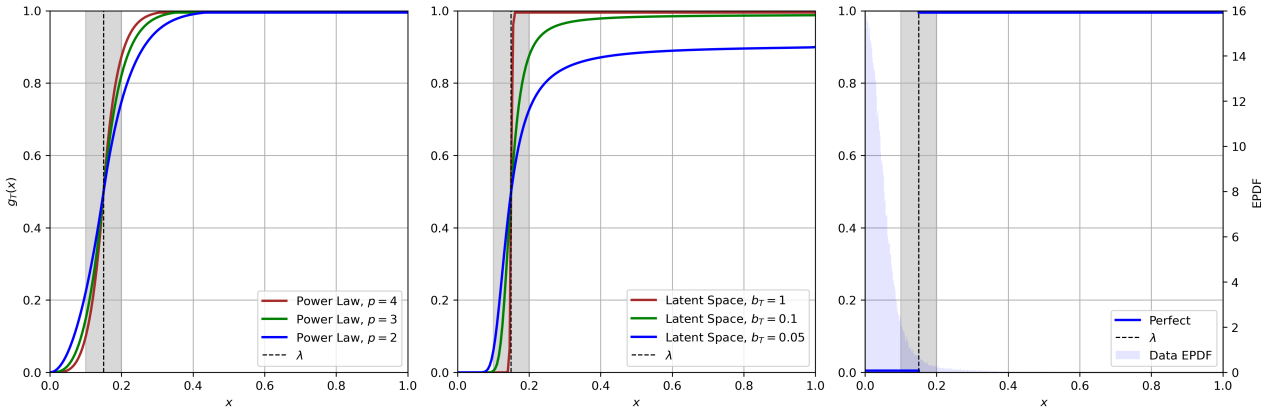


Figure 1: Representative specifications of g_T. We set $\lambda = 0.15$. The (piecewise) power-law family is defined by $0.5x^p I_{\{x\leq\lambda\}} + (1 - 0.5\,(1-x)^q)\, I_{\{x>\lambda\}}$, with $p \in \{2, 3, 4\}$ where the exponent p applies when $x \leq \lambda$, while for $x > \lambda$ the power q is chosen automatically to ensure g_T is smooth. The latent-space specification follows Handcock et al. (2007) and is given by $g_T(x) = \frac{\exp\{b_T(d(\lambda)-d(|r_{ij}|))\}}{1+\exp\{b_T(d(\lambda)-d(|r_{ij}|))\}}$, where $d(r) = -\left[\log(1-r^2)\right]^{-1}$ is the inverse Kullback–Leibler divergence distance (see Appendix A.3 for details). The light blue shaded region in the right panel displays the empirical probability density function (EPDF) of the absolute sample correlation between firms' idiosyncratic returns.

We can then measure the quality of any given $\widehat{\boldsymbol{L}}$ constructed from the auxiliary network information by the difference between the observation probability g_T associated with $\widehat{\boldsymbol{L}}$ and the optimal g_T^*. There are two possible errors in $\widehat{\boldsymbol{L}}$, false positives and false negatives. Specifically, we assume that for some positive $\nu_N = o(\lambda)$, there exist $\kappa_0(N,T)$, ϱ_{0T} and ϱ_{1T}, such that

$$|g_T^*(x) - g_T(x)| = \begin{cases} g_T(x) = O(\varrho_{0T}), & x < \lambda - \nu_N, \\ 1 - g_T(x) = O(\varrho_{1T}), & x > \lambda + \nu_N, \end{cases} \tag{3}$$

$$Q_N(\lambda - \nu_N, \lambda + \nu_N) = O(\kappa_0(N,T)c_0(N))\,,$$

[4] See Rinaldo et al. (2013) and Handcock et al. (2007) for detailed discussions.

where $c_0(N)$ is a parameter for the overall magnitude of the correlation matrix (see Equation 6 for more details), and the parameter ν_N is a technical device introduced to handle the discontinuity in the optimal observation probability g_T^*. When g_T is continuous, there is a "gray region" around λ in which entries are difficult to classify correctly. Here $Q_N(\lambda - \nu_N, \lambda + \nu_N) = \max_i \sum_{j=1}^{N} I_{\{\lambda-\nu_N<|r_{ij}|\leq\lambda+\nu_N\}}$ denotes the maximum (row-wise) number of elements lying within a shrinking gray region of radius $\nu_N = o(\lambda)$. If too many elements concentrate near λ, the methodology may fail since it is difficult for the auxiliary network to clearly identify the true underlying latent network structure. The false positive rate ϱ_{0T} and the false negative rate ϱ_{1T} will determine the quality of the auxiliary information and affect the statistical properties of the proposed estimators, which will be discussed in Section 4.

2.3.2 Estimated Relative Importance Matrix

Similarly, the discrepancy between $\widehat{\boldsymbol{C}}$ and $\boldsymbol{C}$ introduces error. Analogous to the thresholding framework, we assume the existence of functions g_T such that

$$P\left(j \in S_k^{\widehat{c}_i}\right) = g_T(c_{ij}), \qquad P\left(j \notin S_k^{\widehat{c}_i}\right) = 1 - g_T(c_{ij}),$$

where c_{ij} denotes the rank of $|r_{ij}|$ in the set $\text{abs}(\boldsymbol{r}_i)$. Here, $g_T : \{1, 2, \ldots, N\} \to [0, 1]$, whereas in the thresholding framework g_T is defined on $[0, 1]$. The key difference is that, in thresholding, the observation probability that a pair is classified as important depends directly on the correlation coefficient, while here it depends on the *rank* of the correlation coefficient, reflecting that the auxiliary data provide relative information about the neighbor importance.

With rank statistics, if there are no ties, then there is no "gray region". Moreover, once we control misclassification on one side, the error on the other side vanishes automatically. Accordingly, we only assume that

$$1 - g_T(m) = O(\varrho_{1T}), \quad \forall m \geq N - k + 1,$$

for some sequence ϱ_{1T} that measures the errors in the ranking. A higher-quality network corresponds to a smaller ϱ_{1T}. We also provide representative examples in Figure 2, where both the piecewise power law and the latent space models converge rapidly to the perfect case.

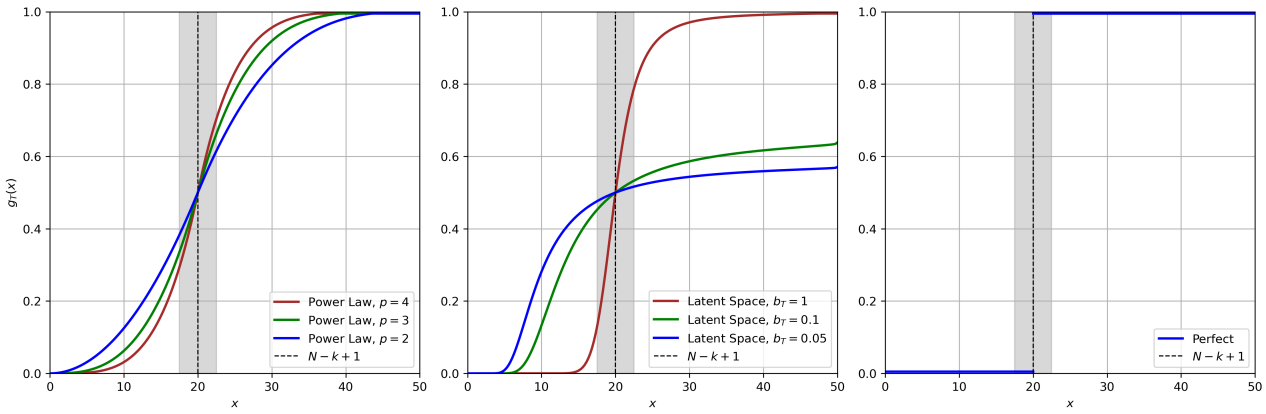


Figure 2: Representative specifications for g_T. We set $N = 50$ and $k = 31$. In the piecewise power law model, p denotes the power for $m < N-k+1$ while the power for $m \geq N-k+1$ is chosen automatically to ensure that g_T is smooth. In the latent space model, the distance is specified as $d_{ij} = -[\log(1-(c_{ij}/N)^2)]^{-1}$.

3 Network-Guided Estimators

Given the estimated Location Indicator Matrix and Relative Importance Matrix, we propose two methods in this section, Network-Guided Thresholding and Network-Guided Banding, which incorporate these two network structures, respectively. By combining asset return data $\boldsymbol{X}_t$ with auxiliary network data that help reveal the latent structure of the population correlation matrix, we show that these methods achieve superior theoretical and empirical performance.

3.1 Network-Guided Thresholding

Suppose $\boldsymbol{\Sigma} = [\sigma_{ij}]_{N\times N}$ is a covariance matrix and $\boldsymbol{R} = [r_{ij}]_{N\times N}$ is the corresponding correlation matrix.[5] We first recall the classical thresholding estimator introduced in Bickel and Levina (2008a):

$$T_\lambda\left(\widehat{\boldsymbol{R}}\right) = [s_\lambda(\widehat{r}_{ij})]_{N\times N}\,, \quad \text{where} \quad \widehat{r}_{ij} = \frac{\widehat{\sigma}_{ij}}{\sqrt{\widehat{\sigma}_{ii}\,\widehat{\sigma}_{jj}}}, \tag{4}$$

and $s_\lambda(x)$ denotes a generalized thresholding operator.[6] A key property is that $s_\lambda(x) = 0$ whenever $|x| \leq \lambda$, which shrinks small elements to zero and mitigates poor performance in the high-dimensional setting with $N \gg T$.

Bickel and Levina (2008a) showed that this thresholding estimator achieves a favorable rate of convergence when estimating covariance matrices in the following class:

$$\mathcal{U}_\tau(q, c_0, M) = \left\{\boldsymbol{\Sigma} : \sigma_{ii} \leq M,\ \sum_{j=1}^{N} |r_{ij}|^q \leq c_0(N),\ \text{for all } i\right\}. \tag{5}$$

[5] $\boldsymbol{\Sigma} = \boldsymbol{DRD}$, where $\boldsymbol{D} = \text{diag}\{\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{NN}}\}$.

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

Here, the sparsity pattern parameter $q \in [0, 1)$ and the sparsity magnitude parameter $c_0(N)$ jointly characterize the restrictions on the off-diagonal elements required for the covariance matrix to be sparse. The terminology for q and $c_0(N)$ can be understood by contrasting 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 whose off-diagonal elements can be all nonzero but individually small in magnitude. For $q = 0$, the sum $\sum_{j=1}^{N} |r_{ij}|^q$ simply counts the number of nonzero off-diagonal elements in each row and ignores their magnitudes. Hence, for the first type of covariance matrix, $c_0(N)$ can be small, but $c_0(N)$ would 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 can have a much smaller $c_0(N)$. Intuitively, q determines which pattern of sparsity is emphasized, whereas $c_0(N)$ measures how sparse the covariance matrix is under that pattern.

With these two examples in mind, it becomes clear that a limitation of $\mathcal{U}_\tau$ is that it attempts to capture two distinct sparsity patterns using a single pair of parameters $(q, c_0(N))$. Cai and Zhou (2012) showed that the minimax optimal convergence rate is faster for covariance matrices of the first type, particularly when the sparsity magnitude parameter $c_0(N)$ is smaller. By incorporating auxiliary information about the sparsity pattern—especially the locations of the few large elements—we can treat these two types of sparsity separately. This is a driving factor for our *Network-Guided Thresholding Estimator* to have superior performance both theoretically and empirically when the auxiliary network information is of reasonably high quality.

Given the *Location Indicator Matrix* defined in Equation 2 and its complement $\boldsymbol{L}^0$, we consider the following uniformity class:

$$\mathcal{U}_1(q, c_0, M) = \Bigg\{ \boldsymbol{\Sigma} = \boldsymbol{DRD} : \max_{1\le i\le N} \sigma_{ii} \le M, \max_{1\le i\le N} \sum_j L_{ij} \le c_0(N), \\ \max_{1\le i\le N} \sum_j L^0_{ij} |r_{ij}|^q \le c_0(N) \Bigg\}, \tag{6}$$

where $\boldsymbol{D} = \mathrm{diag}\{\sqrt{\sigma_{11}}, \cdots, \sqrt{\sigma_{NN}}\}$, $\boldsymbol{R}$ is the correlation coefficient matrix. Our Location Indicator Matrix $\boldsymbol{L}$ is designed to identify the "large" elements that should be retained (i.e., not shrunk) under the classical thresholding method, so the observation level λ in the population $\boldsymbol{L}$ is set to coincide with the threshold parameter in Equation 4. In practice, λ is treated as a tuning parameter and can be selected via cross-validation. By treating large elements (pairs (i, j) such that $L_{ij} = 1$) and small elements (pairs (i, j) such that $L^0_{ij} = 1$) differently, this uniformity class imposes separate restrictions on the number of large elements and on the growth rate of the remaining small elements.

The Location Indicator Matrix is well-suited for separating the large elements that should be retained and the small elements that should be regularized. However, $\boldsymbol{L}$ is unknown in practice. The estimated Location Indicator Matrix $\widehat{\boldsymbol{L}}$, constructed from auxiliary network information, is then used in practice to estimate the covariance matrix.

With $\widehat{\boldsymbol{L}}$, we propose the following *Network-Guided Thresholding Estimator* for $\boldsymbol{\Sigma} \in \mathcal{U}_1$,

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

where $T_{\widehat{\boldsymbol{L}},\lambda}\left(\widehat{\boldsymbol{R}}\right) = \left[s_{\widehat{\boldsymbol{L}},\lambda}\left(\widehat{r}_{ij}\right)\right]_{N\times N}$ and $s_{\widehat{\boldsymbol{L}},\lambda}\left(\widehat{r}_{ij}\right) = \widehat{r}_{ij}\widehat{L}_{ij} + s_{\lambda}\left(\widehat{r}_{ij}\right)\widehat{L}_{ij}^{0}$.

3.2 Network-Guided Banding

When the auxiliary data reveal the relative importance of neighbors for each node—for example, through frequency counts of co-mentions of firms i and j—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. In that case, Bickel and Levina (2008b) considered the following uniformity class of covariance matrices:

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

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

The original banding and tapering methods are primarily applicable to time series, which have a natural ordering.[7] In many practical applications, however, entities are not ordered, rendering these methods difficult to apply directly. Using the Relative Importance Matrix $\boldsymbol{C}$, we can extend these methods to more general network structures. Recall that $\boldsymbol{C} = [c_{ij}]_{N\times N}$, where c_{ij} is the rank of $|r_{ij}|$ in the vector $\text{abs}(\boldsymbol{r}_i) := (|r_{i1}|, \ldots, |r_{iN}|)$, and $S_k^{\boldsymbol{c}_i}$ denotes the set of the k most strongly correlated neighbors of entity i.

With this, we generalize the uniformity class in Bickel and Levina (2008b) (Equation 8) by directly comparing the relative magnitudes (rather than a real "distance") of entries within

[7] For permuted matrices, there are also methods for estimating a banding structure; see, for example, Giraud et al. (2023).

each row of the matrix. Specifically, we consider the generalized uniformity class of covariance matrices

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

where $\boldsymbol{r}_i$ is the i-th row of correlation matrix $\boldsymbol{R}$, and $\text{abs}(\boldsymbol{r}_i) = (|r_{i1}|, \cdots, |r_{iN}|)$ collects the absolute values of the correlation coefficients. $S_k^{\boldsymbol{c}_i}$ is the set of indexes of the k-biggest elements; when $k = 1$, $S_k^{\boldsymbol{c}_i} = i$ since the self-correlation is always the largest. For $k > 1$, $S_k^{\boldsymbol{c}_i}$ includes i and its $k-1$ nearest neighbors. Under Equation 9, correlations between non-neighboring pairs need to be small. Compared with the original banding approach, this formulation is permutation-invariant and accommodates a more general connectivity (network) structure.

With the estimated Relative Importance Matrix $\widehat{\boldsymbol{C}}$, we define the *Network-Guided Banding Estimator* as follows:

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

To implement the Network-Guided Banding estimator, we also need to choose k, the number of neighbors per node. In practice, the optimal k can be selected via cross-validation.

Remark: The k-neighbor relationship inferred from auxiliary network information can be asymmetric; that is, $i \in S_k^{\boldsymbol{c}_j}$ does not necessarily imply $j \in S_k^{\boldsymbol{c}_i}$. For example, the stock of a large firm with extensive coverage may be among the k nearest neighbors of many other stocks, while the reverse need not hold, leading to asymmetric news co-mention relationships. To ensure that the estimated covariance matrix is symmetric, we retain only mutual k-neighbor pairs.

3.3 Conditional Sparsity

Asset returns are exposed to common risk factors, leading to strong co-movement in their returns, so it is inappropriate to assume sparsity for the return covariance matrix itself. Following Fan et al. (2016b), we instead adopt the more widely used assumption of *Conditional Sparsity*, which states that after accounting for common factors, the residual dependence structure

among asset returns is sparse:

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

for $t = 1, 2, \cdots, T$, where $\boldsymbol{y}_t$ is the $N \times 1$ asset returns 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 component, which may exhibit cross-sectional dependence. Factor models have long been employed in modeling asset returns; see, for example, Ross (1976), Chamberlain and Rothschild (1982), and Fama and French (1993).

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

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

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. Since the factors are observable, the coefficients in Equation 11 can be estimated by ordinary least squares (OLS). After obtaining $\widehat{\boldsymbol{B}}$, the covariance matrix of the common-factor component is estimated by $\widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^\intercal$ 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 main challenge in estimating the return covariance matrix lies in estimating $\boldsymbol{\Sigma}_u$. Using the OLS estimates, we compute 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 obtain estimates of the Location Indicator Matrix or the Relative Importance Matrix, and then apply the corresponding network-guided method to construct the feasible Network-Guided Thresholding Estimator $\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_{u,\widehat{\boldsymbol{L}}}$ or feasible Network-Guided Banding Estimator $\widehat{\boldsymbol{\Sigma}}^{\mathcal{B}}_{u,\widehat{\boldsymbol{C}}}$.

4 Main Results

4.1 Theoretical Results

In this subsection, we first introduce the assumptions and establish the corresponding theoretical properties of the Network-Guided Estimators $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{u,\widehat{\boldsymbol{L}}}$ and $\widehat{\mathbf{\Sigma}}^{\mathcal{B}}_{u,\widehat{\boldsymbol{C}}}$. Building on these results, we then present the convergence theorem for $\widehat{\mathbf{\Sigma}}_y$ later. In the asymptotic analysis, we allow both N and T to approach infinity, and N can be larger than T with $\frac{\log N}{T} \to 0$. Proofs of all theorems are deferred to the Appendix B. 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 $\mathbf{\Sigma}_u$. 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 (thus uniformly mixing over N). Additionally, there are constants $\underline{c}, \bar{c}$, s.t., $0 < \underline{c} < \inf_{i,j} \text{Var}(u_{it}u_{jt}) < \sup_{i,j} \text{Var}(u_{it}u_{jt}) < \bar{c}$, $\underline{c} < \rho_{\min}\left(\mathbf{\Sigma}_u\right) < \rho_{\max}\left(\mathbf{\Sigma}_u\right) < \bar{c}$.*

(b) The tail of the distribution of u_{it} is uniformly bounded by an exponential-type tail, i.e., for some constant $\phi_3, \phi_4 > 0$, 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, assume that $P\left(\max_i \frac{1}{T}\sum_{t=1}^{T} |u_{it} - \widehat{u}_{it}|^2 > A a_T^2\right) \leq O\left(\kappa_1(N,T)\right)$ and $P\left(\max_{i,t} |u_{it} - \widehat{u}_{it}| > A\right) = o(1)$.

(d) Suppose $\gamma^{-1} = \phi_2^{-1} + 3\phi_4^{-1} > 1$ and $\left(\frac{\log N}{T}\right)^{\frac{1}{2(1-q)}} = o\left(T^{-1+\frac{3}{6-\gamma}}\right)$. For the banding estimator where q is not defined, we assume $(\log N)^{6/\gamma-1} = o(T)$, which actually corresponds to the $q = 0$ case.

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 extension to Fan et al. (2011) to allow $q > 0$. Given these assumptions, one can easily show that

$$P\left(\max_{i,j} |\widehat{\sigma}_{ij} - \sigma_{ij}| > A\sqrt{\frac{\log N}{T}}\right) = O\left(\frac{1}{N^2} + \kappa_1\left(N,T\right)\right)$$

for some constant A. The proof can be found in Lemma A.3 of Fan et al. (2011).

Remark: It is assumed that the error term $\boldsymbol{u}_t$ is stationary, and $\mathbf{\Sigma}_u$ is a static matrix over

t. Sometimes, in financial studies, a dynamic covariance matrix or network is needed (see Engle et al. (2019a) and Chen et al. (2025) for time-varying covariance matrix and network, respectively). For that, our framework can be further extended into the locally stationary case (see, for example, Inoue et al. (2017)). Specifically, we can allow the covariance matrix $\mathbf{\Sigma}_u(t/T_{\text{all}})$ to be slowly time-varying, where T_{all} is the whole time horizon. Moreover, given the high-dimensional techniques we develop, the models are in practice estimated using rolling windows with sample length $T \ll T_{\text{all}}$, which naturally accommodates the slow time variation.

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

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

(b) $\rho_{\min}(\mathbf{\Sigma}_f) > 0$ uniformly. In addition, there exists a positive definite matrix $\mathbf{\Omega}$, s.t., $\left\|\frac{1}{N}\boldsymbol{B}^\intercal\boldsymbol{B} - \mathbf{\Omega}\right\| = o(1)$.
(c) The number of factors $K = o(N)$, $K^4(\log N)^2 = o(T)$, 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 are pervasive. These conditions on 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 additional assumptions required to establish the asymptotic properties of each estimator and present the corresponding theoretical results.

4.1.1 Network-Guided Thresholding Estimator

We assume that $\mathbf{\Sigma}_u \in \mathcal{U}_1(q, c_0, M)$ as defined in Equation 6, which extends the class of sparse covariance matrices from Bickel and Levina (2008a). For notational simplicity, we will write $\mathbf{\Sigma}$, $\boldsymbol{R}$, and $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{\widehat{\boldsymbol{L}}}$ for $\mathbf{\Sigma}_u$, $\boldsymbol{R}_u$, and $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{u,\widehat{\boldsymbol{L}}}$, respectively. Suppose we have an auxiliary dataset $\mathcal{I}(N,T)$, collected separately from the return sample $\mathcal{X}(N,T)$, and the corresponding estimated $\widehat{\boldsymbol{L}}$. To derive the asymptotic results for our Network-Guided Thresholding estimator $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{\widehat{\boldsymbol{L}}} = \widehat{\boldsymbol{D}}T_{\widehat{\boldsymbol{L}},\lambda}\left(\widehat{\boldsymbol{R}}\right)\widehat{\boldsymbol{D}}$, in addition to Assumptions 1 and 2, we impose the following conditions.

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

(b) With ϱ_{0T}, ϱ_{1T}, and $\kappa_0(N,T)$ are defined in Equation 3, we assume: (i) the total observation error $\varrho_T := \lambda\, c_0(N)\, \varrho_{1T} + \sqrt{\frac{\log N}{T}}\, N\, \varrho_{0T} \to 0$; (ii) $\kappa_0(N,T) = O\left(\sqrt{\frac{\log N}{T}}\, \lambda^{-1}\right)$, which bounds the proportion of entries falling in the gray region.

Remark: Condition (a) is condition (iii) in Rothman et al. (2009), which is a standard assumption for thresholding estimation. The requirement in Condition (b) that $\varrho_T \to 0$ is not restrictive. Recall that the quantity ϱ_T measures the total observation error in $\widehat{\boldsymbol{L}}$, which has two components, the false negative probability $P\left(\widehat{L}_{ij}=0 \middle| L_{ij}=1\right) = O(\varrho_{1T})$ and the false positive probability $P\left(\widehat{L}_{ij}=1 \middle| L_{ij}=0\right) = O(\varrho_{0T})$. The first component (FN) is small in general because $\lambda c_0(N)$ is typically assumed to be $o(1)$ in the literature (Bickel and Levina (2008a)). For the second part (FP), when we construct $\widehat{\boldsymbol{L}}$ from the auxiliary network information $\mathcal{I}(N,T)$, we can impose conditions on the number of pairs with $\widehat{L}_{ij}=1$ and bound the total FP error $N\varrho_{0T}$. Consequently, it is reasonable to assume that $\varrho_T \to 0$ and the improvement from the auxiliary dataset is largely determined by whether it can correctly locate the large entries and control the FN error ϱ_{1T}.[8] Condition (b)(ii) restricts the mass of correlations in the "gray region". Recall that $\kappa_0(N,T)$ is proportional to the ratio between the number of entries in the gray region and $c_0(N)$. In particular, under the choice $\lambda \asymp \left(\frac{\log N}{T}\right)^{\frac{1-q}{2}}$ suggested in Theorem 1, this condition implies that $Q_N(\lambda-\nu_N, \lambda+\nu_N) = O\left(c_0(N)\left(\frac{\log N}{T}\right)^{\frac{q}{2}}\right)$, which requires that the number of elements in the shrinking interval $(\lambda-\nu_N, \lambda+\nu_N]$ cannot be too large.

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 $a_T = O(\lambda)$, then we have:*

$$\left\|\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_{\widehat{\boldsymbol{L}}} - \boldsymbol{\Sigma}\right\| = O_P\left(c_0(N)\left(\lambda^{1-q} + \sqrt{\frac{\log N}{T}}\right) + \sqrt{\varrho_T}\right), \tag{14}$$

where $\|\cdot\|$ represents the operator norm.

Remark: Similar to Bickel and Levina (2008a), the optimal choice of λ is $\lambda_T \asymp \left(\frac{\log N}{T}\right)^{1/2(1-q)}$, which yields

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

[8] Figure 7 shows that firms i and j exhibit substantially larger correlation coefficients when $\widehat{L}_{ij}=1$ (as identified by the auxiliary information) than when $\widehat{L}_{ij}=0$, indicating that the auxiliary network information we use in this paper is of high quality (i.e., FN error is low).

and, provided $c_0(N)\sqrt{\frac{\log N}{T}} \to 0$ and $\varrho_T \to 0$, we obtain $\left\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{\widehat{\boldsymbol{L}}} - \mathbf{\Sigma}\right\| = o_P(1)$. Compared to the standard thresholding estimator (see, e.g., Bickel and Levina (2008a) and Rothman et al. (2009)), which converges at rate $c_0(N)\left(\frac{\log N}{T}\right)^{\frac{1-q}{2}}$, our estimator attains a faster rate when the auxiliary information is of sufficiently high quality (i.e., when ϱ_T goes to 0 fast enough). In the previous Remark, we have argued that $\varrho_T \to 0$ is reasonable. Also notice that, since the traditional estimators treat all large elements as if they are small elements ($\varrho_{1T} = 1$), our proposed estimator will not have worse convergence rate than the traditional estimators once the FP error ϱ_{0T} is controlled.

4.1.2 Network-Guided Banding Estimator

We assume that $\mathbf{\Sigma}_u \in \mathcal{U}_2(\varepsilon, \alpha, b_0, M)$ as defined in Equation 9, which extends the class of bandable covariance matrix in Bickel and Levina (2008b). For notational simplicity, we write $\mathbf{\Sigma}$, $\boldsymbol{R}$, and $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{\widehat{\boldsymbol{C}}}$ to denote $\mathbf{\Sigma}_u$, $\boldsymbol{R}_u$, and $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{u,\widehat{\boldsymbol{C}}}$, respectively. As before, we must impose additional assumptions on the auxiliary network information to derive the asymptotic properties of our Network-Guided Banding Estimator $\widehat{\mathbf{\Sigma}}^{\mathcal{B}}_{\widehat{\boldsymbol{C}}} = \widehat{\boldsymbol{D}} B_{\widehat{\boldsymbol{C}},k}\left(\widehat{\boldsymbol{R}}\right)\widehat{\boldsymbol{D}}$.

Assumption 4. *(a) For $\boldsymbol{R}$ and $\boldsymbol{C}$, there exists b_1, s.t. $\max_{1\le i\le N}\sum_{j=1}^{N}|r_{ij}|\, I_{\{i\notin S_k^{c_j}, j\in S_k^{c_i}\}} < b_1(N)$, when $k = k_T \to \infty$;*[9]

(b) The total observation error $\varrho_T := k\varrho_{1T}\min\{1, k^{-\alpha}b_0(N)\} \to 0$.

Remark: Because the k-neighbor relation inferred from auxiliary networks can be asymmetric, we retain only mutual k-neighbor pairs to ensure a symmetric covariance estimator. Condition (a) restricts the degree of asymmetry, requiring that most asymmetric terms be sufficiently small. Condition (b) requires that the total error induced by the misclassification of large elements vanishes asymptotically. When $k^{-\alpha}b_0(N) > 1$, we have $\varrho_T = k\,\varrho_{1T}$, indicating that the total error scales with the number of selected neighbors multiplied by the per-neighbor observation error. In much of the existing literature, $k^{-\alpha}b_0(N)$ is typically assumed to be $o(1)$ (see, e.g., Bickel and Levina (2008a)), in which case this requirement is easy to satisfy. Unlike in the Network-Guided Thresholding case, no condition on ϱ_{0T} is needed here: when most true neighbors occupy the top ranks, non-neighbors can only appear in lower ranks. In other words, if ϱ_{1T} is small, then ϱ_{0T} cannot be too large.

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

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

Then,

$$\left\|\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{C}}}^{\mathcal{B}}-\boldsymbol{\Sigma}\right\|=O_P\left(b_0\left(N\right)k^{-\alpha}+k\sqrt{\frac{\log N}{T}}+b_1\left(N\right)+\sqrt{\varrho_T}\right), \tag{16}$$

where $\|\cdot\|$ *represents the operator norm.*

Remark: In the error bound, the first two terms, $k\sqrt{\frac{\log N}{T}}+b_0\left(N\right)k^{-\alpha}$ are the same as Bickel and Levina (2008a), while the term $b_1\left(N\right)$ arises from the "asymmetry" introduced in Assumption 4. Additionally, the term ϱ_T captures the additional error from using the estimated Relative Importance Matrix $\widehat{\boldsymbol{C}}$. Bickel and Levina (2008a) suggested an optimal choice of $k\asymp\left(\frac{\log N}{T}\right)^{-1/2(\alpha+1)}$, which yields

$$\left\|\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{C}}}^{\mathcal{B}}-\boldsymbol{\Sigma}\right\|=O_P\left(\left(1+b_0\left(N\right)\right)\left(\frac{\log N}{T}\right)^{\frac{\alpha}{2(\alpha+1)}}+b_1\left(N\right)+\sqrt{\varrho_T}\right). \tag{17}$$

If the matrix $\boldsymbol{C}$ implies a symmetric neighbor network, or if $b_1(N)$ converges to 0 faster than the first term, then our bound in Equation 17 simplifies to $O_P\left(\left(1+b_0\left(N\right)\right)\left(\frac{\log N}{T}\right)^{\frac{\alpha}{2(\alpha+1)}}+\sqrt{\varrho_T}\right)$. If the observation error term ϱ_T is negligible compared to the first term, i.e., the ordering within each column is well specified, this bound matches the result in Bickel and Levina (2008a), which is infeasible in our settings as it requires the order to be known.

4.1.3 Convergence of $\widehat{\boldsymbol{\Sigma}}_y$

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

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

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 $\widehat{\boldsymbol{\Sigma}}_y$. To this end, they consider the entropy loss norm,[10] defined as

$$\left\|\widehat{\boldsymbol{\Sigma}}_y-\boldsymbol{\Sigma}_y\right\|_E=\left(\frac{1}{N}\mathrm{tr}\left\{\left(\widehat{\boldsymbol{\Sigma}}_y\boldsymbol{\Sigma}_y^{-1}-\boldsymbol{I}_{N\times N}\right)^2\right\}\right)^{1/2}, \tag{18}$$

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

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

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

$\sqrt{\varrho_T}$, *we have*

$$\left\|\widehat{\boldsymbol{\Sigma}}_y - \boldsymbol{\Sigma}_y\right\|_E = O_P\left(K\frac{\sqrt{N}\log N}{T} + \sqrt{K}\sqrt{\frac{\log N}{T}} + \frac{c_0(N)\sqrt{\frac{\log N}{T}} + \sqrt{\varrho_T}}{\sqrt{N}}\right).$$

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

$$\left\|\widehat{\boldsymbol{\Sigma}}_y - \boldsymbol{\Sigma}_y\right\|_E = O_P\left(K\frac{\sqrt{N}\log N}{T} + \sqrt{K}\sqrt{\frac{\log N}{T}} + \frac{(1+b_0(N))\left(\frac{\log N}{T}\right)^{\frac{\alpha}{2(\alpha+1)}} + b_1(N) + \sqrt{\varrho_T}}{\sqrt{N}}\right).$$

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

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

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

$$\widehat{\boldsymbol{\Sigma}}_{M_0} = \widehat{\boldsymbol{\Sigma}} + (m_T - \widehat{\rho}_N)\cdot \boldsymbol{I}_{N\times N}, \tag{19}$$

where $\boldsymbol{I}_{N\times N}$ is the $N \times N$ identity matrix and $m_T > 0$ is a tuning parameter. This adjustment makes the smallest eigenvalue positive and thus ensures that $\widehat{\boldsymbol{\Sigma}}_{M_0}$ is invertible. Chen et al. (2019) further refined Equation 19 by

$$\widehat{\boldsymbol{\Sigma}}_M = \widehat{\boldsymbol{\Sigma}}\cdot I_{\{\widehat{\rho}_N > 0\}} + \widehat{\boldsymbol{\Sigma}}_{M_0}\cdot I_{\{\widehat{\rho}_N \leq 0\}} = \widehat{\boldsymbol{\Sigma}} + (m_T - \widehat{\rho}_N)\cdot \boldsymbol{I}_{N\times N}\cdot I_{\{\widehat{\rho}_N \leq 0\}}, \tag{20}$$

which retains $\widehat{\boldsymbol{\Sigma}}$ when it is already positive definite and applies the modified version $\widehat{\boldsymbol{\Sigma}}_{M_0}$ only when non-positive eigenvalues arise.

To invert the return covariance matrix, we apply the Sherman-Morrison-Woodbury formula to $\widehat{\boldsymbol{\Sigma}}_y$ and obtain

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

where $\widehat{\boldsymbol{\Sigma}}_f$ is naturally invertible in a (finite) factor structure while $\widehat{\boldsymbol{\Sigma}}_u$ may not be invertible in

a given finite sample. We therefore modify $\widehat{\mathbf{\Sigma}}_u$ using Equation 20. Note that

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

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

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

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

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

5 Simulation

5.1 True Covariance Matrix

To examine the properties of the proposed estimators in a finite sample, we run extensive numerical experiments. We consider four setups of sparse covariance matrices in our simulation studies. Setups 1 and 2 have configurations similar to those in Cai and Liu (2011), corresponding to two typical sparse cases—with and without a bandable structure, respectively. To make the simulation design more empirically relevant, we consider two additional covariance matrices calibrated from real data: we estimate the residual covariance matrix of de-factored returns, and induce sparsity by removing either non-neighbor connections (Setup 3) or small-magnitude entries (Setup 4).

- **Setup 1 (banded matrix with ordering)**: $\mathbf{\Sigma} = \text{diag}\{\boldsymbol{A}_1, \boldsymbol{A}_2\}$, where $\boldsymbol{A}_1 = \left(a_{ij}\right)_{\frac{N}{2}\times\frac{N}{2}}$ with $a_{ij} = \left(1 - \frac{|i-j|}{10}\right)^+$, and $\boldsymbol{A}_2 = 4\boldsymbol{I}_{\frac{N}{2}\times\frac{N}{2}}$. Here, $\boldsymbol{A}_1$ is a bandable sparse covariance matrix, while $\boldsymbol{A}_2$ is an identity matrix scaled by 4.

- **Setup 2 (sparse matrix without ordering)**: $\mathbf{\Sigma} = \text{diag}\{\boldsymbol{A}_1, \boldsymbol{A}_2\}$, where $\boldsymbol{A}_2 = 4\boldsymbol{J}_{\frac{N}{2}\times\frac{N}{2}}$, and $\boldsymbol{A}_1 = \boldsymbol{B} + \epsilon\boldsymbol{I}_{\frac{N}{2}\times\frac{N}{2}}$ with $\boldsymbol{B} = \left(b_{ij}\right)_{\frac{N}{2}\times\frac{N}{2}}$. The entries of $\boldsymbol{B}$ are generated independently

as follows:

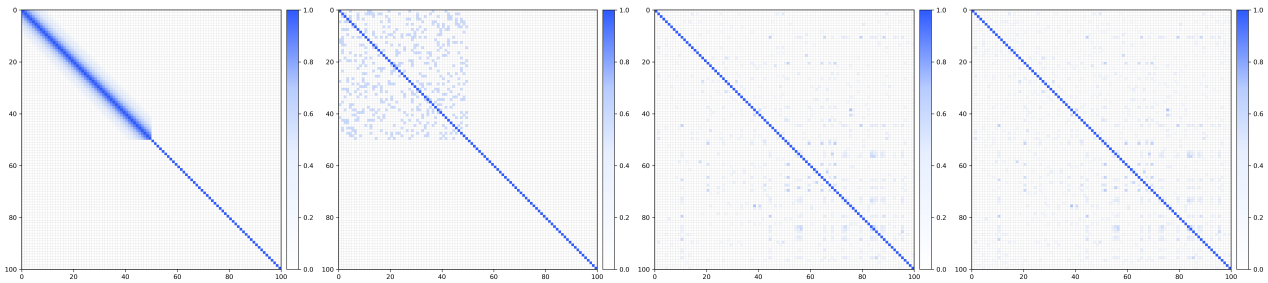
$$b_{ij} = \begin{cases} \text{Ber}\left(\frac{20}{N}\right), & \text{for } i < j, \\ 1, & \text{for } i = j, \\ b_{ji}, & \text{for } i > j. \end{cases} \tag{21}$$

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

- **Setup 3 (real-data covariance matrix with k-nearest-neighbor sparsification)**: We construct $\boldsymbol{\Sigma}$ by sparsifying the residual covariance matrix from the CH-4 factor model Liu et al. (2019). Specifically, in each row we retain only the k largest (in absolute value) off-diagonal entries and set the rest to zero, where $k = \lfloor 0.2N \rfloor$ and $\lfloor \cdot \rfloor$ denotes the floor function. The calibration uses daily returns of the first N constituent stocks of the CSI 800 index over 2012–2021, with $N \in \{100, 300, 500\}$.

- **Setup 4 (real-data covariance matrix with hard thresholding)**: Similarly, we obtain $\boldsymbol{\Sigma}$ by applying hard thresholding to the residual covariance matrix. The threshold is set to $\lambda = \sqrt{\frac{\log N}{T}}$, as suggested by Bickel et al. (2008).

The heatmaps of the four data-generating processes are presented in Figure 3. The two real-data cases (Setups 3 and 4) appear sparser than Setups 1 and 2. Setups 3 and 4 also look similar, as both are obtained by truncating the same residual covariance matrix of idiosyncratic returns, albeit using different rules.



(a) Setup 1 **(b)** Setup 2 **(c)** Setup 3 **(d)** Setup 4

Figure 3: Typical Heatmaps of Four Data Generating Processes

5.2 Auxiliary Network Information

To apply both Network-Guided methods, we need to generate *estimated* versions of the Location Indicator Matrix $\boldsymbol{L}$ and the Relative Importance Matrix $\boldsymbol{C}$. In empirical applications, the

auxiliary network information is observed with error and does not directly reveal the true correlation matrix $\boldsymbol{R}$. To reflect this in our simulation design, we generate an intermediate variable $\boldsymbol{R}^A = [r_{ij}^A]_{N\times N}$ by perturbing $\boldsymbol{R}$, which mimics the auxiliary network data that we have in the empirical study. Specifically, for $i \neq j$, we assume that,

$$r_{ij}^A \sim \text{Uniform}\big(r_{ij} - \eta,\ r_{ij} + \eta\big).$$

Given $\boldsymbol{R}^A$, we then construct $\widehat{\boldsymbol{L}}$ and $\widehat{\boldsymbol{C}}$ as follows:

1. **Estimated Location Indicator Matrix $\widehat{\boldsymbol{L}}$**: Given $\boldsymbol{R}^A$, we define $\widehat{L}_{ij} = 1$ if $|r_{ij}^A| > \lambda$ and $\widehat{L}_{ij} = 0$ otherwise, where the threshold is set to $\lambda = \sqrt{\frac{\log N}{T}}$, matching the choice used in Setup 4.

2. **Estimated Relative Importance Matrix $\widehat{\boldsymbol{C}}$**: In the banding cases, we rank each row of $\boldsymbol{R}^A$ and obtain the corresponding $\widehat{\boldsymbol{C}}$.

The parameter η controls the quality (noise level) of auxiliary network information. When $\eta = 0$, we have $\boldsymbol{R}^A = \boldsymbol{R}$, corresponding to perfect auxiliary information; larger values of η represent lower-quality auxiliary information. Since $|\widehat{r}_{ij} - r_{ij}| = O_P\left(\sqrt{\frac{\log N}{T}}\right)$, the quality measure of the auxiliary network information is the ratio $\eta/\sqrt{\frac{\log N}{T}}$. Table 1 shows the four cases of auxiliary information qualities we consider, from very low to perfect, according to the quality measure ratio, as well as the values of η, the false positive and false negative probabilities in these cases under Setup 1. Notably, even under the *High* quality setting, we still allow non-negligible false positive and false negative probabilities of 0.269 and 0.205 in Setup 1.[11] As shown below, in this case, the performance of our proposed estimators is still stronger than benchmark methods, which attests to the robustness of our methods.

5.3 Calibration with Empirical Data

To better match the real asset return data in our empirical analysis, we generate asset returns using the CH-4 factor model of Liu et al. (2019), which consists of four factors: Market, Value-

[11] For each candidate value η, we compute the false positive and false negative probabilities under Setup 1 for illustration, as the rates are easy to compute in this setup. In Setup 1, the off-diagonal elements of r_{ij} can only take values of $\{0, 0.1, 0.2, \ldots, 0.9\}$. For example, in the high-quality case for $(N, T) = (500, 300)$, if $\lambda = 0.144$ and $\eta = 0.095$, then given $r_{ij}^A \sim \text{Uniform}\,(r_{ij} - 0.095, r_{ij} + 0.095)$, the upper bound for false positive rate is $P\Big(\widehat{L}_{ij} = 1 \mid r_{ij} = 0.1\Big) = P\left(r_{ij}^A > \lambda \mid r_{ij} = 0.1\right) = 0.269$. Similarly, we obtain the upper bound for the false negative rate by $P\Big(\widehat{L}_{ij} = 0 \mid r_{ij} = 0.2\Big) = 0.205$.

Table 1: Auxiliary Network Information Quality. We report the auxiliary-network quality parameter η under different regimes. The false positive and false negative probabilities, denoted by ϱ_{0T} and ϱ_{1T}, are computed under $(N, T) = (500, 300)$ in Setup 1.

Case	**Auxiliary Information Quality**			
	Very Low	*Low*	*High*	*Perfect*
Ratio $\eta/\sqrt{(\log N)/T}$	2	1.5	1	0
η when $(N, T) = (100, 300)$	0.163	0.122	0.082	0
η when $(N, T) = (300, 300)$	0.182	0.136	0.091	0
η when $(N, T) = (500, 300)$	0.190	0.142	0.095	0
False discoveries $(\varrho_{0T}, \varrho_{1T})$	(0.384, 0.352)	(0.345, 0.303)	(0.269, 0.205)	(0, 0)

Minus-Growth, Small-Minus-Big, and Pessimistic-Minus-Optimistic.[12]

$$\boldsymbol{y}_t = \boldsymbol{\beta}_1 f_{\text{MKT,t}} + \boldsymbol{\beta}_2 f_{\text{VMG,t}} + \boldsymbol{\beta}_3 f_{\text{SMB,t}} + \boldsymbol{\beta}_4 f_{\text{PMO,t}} + \boldsymbol{u}_t. \tag{22}$$

In the simulations, we randomly draw factor loadings, factor returns and the residual $\boldsymbol{u}_t$ from distributions calibrated to empirical data. The cross-sectional distribution of factor loadings is calibrated using weekly returns for all listed stocks and CH-4 factor data from 2000 to 2021, resulting in

$$\beta_{1i} \sim N\left(0.7013, 0.1961^2\right), \quad \beta_{2i} \sim N\left(-0.1582, 0.2055^2\right),$$
$$\beta_{3i} \sim N\left(-0.1200, 0.2182^2\right), \quad \beta_{4i} \sim N\left(-0.0050, 0.2245^2\right).$$

We sample factor returns from multivariate normal distributions with means and covariance matrices calibrated to historical data from 2000 to 2021, as reported in Table 2. The residuals are drawn from a multivariate normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}_u$, which belongs to one of the four Setups we consider, scaled to match the aggregated variance of weekly returns in the real data.

Using the simulated data, we estimate the CH-4 factor model, collect the residuals $\widehat{\boldsymbol{u}}_t$, and then apply various methods to estimate their covariance matrix. For each scenario and each combination of sample size and auxiliary quality measure (N, T, η), the experiment is repeated 1000 times. Our analysis focuses on the finite-sample performance of the Network-Guided Thresholding Estimator and the Network-Guided Banding Estimator, compared with an extensive set of existing benchmark methods that do not exploit the auxiliary information:

1. **Sample Covariance Estimator**: Take the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_u$ computed from

[12]The CH-4 factor model has been shown to fit the Chinese stock market well and to outperform the Fama–French five-factor model.

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

	Descriptive Statistics					Correlation			
	Count	Mean	Std. Dev.	Skewness	Kurtosis	MKT	VMG	SMB	PMO
MKT	1119	0.147%	3.380%	-0.102	2.518	1.000	-0.237	0.159	-0.283
VMG	1119	0.277%	1.735%	1.048	6.909		1.000	-0.637	0.215
SMB	1119	0.119%	2.031%	-0.520	5.100			1.000	-0.137
PMO	1119	0.189%	1.588%	0.599	8.181				1.000

factor-model residuals.

2. **Hard Thresholding Estimator**: Apply $s_\lambda(x) = xI_{\{|x|>\lambda\}}$ to the elements in the sample covariance matrix $\widehat{\mathbf{\Sigma}}_u$.

3. **Soft Thresholding Estimator**: Apply $s_\lambda(x) = \text{sign}(x) \cdot \max\{|x| - \lambda, 0\}$ to the sample covariance matrix $\widehat{\mathbf{\Sigma}}_u$.

4. **Linear Shrinkage Estimator**: Apply linear shrinkage of Ledoit and Wolf (2004) on $\widehat{\mathbf{\Sigma}}_u$.

5. **Nonlinear Shrinkage Estimator**: Apply nonlinear shrinkage of Ledoit and Wolf (2012) on $\widehat{\mathbf{\Sigma}}_u$.[13]

6. **Self-Banding Estimator**: Construct $\widehat{\boldsymbol{C}}$ based on $\widehat{\boldsymbol{R}}$ itself and then apply our Network-Guided Banding.[14]

7. **POET**: Soft-thresholding estimator on the sample covariance matrix of residuals after taking out PCA factors rather than observed factors (Fan et al. (2013)).

The tuning parameters λ and k in the benchmark methods and the two Network-Guided methods are selected via the cross-validation procedure of Cai and Liu (2011). Specifically, we randomly split the sample $\{\boldsymbol{X}_t : 1 \leq t \leq T\}$ into two subsamples. Let $\widehat{\mathbf{\Sigma}}_1^j$ and $\widehat{\mathbf{\Sigma}}_2^j$ denote the corresponding sample covariance matrices from the j-th split, for $j = 1, \ldots, H$, where we take $H = 5$ in practice. For a candidate tuning parameter λ (or k), let $\widehat{\mathbf{\Sigma}}_1^j(\lambda)$ (or $\widehat{\mathbf{\Sigma}}_1^j(k)$) and $\widehat{\mathbf{\Sigma}}_2^j(\lambda)$ (or $\widehat{\mathbf{\Sigma}}_2^j(k)$) be the estimator computed from the first and second subsample in split j,

[13] We use Python packages `sklearn.covariance.LedoitWolf` and `nonlinshrink`.

[14] We do not require a separate Self-Thresholding method, since Hard Thresholding and Soft Thresholding are themselves thresholding approaches that utilize network information directly from $\widehat{\boldsymbol{R}}$.

respectively. We then choose λ and k by minimizing the cross-validation criteria

$$\frac{1}{H}\sum_{j=1}^{H}\left\|\widehat{\boldsymbol{\Sigma}}_1^j(\lambda)-\widehat{\boldsymbol{\Sigma}}_2^j\right\|_F^2, \qquad \frac{1}{H}\sum_{j=1}^{H}\left\|\widehat{\boldsymbol{\Sigma}}_1^j(k)-\widehat{\boldsymbol{\Sigma}}_2^j\right\|_F^2,$$

respectively.

The results, evaluated using both the Frobenius norm and the operator norm, are presented in Table 4, which compares the performance of the various estimation methods.

Panel A of Table 4 reports the results for Setup 1, where the true covariance matrix is banded with a natural ordering. We find that both Network-Guided estimators outperform their counterparts whenever the auxiliary network information is of reasonable quality. For the Network-Guided Thresholding estimator, except in the case $N < T$ combined with $\eta = 2\sqrt{(\log N)/T}$ (corresponding to very poor auxiliary information), the method consistently outperforms the benchmarks across most (η, N) combinations. Turning to the Network-Guided Banding estimator, it achieves smaller estimation errors than all other purely statistical methods, provided that the auxiliary information is not excessively noisy. For instance, when $\eta = \sqrt{(\log N)/T}$, corresponding to moderate accuracy, the Network-Guided Banding estimator dominates across most (N, T) settings, particularly when $N \geq T$.

Panel B presents the results for Setup 2, in which the true covariance matrix is sparse but lacks an ordering structure. In this case, both Network-Guided estimators continue to exhibit superior performance as long as the auxiliary network information is of adequate quality. Note that the Self-Banding estimator performs particularly poorly in this no-ordering scenario, in contrast to Setup 1 where an ordering structure exists. By contrast, our Network-Guided Banding Estimator adapts to a broader class of bandable covariance structures and continues to perform well, provided that the observation error parameter η is not unduly large. Similarly, the Network-Guided Thresholding Estimator remains highly effective, especially when the observation error in $\widehat{\boldsymbol{L}}$ is small.

We also highlight several features of the benchmark methods. First, the Sample covariance estimator performs reasonably well when $N = 100$, but its performance deteriorates substantially as N increases. Second, the Self-Banding method behaves similarly to Hard Thresholding in Panel A, owing to the presence of an ordering structure. In contrast, in Panel B, where no such structure exists, Self-Banding tends to perform worse than Hard Thresholding. Finally, POET—a purely data-driven method that does not use observed factors—performs worse than all other models except the simple Sample covariance estimator. This is expected, since our data-generating process explicitly links individual returns to realized factor returns, making

methods that exploit observed factors naturally superior to POET.

Panels C and D report Setups 3 and 4, where the "true" idiosyncratic covariance is constructed from real data and then truncated to induce sparsity. Although both are truncated versions of the same underlying real-data covariance, they emphasize different structural features: Setup 3 keeps the largest neighbors in each row based on relative importance, while Setup 4 keeps large elements based on absolute importance. Consistent with these constructions, the two Network-Guided estimators exhibit complementary strengths in Panels C and D. In Panel C (Setup 3), the Network-Guided Banding estimator benefits most from accurate auxiliary information and becomes particularly competitive when the auxiliary network reliably captures neighborhood relations, delivering the smallest errors in several (N, T) settings. In Panel D (Setup 4), the Network-Guided Thresholding estimator tends to dominate, reflecting that the target sparsity is generated by entrywise truncation, which aligns more directly with thresholding-type procedures. Across both real-data settings, the performance of the Network-Guided methods improves monotonically as auxiliary quality increases, and both methods remain robust relative to the benchmarks in high-dimensional cases.

In summary, across all four setups, the proposed Network-Guided estimators deliver strong and stable numerical performance. They perform well not only in the artificial designs of Cai and Liu (2011) but also in the two cases calibrated from the data used in our empirical study, consistently outperforming traditional purely statistical approaches whenever the auxiliary network information is of decent quality. We do not attempt to compare the two Network-Guided estimators directly, as they are designed for different types of auxiliary network information. Instead, they exhibit complementary strengths: the Network-Guided Banding estimator excels when the underlying covariance matrix is bandable and the auxiliary network encodes a meaningful neighborhood structure for each entity (as in Panels A and C), whereas the Network-Guided Thresholding estimator is more robust when no clear bandable structure is present.

Table 3: Simulation Results for Setups 1 and 2. We report the results of our methods with different auxiliary qualities (denoted by $\eta/\sqrt{\log N/T}$) and some benchmarks, including Sample Covariance matrix (Sam.), Linear Shrinkage (L-Sh.), Non-linear Shrinkage (N-Sh.), Hard Thresholding (H-Th.), Soft Thresholding (S-Th.), Self-Banding (S-Ba.) and POET, in terms of Frobenius norm and the operator norm. Note that the Python package `nonlinshrink` only works when $T > N$. We perform 1000 simulations and report the mean and standard deviation. The tuning parameters are selected via the cross-validation procedure of Cai and Liu (2011). Results are displayed for different values of $N = 100, 300, 500$, with T fixed at 300.

Setting		Network-Guided Thresholding				Network-Guided Banding				Benchmarks						
		2	1.5	1	0	2	1.5	1	0	Sam.	L-Sh.	N-Sh.	H-Th.	S-Th.	S-Ba.	POET
• Panel A: Setup 1, banded matrix with ordering																
$N = 100$	$\|\cdot\|_F$	5.33	3.42	2.45	2.42	4.34	3.40	2.89	2.29	8.72	8.00	6.42	3.47	4.72	5.78	13.70
		(0.19)	(0.23)	(0.30)	(0.30)	(0.15)	(0.21)	(0.25)	(0.33)	(0.23)	(0.20)	(0.24)	(0.32)	(0.49)	(0.20)	(0.81)
	$\|\cdot\|$	1.65	1.30	1.25	1.23	1.68	1.38	1.26	1.24	3.08	2.94	2.79	1.43	2.88	1.66	8.62
		(0.28)	(0.33)	(0.34)	(0.34)	(0.30)	(0.31)	(0.33)	(0.35)	(0.34)	(0.31)	(0.29)	(0.34)	(0.38)	(0.25)	(0.25)
$N = 300$	$\|\cdot\|_F$	17.12	10.83	4.49	4.40	9.03	6.58	4.84	4.07	25.95	20.68		8.59	8.50	11.73	14.08
		(0.20)	(0.17)	(0.28)	(0.28)	(0.12)	(0.18)	(0.26)	(0.32)	(0.27)	(0.19)		(0.33)	(0.47)	(0.21)	(0.93)
	$\|\cdot\|$	3.07	1.91	1.57	1.55	2.30	1.81	1.57	1.56	6.29	4.99		1.96	3.25	1.96	5.98
		(0.24)	(0.23)	(0.28)	(0.28)	(0.23)	(0.25)	(0.27)	(0.30)	(0.40)	(0.22)		(0.28)	(0.25)	(0.22)	(1.05)
$N = 500$	$\|\cdot\|_F$	29.61	19.47	5.94	5.82	12.13	8.90	6.48	5.28	43.18	30.67		13.58	11.08	16.16	16.01
		(0.22)	(0.17)	(0.28)	(0.29)	(0.12)	(0.17)	(0.26)	(0.33)	(0.31)	(0.18)		(0.33)	(0.46)	(0.22)	(0.50)
	$\|\cdot\|$	4.45	2.44	1.68	1.67	2.51	1.97	1.69	1.68	8.81	6.03		2.26	3.38	2.07	3.08
		(0.22)	(0.18)	(0.26)	(0.26)	(0.22)	(0.24)	(0.25)	(0.27)	(0.41)	(0.17)		(0.25)	(0.23)	(0.19)	(0.37)
• Panel B: Setup 2, sparse matrix without ordering																
$N = 100$	$\|\cdot\|_F$	18.94	14.13	11.50	8.02	20.97	19.79	14.83	7.95	27.26	17.67	15.84	14.96	17.89	20.03	22.25
		(0.32)	(0.31)	(0.31)	(0.36)	(0.10)	(0.11)	(0.18)	(0.36)	(0.42)	(0.25)	(0.31)	(0.61)	(0.38)	(0.41)	(0.50)
	$\|\cdot\|$	4.85	4.49	3.58	2.49	9.11	7.89	4.68	2.48	7.41	7.65	6.27	4.35	8.27	4.61	11.07
		(0.36)	(0.37)	(0.28)	(0.28)	(0.15)	(0.21)	(0.26)	(0.28)	(0.53)	(0.47)	(0.75)	(0.42)	(0.30)	(0.31)	(0.71)
$N = 300$	$\|\cdot\|_F$	59.02	40.36	23.56	15.35	37.68	36.42	31.82	14.95	85.21	35.91		35.17	32.93	45.35	38.10
		(0.39)	(0.32)	(0.30)	(0.38)	(0.08)	(0.09)	(0.13)	(0.38)	(0.57)	(0.13)		(0.57)	(0.31)	(0.46)	(0.54)
	$\|\cdot\|$	8.06	5.82	4.51	2.84	9.88	9.36	6.99	2.80	15.81	9.55		5.40	8.78	5.92	7.10
		(0.22)	(0.21)	(0.19)	(0.24)	(0.08)	(0.09)	(0.12)	(0.24)	(0.53)	(0.20)		(0.22)	(0.15)	(0.18)	(0.31)
$N = 500$	$\|\cdot\|_F$	102.65	70.37	32.38	20.99	48.80	47.69	42.27	20.96	146.04	47.87		54.38	43.50	66.39	56.07
		(0.50)	(0.38)	(0.30)	(0.38)	(0.08)	(0.09)	(0.12)	(0.36)	(0.70)	(0.09)		(0.63)	(0.28)	(0.49)	(0.62)
	$\|\cdot\|$	12.04	7.31	5.06	3.04	10.11	9.65	7.63	4.89	23.00	10.10		6.22	9.07	6.71	7.08
		(0.27)	(0.15)	(0.15)	(0.23)	(0.06)	(0.07)	(0.10)	(0.18)	(0.58)	(0.13)		(0.18)	(0.12)	(0.15)	(0.23)

Table 4: Simulation Results for Setups 3 and 4. We report the results of our methods with different auxiliary qualities (denoted by $\eta/\sqrt{\log N/T}$) and some benchmarks, including Sample Covariance matrix (Sam.), Linear Shrinkage (L-Sh.), Non-linear Shrinkage (N-Sh.), Hard Thresholding (H-Th.), Soft Thresholding (S-Th.), Self-Banding (S-Ba.) and POET, in terms of Frobenius norm and the operator norm. Note that the Python package `nonlinshrink` only works when $T > N$. We perform 1000 simulations and report the mean and standard deviation. The tuning parameters are selected via the cross-validation procedure of Cai and Liu (2011). Results are displayed for different values of $N = 100, 300, 500$, with T fixed at 300.

Setting		Network-Guided Thresholding				Network-Guided Banding				Benchmarks						
		2	1.5	1	0	2	1.5	1	0	Sam.	L-Sh.	N-Sh.	H-Th.	S-Th.	S-Ba.	POET
• Panel C: Setup 3, real data covariance matrix with k-nearest neighbor sparsification																
$N = 100$	$\|\cdot\|_F$	4.30	3.45	3.05	2.96	5.64	4.85	3.47	2.27	5.80	4.94	4.65	3.67	4.91	4.09	8.03
		(0.08)	(0.09)	(0.09)	(0.09)	(0.03)	(0.04)	(0.07)	(0.12)	(0.11)	(0.14)	(0.11)	(0.13)	(0.23)	(0.12)	(0.28)
	$\|\cdot\|$	1.19	1.01	0.96	0.92	2.27	1.72	1.03	0.83	1.90	2.22	1.89	1.07	2.57	1.09	5.73
		(0.17)	(0.17)	(0.16)	(0.16)	(0.21)	(0.22)	(0.16)	(0.20)	(0.21)	(0.26)	(0.22)	(0.18)	(0.34)	(0.16)	(0.34)
$N = 300$	$\|\cdot\|_F$	12.16	8.45	5.18	4.71	11.55	9.64	6.83	3.82	17.31	11.63		7.37	8.89	8.15	9.24
		(0.10)	(0.09)	(0.10)	(0.11)	(0.02)	(0.04)	(0.07)	(0.13)	(0.15)	(0.11)		(0.15)	(0.23)	(0.14)	(0.54)
	$\|\cdot\|$	1.97	1.35	1.08	1.04	3.16	2.30	1.37	1.00	3.90	4.01		1.29	2.88	1.29	2.45
		(0.17)	(0.18)	(0.21)	(0.22)	(0.10)	(0.11)	(0.13)	(0.24)	(0.28)	(0.23)		(0.20)	(0.30)	(0.19)	(0.81)
$N = 500$	$\|\cdot\|_F$	20.87	14.22	6.56	6.08	16.43	13.25	8.74	5.33	29.37	18.53		10.93	12.62	11.10	12.71
		(0.13)	(0.10)	(0.13)	(0.14)	(0.03)	(0.05)	(0.09)	(0.16)	(0.19)	(0.12)		(0.18)	(0.28)	(0.17)	(0.31)
	$\|\cdot\|$	2.98	1.80	1.30	1.29	3.53	2.59	1.45	1.28	5.82	5.54		1.61	3.45	1.54	2.19
		(0.21)	(0.21)	(0.27)	(0.28)	(0.17)	(0.09)	(0.21)	(0.28)	(0.32)	(0.23)		(0.24)	(0.33)	(0.24)	(0.32)
• Panel D: Setup 4, real data covariance matrix with hard thresholding																
$N = 100$	$\|\cdot\|_F$	3.99	2.94	2.37	1.88	5.13	3.97	2.88	2.14	5.80	4.88	4.51	2.98	4.36	4.07	7.69
		(0.09)	(0.10)	(0.11)	(0.14)	(0.04)	(0.06)	(0.09)	(0.12)	(0.12)	(0.15)	(0.12)	(0.16)	(0.26)	(0.11)	(0.28)
	$\|\cdot\|$	1.17	1.02	0.91	0.84	2.48	1.51	1.13	0.92	1.91	2.25	1.87	1.05	2.59	1.13	5.79
		(0.18)	(0.18)	(0.20)	(0.21)	(0.21)	(0.20)	(0.18)	(0.19)	(0.22)	(0.28)	(0.23)	(0.21)	(0.35)	(0.18)	(0.32)
$N = 300$	$\|\cdot\|_F$	12.62	8.22	3.65	3.26	10.38	8.37	5.26	3.44	18.60	11.32		6.49	7.52	8.88	8.27
		(0.10)	(0.09)	(0.12)	(0.14)	(0.03)	(0.04)	(0.08)	(0.13)	(0.14)	(0.10)		(0.17)	(0.26)	(0.11)	(0.54)
	$\|\cdot\|$	2.00	1.35	1.01	0.99	3.30	2.37	1.23	1.03	4.01	4.25		1.25	2.91	1.33	2.37
		(0.14)	(0.18)	(0.22)	(0.23)	(0.15)	(0.15)	(0.18)	(0.21)	(0.26)	(0.20)		(0.21)	(0.29)	(0.18)	(0.75)
$N = 500$	$\|\cdot\|_F$	28.40	19.35	8.91	6.62	20.70	18.53	14.82	11.15	40.05	20.78		14.76	15.37	17.86	17.22
		(0.14)	(0.12)	(0.16)	(0.19)	(0.02)	(0.03)	(0.05)	(0.09)	(0.22)	(0.13)		(0.24)	(0.41)	(0.15)	(0.83)
	$\|\cdot\|$	3.99	2.78	2.21	1.83	8.50	7.37	6.12	4.75	7.60	8.57		2.60	6.58	4.71	4.86
		(0.29)	(0.39)	(0.40)	(0.39)	(0.22)	(0.29)	(0.36)	(0.42)	(0.44)	(0.32)		(0.47)	(0.64)	(0.44)	(1.46)

6 Empirical Study

6.1 Data

6.1.1 Asset Returns

Stocks in our sample are constituent stocks of three main indices in China in 2021, namely HS300 (000300.SH), CSI500 (000905.SH), and CSI800 (000906.SH), comprising 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.[15]

6.1.2 News Co-mention Linkage Data

We analyze millions of articles from RESSET's Financial Text Intelligent Analysis Platform and the Juyuan database, spanning 2006–2021. From these, we select articles that mention at least one publicly listed company in China's A-share market, yielding a total of 1,138,247 news items.

Ge et al. (2025) documented that news co-mentions capture a wide range of economically important linkages. Following their approach, we identify news-implied links based on co-mentions within the same news article, using four methods tailored to different co-mention definitions: the `one2one_passage`, `all_passage`, `one2one_sentence`, and `all_sentence` approaches. In Table 5, we summarize the differences between these link identification strategies.

Table 5: 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`

At time t, we use the latest τ_0 days as the identification window.[16] For each of the four link identification strategies, we count the number of co-mentions M_{ij} for each stock pair (i, j), and then we construct the co-mention matrix $\boldsymbol{M} = (M_{ij})_{N\times N}$.

[15] 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 the 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.

[16] 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.

6.1.3 Analyst Coverage Linkage

We also explore stock linkages based on analyst coverage, denoted as `Analyst`. This approach is supported by a large body of finance literature showing 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 use data from the Chinese Research Data Services Platform (CNRDS), covering analyst reports from January 2005 to December 2020. After data cleaning, we identify 530,696 unique analyst reports to trace connections based on shared coverage. Starting in 2006, at each 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 construct the analyst co-coverage linkage matrix $\boldsymbol{M} = (M_{ij})_{N\times N}$.

6.1.4 Industry-based Linkage

Stocks within the same sector or industry often comove beyond their exposure to common risk factors (Fan et al. (2016a)). Motivated by this, we examine linkages based on industry classifications, denoted as `Industry`. We analyze three major industry classification systems in China—CSRC, CITIC, and Shenwan—updated annually using the RESSET database. Our primary focus is the Shenwan primary classification, which is widely regarded as the leading system in China's financial industry. For each pair of stocks (i, j), $M_{ij} = 1$ if the two stocks are in the same industry, and 0 otherwise.

6.1.5 Summary Statistics of All Types of Auxiliary Network

We report the summary statistics of these different networks in Table 6. Under `sentence_1`, each focal firm has 16 peer firms on average, fewer than 29 peers from `article_1`. This is consistent with our expectation: the same-sentence strategy removes noisy links that may arise under the same-article strategy, leading to fewer identified connections. In addition, the number of peer firms naturally increases with the length of the identification window. For the other linkage types, we generally observe denser networks, with each sample stock having more linked stocks on average.

Table 6: 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

6.2 Methodology

We use the Global Minimum Variance (GMV) portfolio as a testing ground to evaluate different covariance matrix estimation techniques. Our main interest is whether the GMV portfolio constructed with the help of auxiliary network information outperforms portfolios based on alternative estimators. This subsection describes the procedure for applying our proposed Network-Guided estimators to the stock return covariance matrix, followed by an out-of-sample performance comparison.

6.2.1 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)[17]:

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 f_{\text{MKT,t}} + \boldsymbol{\beta}_2 f_{\text{VMG,t}} + \boldsymbol{\beta}_3 f_{\text{SMB,t}} + \boldsymbol{\beta}_4 f_{\text{PMO,t}} + \boldsymbol{u}_t \\ &= \boldsymbol{\beta}_0 + \boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{u}_t. \end{aligned} \tag{23}$$

The estimator of $\boldsymbol{\Sigma}_y = \text{Cov}(\boldsymbol{y}_t, \boldsymbol{y}_t)$ is given by

$$\widehat{\boldsymbol{\Sigma}}_y = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^\intercal + \widehat{\boldsymbol{\Sigma}}_u, \tag{24}$$

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

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

Depending on the nature of the auxiliary network information, we choose which Network-Guided method to apply. In general, if the auxiliary dataset provides unweighted linkage information—that is, it only indicates 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. By contrast, if the auxiliary dataset provides weighted linkage information, revealing the relative importance of neighbors for each node, then both the Network-Guided Thresholding and Network-Guided Banding methods can be used.

To apply the Network-Guided Thresholding method, we first estimate the Location Indicator Matrix $\boldsymbol{L}$. Since $\boldsymbol{L}$ is a 0–1 matrix, we tune a threshold parameter m and set $\widehat{L}_{ij} = 1$ only if stocks i and j are co-mentioned more than m times within a given identification window; that is, $\widehat{L}_{ij} = I_{\{M_{ij}>m\}}$.[18] The tuning parameter m is selected by in-sample cross-validation (targeting the minimization of the GMV portfolio's standard deviation, as detailed in the next subsection). To apply the Network-Guided Banding estimator, we need to construct an estimate of the Relative Importance Matrix $\boldsymbol{C}$ from a weighted auxiliary network (news co-mentions or analyst coverage). We use the news co-mention matrix $\boldsymbol{M}$ as an illustrative example, where each entry $0 \le M_{ij} < \infty$ is the integer count of co-mentions. For each row $\boldsymbol{M}_i$ of $\boldsymbol{M}$, we set

[17]The time series of the four factors can be obtained from the author's website.

[18]For industry link, $M_{ij} = 0/1$, we also tune $m \in \{0, 1\}$. Specifically, if $m = 0$, then pairs within the same industry have $\widehat{L}_{ij} = 1$; if $m = 1$, then all pairs have $\widehat{L}_{ij} = 0$, so we effectively disregard the auxiliary industry network information and revert to traditional thresholding.

$\widehat{c}_{ij}$ to be the rank of M_{ij} within that row, and we likewise determine the number of neighbors k for each asset via in-sample cross-validation.

The analyst co-coverage linkage network has the same properties as the news co-mention network. Therefore, all procedures are identical to those described above. By contrast, the industry-based linkage network is unweighted, with $M_{ij} = I_{\{i \text{ and } j \text{ are in the same industry}\}}$. Given the nature of the industry network, we can apply only the Network-Guided Thresholding method: the industry network does not provide relative-importance information, so it cannot be used for the Network-Guided Banding method.

6.2.3 Down-weighting Uninformative Auxiliary Network Information with Cross-Validations

Note that using cross-validation to select tuning parameters is a key step in our empirical analysis. When the auxiliary network information has low quality, we may prefer to disregard it; the cross-validation procedure provides exactly this safeguard. For example, when the auxiliary network is extremely poor, cross-validation may even choose $m = \max_{i,j} M_{ij}$, implying $\widehat{L}_{ij} \equiv 0$ for all pairs of (i, j). In this case, the auxiliary network is ignored and the Network-Guided Thresholding estimator collapses to the traditional thresholding method. Similarly, for the Network-Guided Banding approach, when $k = N$, the auxiliary network is effectively disregarded and the estimator reduces to the sample covariance matrix.

We compute the proportion of time points at which the Network-Guided estimation is active (i.e., the auxiliary network information is not disregarded) under different settings. We find that the news- and analyst-based networks are frequently and persistently retained—especially for CSI800—whereas the industry network is more often ignored, indicating that it adds little incremental information. Details can be found in Appendix A.1.

6.2.4 Comparing the Out-of-sample Portfolios

As discussed in Engle et al. (2019b) and Chen et al. (2019), constructing a global minimum variance (GMV) portfolio is an effective way to evaluate the performance of covariance matrix estimators. Unlike the optimal mean–variance (MV) portfolio, the GMV portfolio avoids the need to estimate expected 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 following Ledoit and Wolf (2004). The

theoretical weights for a GMV portfolio are given by

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

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

Under the factor structure, the estimator of the return covariance matrix can be decomposed as $\widehat{\boldsymbol{\Sigma}}_y = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^\intercal + \widehat{\boldsymbol{\Sigma}}_u$. The factor component can be readily estimated using the CH-4 factor model. Our goal is to show that the proposed method estimates $\boldsymbol{\Sigma}_u$ more accurately and thereby improves GMV portfolio performance.

For robustness checks, we also consider (i) the maximum-return portfolio for a given variance level σ_0^2 and (ii) the minimum-variance portfolio for a given expected-return level μ_0. Recall the construction of the classical optimal portfolio. For example, given a return constraint μ_0, we solve:

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

$\boldsymbol{\mu} = E(\boldsymbol{y}_t)$. The optimal weight vector is

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

where the matrix $\boldsymbol{\Psi}$ is defined as[19]

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

Given the factor structure of asset returns, we have $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{B}}\bar{\boldsymbol{f}}$ and $\widehat{\boldsymbol{\Sigma}}_y = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^\intercal + \widehat{\boldsymbol{\Sigma}}_u$. Plugging the optimal weights into the variance expression yields the minimum variance associated with μ_0 is

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

which defines the mean-variance efficient frontier $\{(\sigma_0, \mu_0), \mu_0 \geq 0\}$. Starting from the maximization problem for any given σ_0^2 leads to the same efficient frontier. Note, however, that this efficient frontier is an in-sample object. When we fix an in-sample σ_0 or μ_0, the corresponding out-of-sample portfolio can deliver different realized volatility and mean return, giving rise to

[19] Details and proofs can be found in Chapter 1.6 of Linton (2019).

an out-of-sample efficient frontier. As with the GMV portfolio, we select tuning parameters via in-sample training and then construct out-of-sample efficient frontiers under different models.

Importantly, although most estimated covariance matrices are positive definite, we modify any non-positive definite covariance estimates $\widehat{\mathbf{\Sigma}}_u$ using the method outlined in Equation 20.

Remark: The portfolio construction is one of the direct applications (but not only) of covariance matrix estimation. Here, we highlight and compare the method with some other candidate methodologies in portfolio construction without estimating the covariance matrix: (a) Many papers have proposed methods for directly estimating the precision matrix $\mathbf{\Sigma}^{-1}$ (see, for example, Cai et al. (2011)). However, in many practical problems, portfolio weights are often subject to additional constraints, such as non-negativity, in which case the covariance matrix itself is required. (b) Meanwhile, there are also approaches that estimate portfolio weights directly (typically via sparse regression) to construct the maximum-Sharpe-ratio portfolio (see, for example, Ao et al. (2019)). From a purely portfolio-construction perspective, such methods can be fast and robust because they bypass covariance matrix estimation altogether. However, they also discard information about cross-asset dependence. When the goal is to study asset co-movements or to attain cross-firm predictability, the cross-asset dependence is still of importance.

6.3 Empirical Results

6.3.1 Comparing GMV Portfolios

Table 7 reports the out-of-sample volatility (measured by standard deviation) of GMV portfolios constructed using different methods and stock universes, 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 $\widehat{\mathbf{\Sigma}}_u$ with a positive definite correction if necessary, and compute $\widehat{\mathbf{\Sigma}}_y = \widehat{\boldsymbol{B}}\widehat{\mathbf{\Sigma}}_f\widehat{\boldsymbol{B}}^\intercal + \widehat{\mathbf{\Sigma}}_u$.

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

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

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

5. **Self-Banding**: Construct $\widehat{\boldsymbol{C}}$ based on $\widehat{\boldsymbol{R}}$ itself and then apply our Network-Guided Banding.

6. **POET**: Implement the soft thresholding method using PCA factors rather than the observed factors.

As noted by Ge et al. (2025), the news-implied network data are abundant only after 2011. Accordingly, we conduct our empirical analysis over the period 2012–2021. Each month, using the preceding one-year window of daily stock and factor returns, together with the corresponding auxiliary dataset, we estimate the asset covariance matrix under the competing methods described above. We then evaluate performance by forming the GMV portfolio from each estimated covariance matrix and comparing the out-of-sample return volatility. All tuning parameters are chosen via in-sample cross-validation, with the objective of minimizing the portfolio standard deviation.

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 exhibits high estimation error due to the large number of parameters, leading to poor out-of-sample performance. Shrinkage methods have the potential to improve portfolio allocation. For example, for the CSI500 index, the "Linear Shrinkage" method offers a competitive reduction in standard deviation relative to "Factors Only".

By contrast, the results for POET and "Self-Banding" are comparatively weak. Unlike our Network-Guided methods, "Self-Banding" shrinks $\widehat{\boldsymbol{R}}$ solely by exploiting its own rank structure, without incorporating auxiliary information beyond $\boldsymbol{X}_t$. Moreover, in financial applications, there is typically no natural ordering of assets; in the absence of an external ordering indicator, "Self-Banding" entails greater misspecification risk than traditional thresholding methods. As for POET, which relies on latent PCA factors, its performance is weaker than the "Factors Only" approach, suggesting that stable observable factors may yield better out-of-sample performance than PCA-based latent factors, despite the latter's stronger in-sample explanatory power.

For Network-Guided Thresholding (Panel B), the results reveal a mixed performance pattern. When incorporating `analyst` and `industry` network information, Network-Guided Thresholding does not outperform the "Factors Only" method. By contrast, news-implied linkages generally appear more effective, reducing out-of-sample volatility in most cases.

Turning to the Network-Guided Banding results (Panel C), news-implied networks again

perform better than other sources of auxiliary network information in improving covariance matrix estimation. Note that `industry` provides only unweighted linkage information, making it unsuitable for Banding.

Finally, we recognize that different sources of auxiliary network information can vary in quality. Among the three information sets used in our empirical study, news co-mention links appear most effective at identifying economically linked pairs relative to the other sources. In the Network-Guided Thresholding setting, using `one2one_passage_12` outperforms all purely data-driven benchmarks across all indices considered. This finding is consistent with the evidence reported in Ge et al. (2025). At the same time, we also find that the best-performing strategy for extracting news-implied linkages varies across indices.

This suggests that while news-based auxiliary network information is valuable, its implementation should be tailored to market conditions and asset characteristics, reflecting the inherent complexity of financial markets.

Table 7: 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 $\mathbf{\Sigma}_u$. "Sample" refers to a simple sample estimator of $\mathbf{\Sigma}_u$; "Factors Only" refers to setting $\widehat{\mathbf{\Sigma}}_u = \text{diag}\{\widehat{\sigma}_1^2, \cdots, \widehat{\sigma}_N^2\}$; 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: Data Driven Benchmarks						
	Sample	Linear Shrinkage	Factors Only	Equal Weights	Self-Banding	POET
CSI800	0.0593	0.0575	0.0547	0.0769	0.0589	0.0598
CSI500	0.0739	0.0703	0.0732	0.0820	0.0797	0.0774
HS300	0.0513	0.0480	0.0440	0.0717	0.0558	0.0528
• Panel B: Network-Guided Thresholding						
	analyst	industry	all_passage_1	all_sentence_1	one2one_passage_1	one2one_sentence_12
CSI800	0.0558	0.0548	0.0508	0.0503	0.0505	0.0510
CSI500	0.0722	0.0737	0.0685	**0.0683**	0.0686	0.0684
HS300	0.0507	0.0458	0.0457	0.0457	0.0447	0.0470
	one2one_sentence_1	all_passage_12	all_sentence_12	one2one_passage_12		Best Thresholding
CSI800	0.0506	0.0582	**0.0499**	0.0500		0.0499
CSI500	0.0685	0.0756	0.0700	0.0687		0.0683
HS300	0.0448	0.0508	0.0452	**0.0426**		0.0426
• Panel C: Network-Guided Banding						
	analyst	industry	all_passage_1	all_sentence_1	one2one_passage_1	one2one_sentence_12
CSI800	0.0598		0.0558	0.0556	**0.0532**	0.0537
CSI500	0.0742		0.0756	**0.0733**	0.0744	0.0768
HS300	0.0460		0.0483	0.0469	**0.0445**	0.0489
	one2one_sentence_1	all_passage_12	all_sentence_12	one2one_passage_12		Best Banding
CSI800	0.0538	0.0605	0.0588	0.0547		0.0532
CSI500	0.0737	0.0741	0.0735	0.0765		0.0733
HS300	0.0467	0.0488	0.0504	0.0513		0.0445

6.3.2 Other Mean-Variance Portfolios

We also compare the performance of various optimal portfolios under different covariance matrix estimators, focusing on the CSI500 index for illustration. We compute out-of-sample efficient frontiers using different methods, as shown in Figure 4, with all returns and volatilities annualized.

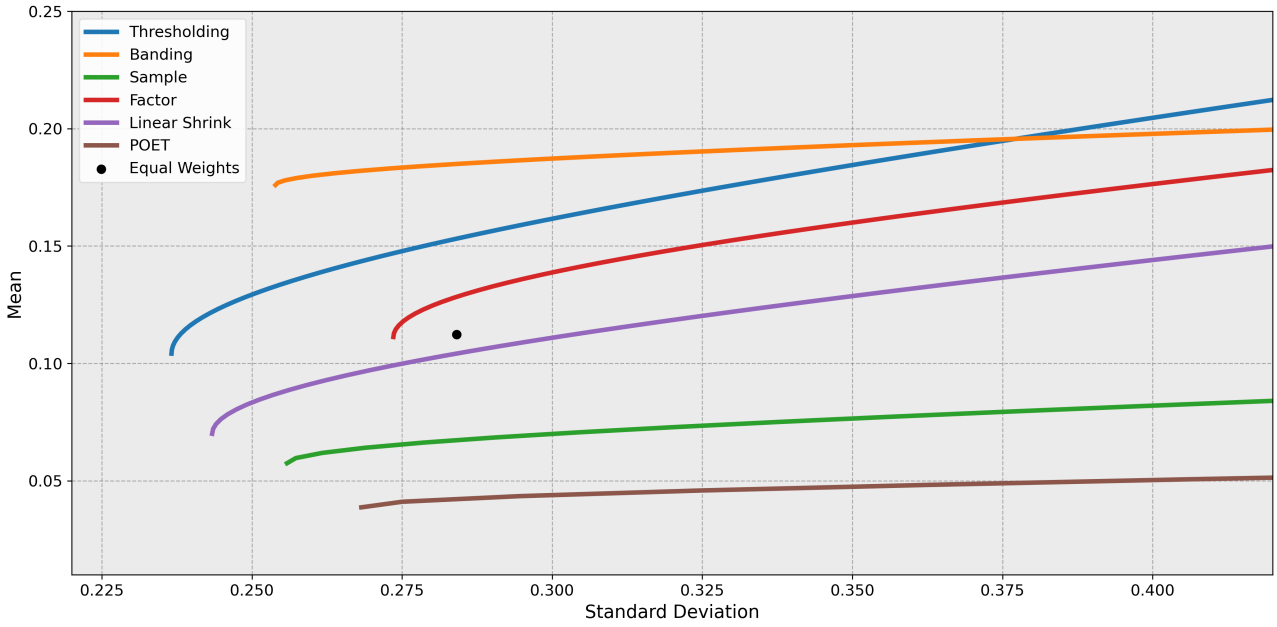


Figure 4: Out-of-sample Efficient Frontiers. For the two Network-Guided methods, we show the efficient frontiers using the best auxiliary information. "Thresholding" refers to Network-Guided Thresholding, while "Banding" refers to Network-Guided Banding.

From Figure 4, we see that Network-Guided Thresholding achieves the lowest variance, consistent with the results in Table 7. In terms of out-of-sample mean returns, the portfolio constructed using Network-Guided Banding outperforms both Network-Guided Thresholding and all other baseline models across most volatility levels. Thresholding yields lower mean returns than Banding at low volatility levels, but higher mean returns when volatility is relatively high. Aside from Thresholding and Banding, only the "Factor" method generates higher average returns than the "Equal Weights" portfolio. Linear Shrinkage performs better than the simple "Sample" method, but does not produce portfolios that outperform the "Factor" method. Finally, the sample covariance matrix and POET consistently underperform in constructing mean–variance portfolios in our study, highlighting the importance of using observed factors and improving the estimation of large covariance matrices for the remaining residuals. As documented by Fan et al. (2013), POET is primarily designed for settings in which the factor $\boldsymbol{f}_t$ is unobserved and must be inferred from the data, so estimating the latent factors introduces

an additional error term in the convergence rate. Consequently, POET is disadvantaged in our setting where robust observed factors are available.

Furthermore, we analyze the maximum Sharpe ratio (mean–variance optimal) portfolios under different models. For each model, we search the efficient frontier in Figure 4 to identify the mean–variance optimal portfolio for comparison. Figure 5 plots the backtest performance of these portfolios over a 10-year out-of-sample window, and the corresponding evaluation statistics are reported in Table 8. Due to the crash of the Chinese stock market in May and June 2015, none of the portfolios achieves a Sharpe ratio greater than 1. However, relative to benchmark methods, the Network-Guided portfolios perform better over the full period—especially the Banding approach. For example, the Network-Guided Thresholding strategy achieves a higher average return, a lower return standard deviation, and a higher Sharpe ratio than the "Factor Only" method. In addition, the Network-Guided Banding portfolio delivers the strongest backtest performance, with the highest return and the lowest maximum drawdown. It is also worth noting that a simple equal-weight portfolio performs better than many benchmark methods. This is not surprising: DeMiguel et al. (2009) documented that, in practice, the gains from mean–variance optimization are often more than offset by estimation error, making a naive equal-weight strategy more effective than commonly expected.

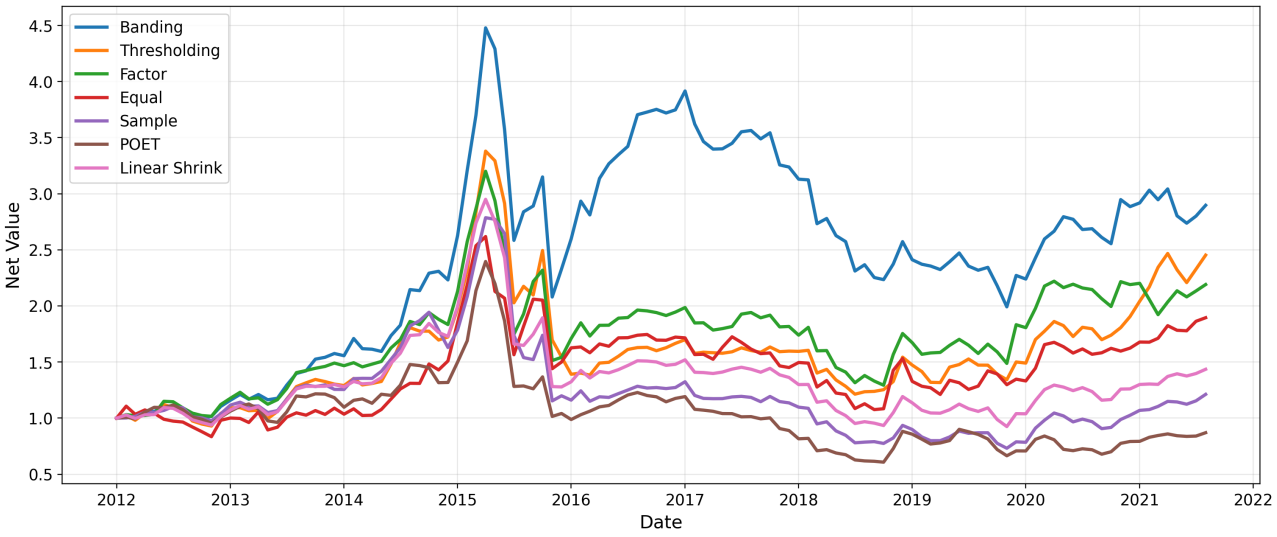


Figure 5: Out-of-sample Mean-Variance Optimal Portfolios. This figure tracks the net value of mean-variance portfolios constructed using different covariance matrix estimation methods. "Thresholding" refers to Network-Guided Thresholding, while "Banding" refers to Network-Guided Banding.

In conclusion, these empirical results validate the usefulness of incorporating network information into covariance matrix estimation for portfolio optimization. While the factor model primarily captures strong cross-sectional dependence among asset returns, auxiliary information—such as news, as discussed in Ge et al. (2022)—can help identify local or weak cross-sectional

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	**Factor**	**Equal**	**Linear Shr.**	**POET**	**Thresholding**	**Banding**
Mean Return	5.09%	11.02%	11.23%	6.35%	2.01%	12.57%	14.24%
Std. Dev.	25.58%	27.36%	28.41%	24.34%	26.82%	26.88%	27.19%
Sharpe Ratio	0.199	0.403	0.396	0.261	0.075	0.468	0.524
Max Draw-down	73.74%	59.58%	58.90%	68.65%	74.63%	64.09%	55.52%

dependencies. This is why auxiliary information improves the estimation of $\mathbf{\Sigma}_u$, the covariance matrix of de-factored returns. At the same time, these findings highlight the complexity of financial markets: the effectiveness of auxiliary information can vary across environments and conditions. We see at least two promising directions for future research. First, one could investigate the mechanisms driving these cross-market variations to further refine the estimation process. Second, following Gao et al. (2025), who employed penalized matrix regression to optimally combine multiple networks for covariance matrix forecasting, our framework could be extended to settings where a collection of networks $\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{M}^{(Q)}$ is available, enabling the construction of an optimal combined network.

7 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 improve the performance of conventional statistical and econometric models. Our study provides theoretical results showing that integrating auxiliary network data—when tailored to conventional thresholding and banding methods—yields improved properties. Both the simulation studies and the empirical illustrations confirm 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 yes: incorporating auxiliary network information of decent quality into conventional covariance matrix estimation methods is recommended.

In this paper, we focus primarily on estimating static covariance matrices. However, a similar approach can be extended to other settings, such as the estimation of large dynamic covariance matrices. For instance, time-varying network information could be 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.

M. Ao, Y. Li, and X. Zheng. Approaching mean-variance efficiency for large portfolios. *The Review of Financial Studies*, 32(7):2890–2919, 2019. doi: 10.1093/rfs/hhy090.

P. J. Bickel and A. Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

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.

P. J. Bickel, E. Levina, et al. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

M. Billio, M. Getmansky, A. Lo, and L. Pelizzon. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of financial economics*, 104(3): 535–559, 2012. doi: 10.1016/j.jfineco.2011.12.010.

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.

T. T. Cai, W. Liu, and X. Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011. doi: https://doi.org/10.1198/jasa.2011.tm10155.

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.

J. Chen, D. Li, Y. Li, and O. Linton. Estimating time-varying networks for high-dimensional time series. *Journal of Econometrics*, page 105941, 2025. doi: 10.1016/j.jeconom.2024.105941.

M. Chen, K. Kato, and C. Leng. Analysis of networks via the sparse β-model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):887–910, 2021.

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

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

P. Erdős, A. Rényi, et al. On the evolution of random graphs. *Publications of the*, 1960.

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, 2016a. doi: 10.1080/07350015.2015.1052458.

J. Fan, Y. Liao, and W. C. Wang. Projected principal component analysis in factor models. *Annals of statistic*, 44(1):219–254, 2016b. doi: 10.1214/15-AOS1364.

N. Flammarion, C. Mao, and P. Rigollet. Optimal rates of statistical seriation. *Bernoulli*, 25 (1):623–653, 2019. doi: 10.3150/17-BEJ1000.

Y. Gao, Z. Y. Zhang, Z. R. Cai, X. N. Zhu, T. Zou, and H. S. Wang. Penalized sparse covariance regression with high dimensional covariates. *Journal of Business & Economic Statistics*, 43 (3):615–626, 2025. doi: 10.1080/07350015.2024.2415109.

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. *Journal of Banking & Finance*, 171:107356, 2025.

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.

M. Handcock, A. Raftery, and T. J.M. Model-based clustering for social networks. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 170(2):301–354, 2007. doi: 10.1111/j.1467-985X.2007.00471.x.

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.

A. Inoue, L. Jin, and B. Rossi. Rolling window selection for out-of-sample forecasting with time-varying parameters. *Journal of econometrics*, 196(1):55–67, 2017. doi: 10.1016/j.jeconom.2016.03.006.

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.

C. Lam. High-dimensional covariance matrix estimation. *WIREs Computational Statistics*, 2019. doi: 10.1002/wics.1485.

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.

E. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. Springer, 2006.

S. Li, T. T. Cai, and H. Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84:149–173, 2020.

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

Y. Lin, Q. Zhu, and G. Li. Improving time series estimation and prediction via transfer learning, Oct. 2025. arXiv:2510.25236v1.

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.

F. Merlevède, M. Peligrad, and E. Rio. A bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3):435–474, 2011. doi: 10.1007/s00440-010-0304-9.

A. Rinaldo, S. Petrović, and S. Fienberg. Maximum lilkelihood estimation in the β-model. *The Annals of Statistics*, 41(3):1085–1110, 2013. doi: 10.1214/12-AOS1078.

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.

V. Solo. Pearson distance is not a distance. *arXiv:1908.06029*, 2019.

M. Tumminello, F. Lillo, and R. Mantegna. Correlation, hierarchies, and networks in financial markets. *Journal of economic behavior & organization*, 75(1):40–58, 2010. doi: 10.1016/j.jebo.2010.01.004.

C. Zhang, X. Pu, M. Cucuringu, and X. Dong. Graph-based methods for forecasting realized covariances. *Journal of Financial Econometrics*, 23(2):nbae026, 2025. doi: 10.1093/jjfinec/nbae026.

Appendices

A Additional Figures, Tables and Results

A.1 Cross-Validation Results

We report the proportion of time points at which the network is used in the Network-Guided Thresholding methods in Table 9. Overall, the `analyst` network is chosen with consistently high frequency (around 0.79–0.86) across all indices, whereas the `industry` network is selected much less often (around 0.11–0.16), suggesting that the industry network is relatively less informative in our empirical setting. Among the text-based networks, both the 1-period and 12-period constructions (`all_*` and `one2one_*`) are frequently selected, with particularly high usage for CSI800 (e.g., `all_passage_12` reaches 0.888, the largest value in the table). In contrast, CSI500 exhibits lower usage for some one-to-one text networks (e.g., `one2one_passage_1` and `one2one_sentence_12`), indicating heterogeneity in network quality and usefulness across indices.

Table 9: Network Usage Ratios Selected by Cross-Validation.

Index	**Network**				
	`analyst`	`industry`	`all_passage_1`	`all_sentence_1`	`one2one_passage_1`
HS300	0.793	0.112	0.716	0.767	0.707
CSI500	0.836	0.155	0.750	0.698	0.586
CSI800	0.862	0.112	0.853	0.853	0.776
	`one2one_sentence_12`	`one2one_sentence_1`	`all_passage_12`	`all_sentence_12`	`one2one_passage_12`
HS300	0.759	0.828	0.793	0.802	0.776
CSI500	0.629	0.836	0.853	0.767	0.845
CSI800	0.853	0.879	0.888	0.845	0.862

In addition to the average usage ratios, Figure 6 visualizes the *time-varying* network selection outcomes from cross-validation. Each dot indicates that, in the corresponding month, cross-validation selects the network-guided estimator under that network type; the absence of a dot indicates that the procedure falls back to the benchmark (i.e., ignoring the network). Several additional patterns emerge from the time-series visualization. First, the `industry` network is only selected sporadically, appearing as isolated dots across time, which suggests that its usefulness is not persistent and is confined to occasional periods. Second, most text-based networks (both 1-period and 12-period constructions) are selected frequently, but the figure reveals that their selections are not uniform over time: network usage can exhibit clustering and regime-like

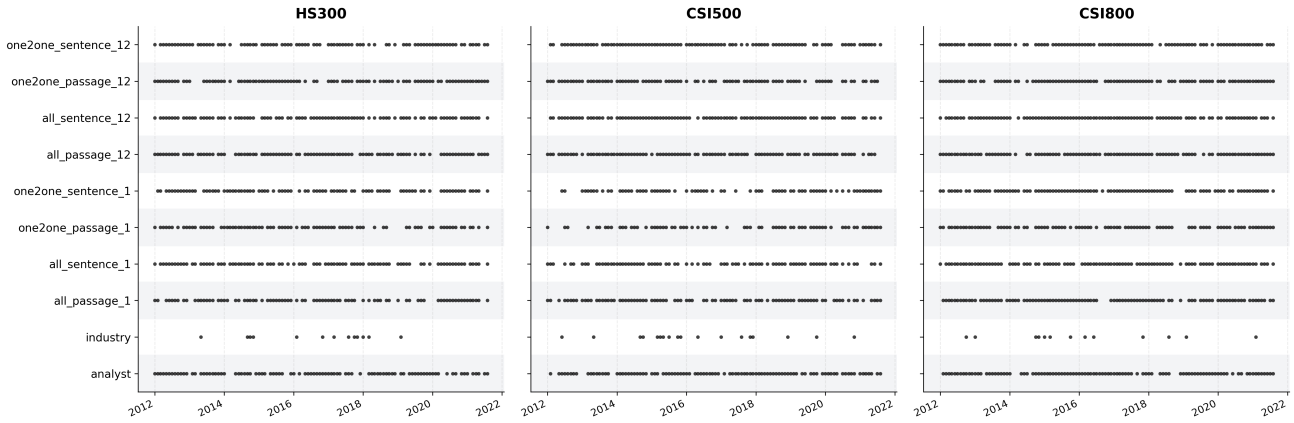


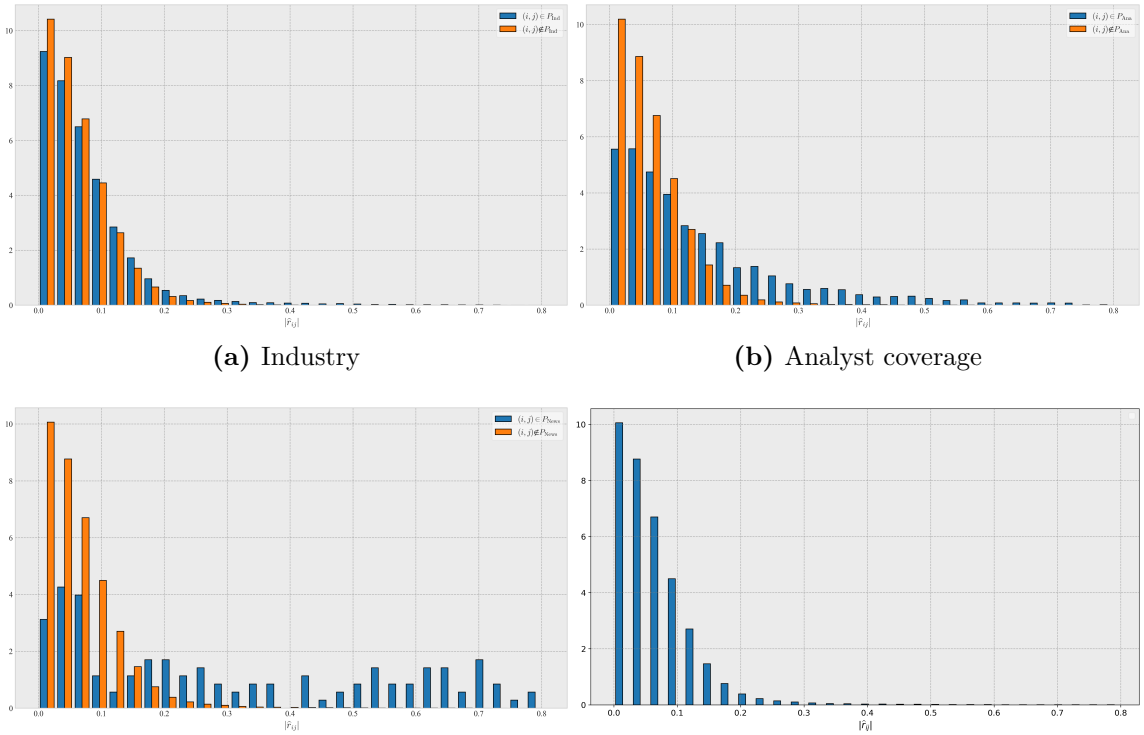
Figure 6: Time-varying Network Usage Selected by Cross-Validation. For each month and each network type on the vertical axis, a dot indicates that cross-validation selects the corresponding network-guided estimator; otherwise, the procedure falls back to the benchmark that ignores the network.

persistence, interspersed with gaps where the fallback estimator is chosen. Third, the temporal selection patterns differ across indices: while HS300 and CSI800 show relatively persistent usage for several text-based networks, CSI500 displays more frequent switching across network types, indicating stronger time variation in the incremental value of network guidance. Overall, the figure complements the table-level averages by revealing when (and how persistently) each network contributes to the estimator across time.

A.2 Realized Residual Correlation and Network Classification

To assess the informativeness of auxiliary network information in practice, we examine how different network sources relate to the realized dependence in idiosyncratic returns. Specifically, we follow Fan et al. (2016a) and compute the realized pairwise sample residual correlations and for each auxiliary network source $s \in \{\text{ind}, \text{ana}, \text{news}\}$, let P_s denote the set of linked pairs implied by that source. We then split all pairs into the linked group $\{(i,j) \in P_s\}$ and the unlinked group $\{(i,j) \notin P_s\}$, and compare the empirical distributions of $|\widehat{r}_{ij}|$ across the two groups.

Figure 7 shows a clear separation between linked and unlinked pairs across the three auxiliary networks: linked pairs tend to exhibit larger realized correlations, whereas unlinked pairs concentrate much more heavily near zero. This pattern supports our modeling premise that auxiliary links are informative about strong latent dependence (i.e., the likelihood of observing a link increases with $|r_{ij}|$), while still being noisy in finite samples. The strength of separation varies by network source. The industry network (Panel (a)) provides a mild yet visible shift toward larger $|\widehat{r}_{ij}|$ for linked pairs, consistent with the idea that firms in the same industry are



(a) Industry **(b)** Analyst coverage

(c) News co-mention **(d)** Overall

Figure 7: Distributions of Sample Correlations for Linked and Unlinked Pairs. Each panel compares the empirical distributions of sample correlations for entity pairs with a link or without a link as defined by different sources of auxiliary network information.

more likely to co-move. The analyst-coverage network (Panel (b)) yields a more pronounced distinction, with linked pairs exhibiting a heavier right tail. Most notably, the news co-mention network (Panel (c)) produces the sharpest contrast: linked pairs display a substantially heavier tail, indicating that co-mentions are strongly associated with large residual correlations. For reference, Panel (d) reports the unconditional distribution of $|\widehat{r}_{ij}|$ when no network classification is applied, which is dominated by small correlations. Some overlap between the two distributions is expected, since $\widehat{r}_{ij}$ is noisy and the auxiliary links are imperfect proxies for dependence. In this sense, $\widehat{L}_{ij}$ provides complementary information to $\widehat{r}_{ij}$ about the same underlying correlation structure. By incorporating auxiliary information $\mathcal{I}(N,T)$, we can therefore improve estimation accuracy in finite samples, consistent with our theoretical results.

A.3 Latent Space Model

As an illustration, we consider a special case in which g_T follows a latent space model with inverse Kullback–Leibler divergence distance[20] $d(r) = -[\log(1-r^2)]^{-1}$ and the shape parameter

[20] Solo (2019) documented that distances based on Pearson correlation are usually semimetrics rather than metrics since the triangle inequality may not always hold.

b_T. The corresponding rates can be taken as $\varrho_{0T} = \varrho_{1T} = O\big(\exp\big\{-\frac{\nu_N b_T}{\lambda^3}\big\}\big)$.

Proposition 1. *[Latent Space Model] Assume $g_T(|r_{ij}|)$ takes the form of a latent space model:*

$$g_T(|r_{ij}|) = \frac{\exp\{b_T d(\lambda) - b_T d_{ij}\}}{1 + \exp\{b_T d(\lambda) - b_T d_{ij}\}}, \tag{25}$$

where the distance function is defined as $d_{ij} = d(|r_{ij}|) := -\left[\log\left(1 - r_{ij}^2\right)\right]^{-1}$, then we have $\varrho_{0T} = \varrho_{1T} = \exp\big\{-\frac{\nu_N b_T}{\lambda^3}\big\}$.

Proof. First, we have

$$g_T(r) = \frac{1}{1 + \exp\{b_T(d(r) - d(\lambda))\}}.$$

A simple calculation gives

$$d'(r) = -\frac{2r}{(1 - r^2)\left[\log(1 - r^2)\right]^2} < 0$$

for $r \in (0, 1)$, and similarly $d''(r) > 0$.

When $r < \lambda - \nu_N$, we define

$$\begin{aligned}\mathcal{X}_T :=& b_T\left(d(r) - d(\lambda)\right) \geq b_T\left(d(\lambda - \nu_N) - d(\lambda)\right) \geq -a_T d'(\lambda)\nu_N \\ =& \frac{2\lambda}{(1 - \lambda^2)\left[\log(1 - \lambda^2)\right]^2} b_T \nu_N.\end{aligned}$$

Therefore

$$g_T(r) = \frac{1}{1 + e^{\mathcal{X}_T}} \leq e^{-\mathcal{X}_T} \leq \exp\left\{-\frac{2\lambda b_T \nu_N}{(1 - \lambda^2)\left[\log(1 - \lambda^2)\right]^2}\right\}.$$

And note that

$$\frac{2\lambda}{(1 - \lambda^2)\left[\log(1 - \lambda^2)\right]^2} = \frac{2\lambda}{(1 + O(\lambda^2))(\lambda^2 + O(\lambda^4))^2} = \frac{2}{\lambda^3} \cdot \{1 + o(1)\},$$

then for (N, T) large, the left hand side must be larger than $1/\lambda^3$. Hence,

$$g_T(r) \leq \exp\left\{-\frac{\nu_N b_T}{\lambda^3}\right\}, \quad r < \lambda - \nu_N.$$

When $r > \lambda + \nu_N$, one has

$$\begin{aligned}\mathcal{Y}_N :=& -b_T\left(d(r) - d(\lambda)\right) \geq -b_T\left(d(\lambda + \nu_N) - d(\lambda)\right) \geq -b_T \nu_N d'(\lambda) \\ =& \frac{2\lambda}{(1 - \lambda^2)\left[\log(1 - \lambda^2)\right]^2} b_T \nu_N.\end{aligned}$$

Therefore,

$$1 - g_T(r) = \frac{e^{-\mathcal{Y}_T}}{1 + e^{-\mathcal{Y}_T}} \leq e^{-\mathcal{Y}_T} \leq \exp\left\{-\frac{\nu_N b_T}{\lambda^3}\right\}, \quad r > \lambda + \nu_N.$$

Combining the two cases establishes the desired result. □

B Proofs

B.1 Proof of Theorem 1

Before constructing the proof of Theorem 1, we first introduce the following lemma.

Lemma 1. *Under Assumption 1, we have*

$$\max_{i,j} |\widehat{r}_{ij} - r_{ij}| \leq O_P\left(\max_{i,j} |\widehat{\sigma}_{ij} - \sigma_{ij}|\right).$$

Proof. For some constant $A > 0$, define the event

$$A_{1T} = \left\{\max_{i,j} |\widehat{\sigma}_{ij} - \sigma_{ij}| > A\sqrt{\tfrac{\log N}{T}}\right\} \cup \left\{\max_{i,j} |\widehat{\sigma}_{ij} - \widetilde{\sigma}_{ij}| > A a_T\right\}, \tag{26}$$

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

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

Consider the function

$$g(\sigma_{ij}, \sigma_{ii}, \sigma_{jj}) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}},$$

which maps covariance entries to correlation coefficients. Its first-order partial derivatives are

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

We can show that these derivatives are uniformly bounded. First, Equation 6 implies $\sigma_{ii} \leq M$ and $\sigma_{jj} \leq M$. Consequently, $|\sigma_{ij}| \leq M$ follows from the Cauchy–Schwarz inequality. Moreover, condition (a) in Assumption 1, which requires $\rho_{\min}(\mathbf{\Sigma}_u) > \underline{c} > 0$, ensures that σ_{ii} and σ_{jj} are bounded away from zero.

When A_{1T}^c occurs and T is sufficiently large, $\widehat{\sigma}_{ij}$ and σ_{ij} are close enough that $|\widehat{\sigma}_{ij}|$ remains bounded, and $\widehat{\sigma}_{ii}, \widehat{\sigma}_{jj}$ are also bounded away from zero. Since the partial derivatives of g are

bounded, it follows that

$$\max_{i,j}|\widehat{r}_{ij}-r_{ij}|\leq O\left(\max_{i,j}|\widehat{\sigma}_{ij}-\sigma_{ij}|\right)\quad \text{under } A_{1T}^{c},$$

which completes the proof. □

We now begin the proof of Theorem 1.

Proof. By the triangle inequality,

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

For the first part, we have

$$\begin{aligned}\left\|\boldsymbol{R}_{\boldsymbol{L}}^{\mathcal{T}}-\boldsymbol{R}\right\| &\leq \max_{1\leq i\leq N}\sum_{j=1}^{N}\left(L_{ij}\left|r_{ij}-r_{ij}\right|+L_{ij}^{0}\left|s_{\lambda}\left(r_{ij}\right)-r_{ij}\right|\right)\\ &=\max_{1\leq i\leq N}\sum_{j=1}^{N}\left(L_{ij}\cdot 0+L_{ij}^{0}\left|0-r_{ij}\right|\right)=\max_{1\leq i\leq N}\sum_{j=1}^{N}L_{ij}^{0}\left|r_{ij}\right|\\ &=\max_{1\leq i\leq N}\sum_{j=1}^{N}L_{ij}^{0}\left|r_{ij}\right|^{q}\left|r_{ij}\right|^{1-q}\leq\lambda^{1-q}\max_{1\leq i\leq N}\sum_{j=1}^{N}L_{ij}^{0}\left|r_{ij}\right|^{q}\\ &=\lambda^{1-q}c_{0}\left(N\right).\end{aligned}$$

And for the second part, we have

$$\begin{aligned}\left\|\widehat{\boldsymbol{R}}_{\widehat{\boldsymbol{L}}}^{\mathcal{T}}-\boldsymbol{R}_{\boldsymbol{L}}^{\mathcal{T}}\right\| \leq \max_{1\leq i\leq N}\sum_{j=1}^{N}&\left(\left|\widehat{r}_{ij}-r_{ij}\right|I_{\left\{L_{ij}=1,\widehat{L}_{ij}=1\right\}}+\left|s_{\lambda}\left(\widehat{r}_{ij}\right)-r_{ij}\right|I_{\left\{L_{ij}=1,\widehat{L}_{ij}=0\right\}}\right.\\ &\left.+\left|s_{\lambda}\left(\widehat{r}_{ij}\right)\right|I_{\left\{L_{ij}=0,\widehat{L}_{ij}=0\right\}}+\left|\widehat{r}_{ij}\right|I_{\left\{L_{ij}=0,\widehat{L}_{ij}=1\right\}}\right)\\ =&\mathrm{I}+\mathrm{II}+\mathrm{III}+\mathrm{IV}.\end{aligned}$$

To bound term I, we write

$$\mathrm{I}=\max_{1\leq i\leq N}\sum_{j=1}^{N}\left|\widehat{r}_{ij}-r_{ij}\right|I_{\left\{L_{ij}=1,\widehat{L}_{ij}=1\right\}}\leq c_{0}\left(N\right)\cdot\max_{i,j}\left|\widehat{r}_{ij}-r_{ij}\right|.$$

If the event A_{1T}^{c} holds, then

$$\mathrm{I}\leq A\,c_{0}(N)\sqrt{\tfrac{\log N}{T}}.$$

For term II, assuming A_{1T}^c holds and let $\nu = \nu_N > 0$, then

$$\begin{aligned}\text{II} &\leq A\lambda \max_{1\leq i\leq N} \sum_{j=1}^{N} \left(I_{\{|r_{ij}|>\lambda+\nu,\, \widehat{L}_{ij}=0\}} + I_{\{\lambda<|r_{ij}|\leq\lambda+\nu,\, \widehat{L}_{ij}=0\}} \right) \\ &\leq A\lambda \left(\max_{1\leq i\leq N} \sum_{j=1}^{N} I_{\{|r_{ij}|>\lambda+\nu,\, \widehat{L}_{ij}=0\}} + Q_N(\lambda, \lambda+\nu) \right).\end{aligned}$$

For the first summation term, note that

$$E\left(\lambda \sum_{j=1}^{N} I_{\{\widehat{L}_{ij}=0,\, |r_{ij}|>\lambda+\nu\}} \right) \leq \varrho_{1T} c_0(N)\lambda \;\leq\; \varrho_T,$$

$$\text{Var}\left(\lambda \sum_{j=1}^{N} I_{\{\widehat{L}_{ij}=0,\, |r_{ij}|>\lambda+\nu\}} \right) \leq \varrho_{1T} c_0(N)\lambda^2(1-\varrho_{1T}) \;\leq\; \varrho_T\lambda.$$

Define the event

$$G_{1T} := \left\{ \lambda \sum_{j=1}^{N} I_{\{\widehat{L}_{ij}=0,\, |r_{ij}|>\lambda+\nu\}} > A\sqrt{\varrho_T} \right\}.$$

By Chebyshev's inequality,

$$P(G_{1T}) = O(\lambda)\,.$$

For the second summation term in II, Equation 3 implies an upper bound of order $O\left(c_0(N)\sqrt{\frac{\log N}{T}} \right)$. Combining these results, on the event $A_{1T}^c \cap G_{1T}^c$ we obtain

$$\text{II} \leq A\left(\sqrt{\varrho_T} + c_0(N)\sqrt{\tfrac{\log N}{T}} \right).$$

For term III, we can write

$$\begin{aligned}\text{III} &\leq \max_{1\leq i\leq N} \sum_{j=1}^{N} |\widehat{r}_{ij}|\, I_{\left\{L_{ij}=0, \widehat{L}_{ij}=0, |\widehat{r}_{ij}|>\lambda\right\}} \\ &\leq \max_{1\leq i\leq N} \sum_{j=1}^{N} |r_{ij}|\, I_{\left\{L_{ij}=0, \widehat{L}_{ij}=0, |\widehat{r}_{ij}|>\lambda\right\}} \\ &\quad + \max_{1\leq i\leq N} \sum_{j=1}^{N} |\widehat{r}_{ij} - r_{ij}|\, I_{\left\{L_{ij}=0, \widehat{L}_{ij}=0, |\widehat{r}_{ij}|>\lambda\right\}} \\ &=: \text{V} + \text{VI}.\end{aligned}$$

Here, we have

$$\mathrm{V} \leq \max_{1\leq i\leq N} \sum_{j=1}^{N} |r_{ij}|\, I_{\{L_{ij}=0\}} \leq \lambda^{1-q} \max_{1\leq i\leq N} \sum_{j=1}^{N} |r_{ij}|^{q}\, I_{\{L_{ij}=0\}} = \lambda^{1-q} c_0\left(N\right).$$

For VI, some $\nu = \nu_N > 0$, we have

$$\begin{aligned}
\mathrm{VI} \leq & \max_{i,j} |\widehat{r}_{ij} - r_{ij}| \cdot \left(\max_{1\leq i\leq N} \sum_{j=1}^{N} \left(I_{\left\{\widehat{L}_{ij}=0, |\widehat{r}_{ij}|>\lambda, |r_{ij}|\leq \lambda-\nu\right\}} + I_{\left\{\widehat{L}_{ij}=0, |\widehat{r}_{ij}|>\lambda, \lambda-\nu<|r_{ij}|\leq\lambda\right\}} \right) \right) \\
\leq & \max_{i,j} |\widehat{r}_{ij} - r_{ij}| \cdot \left(\max_{1\leq i\leq N} \sum_{j=1}^{N} I_{\{|\widehat{r}_{ij}|>\lambda, |r_{ij}|\leq\lambda-\nu\}} + \max_{1\leq i\leq N} \sum_{j=1}^{N} I_{\left\{\widehat{L}_{ij}=0, \lambda-\nu<|r_{ij}|\leq\lambda\right\}} \right).
\end{aligned}$$

The first summation in VI here can be bounded using the Bernstein inequality for α-mixing sequence, proposed by Merlevède et al. (2011). Specifically, we suppose that A_{1T}^{c} happens, and define

$$G_{2T} := \left\{ \max_{1\leq i\leq N} \sum_{j=1}^{N} I_{\{|r_{ij}|\leq\lambda-\nu, |\widehat{r}_{ij}|>\lambda\}} > 0 \right\},$$

we have

$$\begin{aligned}
P\left(G_{2T}\right) = & P\left(\max_{i,j} |\widehat{r}_{ij} - r_{ij}| > (1-\nu)\,\lambda \right) \\
\leq & P\left(\max_{i,j} |\widehat{\sigma}_{ij} - \sigma_{ij}| > \frac{1}{2}\underline{c}\,(1-\nu)\,\lambda \right).
\end{aligned}$$

Since $|\widehat{\sigma}_{ij} - \sigma_{ij}| \leq Aa_T + |\widetilde{\sigma}_{ij} - \sigma_{ij}|$, it suffices to consider $|\widetilde{\sigma}_{ij} - \sigma_{ij}|$. For that, we have

$$P\left(\max_{i,j} |\widetilde{\sigma}_{ij} - \sigma_{ij}| > x \right) \leq N^2 P\left(|\widetilde{\sigma}_{ij} - \sigma_{ij}| > x \right) \leq N^2 P\left(\left| \sum_{t=1}^{T} \left(u_{it}u_{jt} - \sigma_{ij} \right) \right| > Tx \right).$$

By Theorem 1 of Merlevède et al. (2011), for some constants $C_1, \ldots, C_5$ depending on ϕ_2, ϕ_4,

$$\begin{aligned}
P\left(\left| \sum_{t=1}^{T} (u_{it}u_{jt} - \sigma_{ij}) \right| > Tx \right) \leq & \, T\exp\left(-\tfrac{(Tx)^{\gamma}}{C_1} \right) + \exp\left(-\tfrac{T^2x^2}{C_2(1+TC_3)} \right) \\
& + \exp\left(-\tfrac{(Tx)^2}{C_4 T} \exp\left(\tfrac{(Tx)^{(1-\gamma)\gamma}}{C_5(\log(Tx))^{\gamma}} \right) \right).
\end{aligned}$$

Taking $x = \frac{1}{2}\underline{c}(1-\nu)\lambda$ and using condition (d) of Assumption 1, one obtains

$$N^2 \left(Te^{-(Tx)^{\gamma}/C_1} + e^{-T^2x^2/(C_2(1+TC_3))} + e^{-\frac{(Tx)^2}{C_4 T} \exp\{(Tx)^{(1-\gamma)\gamma}/(C_5(\log(Tx))^{\gamma})\}} \right) = O\left(\tfrac{1}{N^2}\right).$$

Together with $a_T = o(\lambda)$, we conclude

$$P(G_{2T}) = O\big(\tfrac{1}{N^2} + \kappa_1(N,T)\big)\,.$$

For the remaining part in VI, we have

$$\max_{i,j}|\widehat{r}_{ij} - r_{ij}| \cdot \max_{1\le i\le N}\sum_{j=1}^{N} I_{\left\{\widehat{L}_{ij}=0,\lambda-\nu<|r_{ij}|\le\lambda\right\}} \le \max_{i,j}|\widehat{r}_{ij} - r_{ij}| \cdot Q_N(\lambda-\nu,\lambda) \le A\sqrt{\frac{\log N}{T}}c_0(N).$$

If A_{1T}^c holds, then it can be bounded by $Ac_0\,(N)\,\sqrt{\frac{\log N}{T}}$. Therefore,

$$\text{VI} \le Ac_0(N)\sqrt{\tfrac{\log N}{T}}.$$

Combining the bounds for V and VI, we obtain

$$\text{III} \le A\Big(\lambda^{\,1-q}c_0(N) + c_0(N)\sqrt{\tfrac{\log N}{T}}\Big).$$

For the term IV, suppose A_{1T}^c holds and let $\nu = \nu_N > 0$, then

$$\begin{aligned}\text{IV} &\le \max_{i,j}|\widehat{r}_{ij} - r_{ij}| \cdot \max_{1\le i\le N}\sum_{j=1}^{N}\left(I_{\{\widehat{L}_{ij}=1,\,|r_{ij}|\le\lambda-\nu\}} + I_{\{\widehat{L}_{ij}=1,\,\lambda-\nu<|r_{ij}|\le\lambda\}}\right)\\ &\le A\sqrt{\tfrac{\log N}{T}}\left(\max_{1\le i\le N}\sum_{j=1}^{N} I_{\{\widehat{L}_{ij}=1,\,|r_{ij}|\le\lambda-\nu\}} + Q_N(\lambda-\nu,\lambda)\right).\end{aligned}$$

For the first summation, we have

$$\begin{aligned}&E\left(\sqrt{\frac{\log N}{T}}\sum_{j=1}^{N} I_{\left\{\widehat{L}_{ij}=1,|r_{ij}|\le\lambda-\nu\right\}}\right) \le \varrho_{0T}N\sqrt{\frac{\log N}{T}} \le \varrho_T,\\ &\text{Var}\left(\sqrt{\frac{\log N}{T}}\sum_{j=1}^{N} I_{\left\{\widehat{L}_{ij}=1,|r_{ij}|\le\lambda-\nu\right\}}\right) \le \varrho_{0T}N\frac{\log N}{T}\,(1-\varrho_{0T}) \le \varrho_T\sqrt{\frac{\log N}{T}}.\end{aligned}$$

Define the event

$$G_{3T} := \left\{\sqrt{\tfrac{\log N}{T}}\sum_{j=1}^{N} I_{\{\widehat{L}_{ij}=1,\,|r_{ij}|\le\lambda-\nu\}} > A\sqrt{\varrho_T}\right\}.$$

By Chebyshev's inequality,

$$P(G_{3T}) \le O\left(\sqrt{\tfrac{\log N}{T}}\right).$$

Thus, with probability $1 - O\left(\sqrt{\frac{\log N}{T}}\right)$, the first summation term is bounded by $A\sqrt{\varrho_T}$. The

remaining term in IV can be bounded exactly as in the analysis of Term VI. By Equation 3, this contributes at most order $c_0(N)\sqrt{\frac{\log N}{T}}$. Therefore, under $A_{1T}^c \cap G_{3T}^c$, we obtain

$$\text{IV} \leq c_0(N)\sqrt{\tfrac{\log N}{T}} + \sqrt{\varrho_T}.$$

Combining the bounds for terms I, II, III and IV, we obtain that on the event $A_{1T}^c \cap G_{1T}^c$,

$$\left\|\widehat{\boldsymbol{R}}_{\widehat{\boldsymbol{L}}}^{\mathcal{T}} - \boldsymbol{R}\right\| \leq A\left(c_0(N)\left(\lambda^{1-q} + \sqrt{\tfrac{\log N}{T}}\right) + \sqrt{\varrho_T}\right).$$

Consequently,

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

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

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

The first term is bounded by $O\left(\left\|\widehat{\boldsymbol{R}}_{\widehat{\boldsymbol{L}}}^{\mathcal{T}} - \boldsymbol{R}\right\|\right)$ provided $\sigma_{ii} < M$ in Equation 6 and the event A_{1T}^c holds (which makes sure $\widehat{\sigma}_{ii} < M + A\sqrt{\frac{\log N}{T}} < 2M$ when T large). For the second term, under A_{1T}^c, again, both $\boldsymbol{D}$ and $\widehat{\boldsymbol{D}}$ are bounded, we have

$$\begin{aligned}\|\widehat{\boldsymbol{D}}\boldsymbol{R}\widehat{\boldsymbol{D}} - \boldsymbol{D}\boldsymbol{R}\boldsymbol{D}\| &\leq \|\widehat{\boldsymbol{D}}\boldsymbol{R}(\widehat{\boldsymbol{D}} - \boldsymbol{D})\| + \|(\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{R}\boldsymbol{D}\| \\ &= O\left(A\sqrt{\tfrac{\log N}{T}}\right).\end{aligned}$$

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

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

which ends the proof. □

B.2 Proof of Theorem 2

Remark: Here, to proceed to theoretical proof, we may need a negative rank correlation assumption. Specifically, note that

$$\sum_{j=1}^{N} I_{\{j\in S_k^{\widehat{\boldsymbol{c}}_i}\}} = k, \quad \text{for all } i,$$

which implies that $I_{\{j_1\in S_k^{\widehat{\boldsymbol{c}}_i}\}}$ and $I_{\{j_2\in S_k^{\widehat{\boldsymbol{c}}_i}\}}$ are negatively dependent. Therefore, it is natural to impose the assumption

$$\operatorname{Cov}\left(I_{\{j_1\in S_k^{\widehat{\boldsymbol{c}}_i}\}}, I_{\{j_2\in S_k^{\widehat{\boldsymbol{c}}_i}\}}\right) \le 0, \quad \text{for all } i, j_1, j_2. \tag{27}$$

This means that when we already know j_1 is a k-largest neighbor of i, then the possibility of any other j_2 being a k-largest neighbor would decrease.

Now we start to prove Theorem 2.

Proof. By triangle inequality, we have $\left\|\widehat{\boldsymbol{R}}^{\mathcal{B}}_{\widehat{\boldsymbol{C}}} - \boldsymbol{R}\right\| \le \left\|\boldsymbol{R}^{\mathcal{B}}_{\boldsymbol{C}} - \boldsymbol{R}\right\| + \left\|\widehat{\boldsymbol{R}}^{\mathcal{B}}_{\widehat{\boldsymbol{C}}} - R^{\mathcal{B}}_{\boldsymbol{C}}\right\|$. For the operator norm, the first part is

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

For the second part, we have

$$\begin{aligned}
\left\|\widehat{\boldsymbol{R}}^{\mathcal{B}}_{\widehat{\boldsymbol{C}}} - \boldsymbol{R}^{\mathcal{B}}_{\boldsymbol{C}}\right\| \le \max_{1\le i\le N}\sum_{j=1}^{N}\Bigg(&\left|\widehat{r}_{ij} - r_{ij}\right|\cdot I_{\left\{(i,j)\in S_k^{\boldsymbol{c}_j}\times S_k^{\boldsymbol{c}_i}, (i,j)\in S_k^{\widehat{\boldsymbol{c}}_j}\times S_k^{\widehat{\boldsymbol{c}}_i}\right\}} + \left|r_{ij}\right|\cdot I_{\left\{(i,j)\in S_k^{\boldsymbol{c}_j}\times S_k^{\boldsymbol{c}_i}, (i,j)\notin S_k^{\widehat{\boldsymbol{c}}_j}\times S_k^{\widehat{\boldsymbol{c}}_i}\right\}} \\
&+ 0\cdot I_{\left\{(i,j)\notin S_k^{\boldsymbol{c}_j}\times S_k^{\boldsymbol{c}_i}, (i,j)\notin S_k^{\widehat{\boldsymbol{c}}_j}\times S_k^{\widehat{\boldsymbol{c}}_i}\right\}} + \left|\widehat{r}_{ij}\right|\cdot I_{\left\{(i,j)\notin S_k^{\boldsymbol{c}_j}\times S_k^{\boldsymbol{c}_i}, (i,j)\in S_k^{\widehat{\boldsymbol{c}}_j}\times S_k^{\widehat{\boldsymbol{c}}_i}\right\}}\Bigg) \\
=: \mathrm{I} + \mathrm{II} + 0 + \mathrm{III}.&
\end{aligned}$$

To bound I, on the event A^c_{1T}, we have

$$\begin{aligned}
\mathrm{I} &\le A\sqrt{\tfrac{\log N}{T}}\max_{1\le i\le N}\sum_{j=1}^{N} I_{\{(i,j)\in S_k^{\boldsymbol{c}_j}\times S_k^{\boldsymbol{c}_i}, (i,j)\in S_k^{\widehat{\boldsymbol{c}}_j}\times S_k^{\widehat{\boldsymbol{c}}_i}\}} \\
&\le Ak\sqrt{\tfrac{\log N}{T}}.
\end{aligned}$$

Here the last inequality follows because each set $S_k^{\boldsymbol{c}_i}$ (and similarly $S_k^{\widehat{\boldsymbol{c}}_i}$) contains at most k indices, so the summation contributes at most k nonzero terms for each i.

For term II, we first have

$$I_{\left\{(i,j)\in S_k^{\boldsymbol{c}_j}\times S_k^{\boldsymbol{c}_i},(i,j)\notin S_k^{\widehat{c}_j}\times S_k^{\widehat{\boldsymbol{c}}_i}\right\}} \leq I_{\left\{i\in S_k^{\boldsymbol{c}_j},i\notin S_k^{\widehat{\boldsymbol{c}}_j}\right\}} + I_{\left\{j\in S_k^{\boldsymbol{c}_i},j\notin S_k^{\widehat{\boldsymbol{c}}_i}\right\}}.$$

Then the summation can be bounded by

For term II, we first have

$$\begin{aligned} P\left((i,j)\in S_k^{\boldsymbol{c}_j}\times S_k^{\boldsymbol{c}_i},(i,j)\notin S_k^{\widehat{\boldsymbol{c}}_j}\times S_k^{\widehat{\boldsymbol{c}}_i}\right) \leq & P\left(i\in S_k^{\boldsymbol{c}_j},i\notin S_k^{\widehat{\boldsymbol{c}}_j}\right)+P\left(j\in S_k^{\boldsymbol{c}_i},j\notin S_k^{\widehat{\boldsymbol{c}}_i}\right) \\ \leq & 2\varrho_{1T}. \end{aligned}$$

Therefore,

$$E(\mathrm{II}) \leq A\varrho_{1T}\max_{1\leq i\leq N}\sum_{j=1}^{N}|r_{ij}|I_{\left\{(i,j)\in S_k^{\boldsymbol{c}_j}\times S_k^{\boldsymbol{c}_i}\right\}} \leq Ak\varrho_{1T}\min\{1,k^{-\alpha}b_0(N)\} = A\varrho_T \to 0.$$

Besides, we have

$$\begin{aligned} \mathrm{Var}\left(\sum_{j=1}^{N}|r_{ij}|\,I_{\left\{(i,j)\in S_k^{\boldsymbol{c}_j}\times S_k^{\boldsymbol{c}_i},(i,j)\notin S_k^{\widehat{\boldsymbol{c}}_j}\times S_k^{\widehat{\boldsymbol{c}}_i}\right\}}\right) \leq & \sum_{j=1}^{N}|r_{ij}|^2\,P\left((i,j)\in S_k^{\boldsymbol{c}_j}\times S_k^{\boldsymbol{c}_i},(i,j)\notin S_k^{\widehat{\boldsymbol{c}}_j}\times S_k^{\widehat{\boldsymbol{c}}_i}\right) \\ & + \text{covariance term}. \end{aligned} \tag{28}$$

From Equation 27, the covariance term is negative. Thus one has

$$\mathrm{Var}(\mathrm{II}) \leq A\varrho_{1T}\max_{1\leq i\leq N}\sum_{j=1}^{N}|r_{ij}|^2 I_{\left\{(i,j)\in S_k^{\boldsymbol{c}_j}\times S_k^{\boldsymbol{c}_i}\right\}} \leq Ak\varrho_{1T}\min\{1,k^{-2\alpha}b_0(N)\} \leq A\varrho_T k^{-\alpha}b_0(N).$$

Then Chebyshev's inequality leads to

$$P\left(\mathrm{II} > A\sqrt{\varrho_T}\right) = O\left(k^{-\alpha}b_0(N)\right).$$

For term III, we can decompose

$$\mathrm{III} \leq \max_{1\leq i\leq N}\sum_{j=1}^{N}\left(|\widehat{r}_{ij}-r_{ij}|+|r_{ij}|\right)I_{\left\{(i,j)\notin S_k^{\boldsymbol{c}_j}\times S_k^{\boldsymbol{c}_i},(i,j)\in S_k^{\widehat{\boldsymbol{c}}_j}\times S_k^{\widehat{\boldsymbol{c}}_i}\right\}} =: \mathrm{IV}+\mathrm{V}.$$

Since the summation contributes at most k nonzero terms, then IV is bounded by $Ak\sqrt{\frac{\log N}{T}}$

on the event A_{1T}^c. For V, we have

$$\mathrm{V} \leq \max_{1\leq i\leq N}\sum_{j=1}^{N}|r_{ij}|I_{\left\{(i,j)\notin S_k^{c_j}\times S_k^{c_i}\right\}} \leq \max_{1\leq i\leq N}\sum_{j=1}^{N}|r_{ij}|I_{\left\{j\notin S_k^{c_i}\right\}} + b_1(N) \leq b_0(N)k^{-\alpha} + b_1(N).$$

Combining the bounds for terms I, II and III, we obtain

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

Applying the same argument as in Theorem 1, we further obtain

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

This completes the proof. □

B.3 Proof of Corollary 1

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

$$\left\|\widehat{\boldsymbol{\Sigma}}_y - \boldsymbol{\Sigma}_y\right\|_E^2 \leq A\left(\|\boldsymbol{B}\mathbf{D}_T\boldsymbol{B}^\intercal\|_E^2 + \left\|\boldsymbol{B}\widehat{\boldsymbol{\Sigma}}_f\mathbf{C}_T^\intercal\right\|_E^2 + \|\mathbf{C}_T\boldsymbol{\Sigma}_f\mathbf{C}_T^\intercal\|_E^2 + \left\|\widehat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u\right\|_E^2\right), \tag{29}$$

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

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

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

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

which is discussed before for both thresholding and banding estimators.

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

$$P\left(\left\|\widehat{\boldsymbol{\Sigma}}_y - \boldsymbol{\Sigma}_y\right\|_E > A\left(K\frac{\sqrt{N}\log N}{T} + \sqrt{K}\sqrt{\frac{\log N}{T}} + \frac{c_0\left(N\right)\sqrt{\frac{\log N}{T}} + \sqrt{\varrho_T}}{\sqrt{N}}\right)\right)$$
$$= O\left(\frac{1}{N^2} + \kappa_1(N,T) + \sqrt{\varrho_T}\right).$$

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

$$P\left(\left\|\widehat{\mathbf{\Sigma}}_y-\mathbf{\Sigma}_y\right\|_E>A\left(K\frac{\sqrt{N}\log N}{T}+\sqrt{K}\sqrt{\frac{\log N}{T}}+\frac{(1+b_0(N))\left(\frac{\log N}{T}\right)^{\frac{\alpha}{2(\alpha+1)}}+b_1(N)+\sqrt{\varrho_T}}{\sqrt{N}}\right)\right)$$
$$=O\left(\frac{1}{N^2}+\kappa_1(N,T)+\sqrt{\varrho_T}\right).$$

These end the proof. □