UNIVERSITY OF CALIFORNIA
RIVERSIDE

Graphical Models in Financial Econometrics and Macroeconomic Forecasting

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Economics

by

Ekaterina Seregina

June 2021

Dissertation Committee:

Dr. Tae-Hwy Lee, Chairperson
Dr. Jang-Ting Guo
Dr. Jean Helwege
Dr. Aman Ullah

The Dissertation of Ekaterina Seregina is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

This dissertation would not be possible without continuous support of my committee members. I am indebted to my advisor, Dr. Tae-Hwy Lee, for his dedication and guidance; for exciting me about financial econometrics and factor modeling; for always being available to discuss research and give advice; and for setting an example by his hard work. I am grateful to Dr. Jang-Ting Guo for his articulated teaching, helping me grow as a more confident person, and of course for admitting me to the program. I am grateful to Dr. Aman Ullah for a rigorous course in nonparametric econometrics which was one of the reasons that encouraged me to pursue econometrics as a field. I sincerely appreciate invaluable advice and support of Dr. Jean Helwege who helped me see financial econometrics from a non-technical perspective. Finally, I would like to thank Gary Kuzas, his infallible assistance and persistent optimism made my PhD journey smoother.

To my parents for all the support.

ABSTRACT OF THE DISSERTATION

Graphical Models in Financial Econometrics and Macroeconomic Forecasting

by

Ekaterina Seregina

Doctor of Philosophy, Graduate Program in Economics
University of California, Riverside, June 2021
Dr. Tae-Hwy Lee, Chairperson

This dissertation provides theoretical and practical guidance for the use of graphical models, a tool from machine learning and network theory, in financial econometrics and macroeconomic forecasting.

Chapter 1 gives a short introduction to the challenges, methods and findings studied in Chapter 2 to Chapter 4.

Chapter 2 studies a framework to estimate a high-dimensional inverse covariance (precision) matrix for a portfolio allocation problem. I integrate two competing streams of literature, graphical models and factor models, to develop a technique, Factor Graphical Lasso (FGL), that combines the benefits of both aforementioned approaches. I prove consistency of FGL for estimating precision matrix, portfolio weights and risk exposure for three formulations of the optimal portfolio allocation. FGL-based portfolios are shown to exhibit superior performance over several prominent competitors in the empirical application for the S&P500 constituents.

Chapter 3 develops a methodology to construct sparse portfolios in high dimensions. Motivated by a stylized fact that portfolios based on holding all assets fail to generate positive return during economic downturns, I hypothesize that holding sparse portfolios is the key to hedging during recessions. Given unrealistic assumptions imposed by the existing allocation techniques, I develop a strategy for constructing sparse portfolios that could be used as a hedging vehicle during economic downturns. I establish consistency properties of the optimal sparse allocations and provide guidance regarding the distribution of portfolio weights. I also examine the merit of sparse portfolios during different market scenarios and show their robustness to recession periods.

Motivated by the stylized fact that forecasters often use common sets of information and hence they tend to make common mistakes, Chapter 4 proposes a new approach to forecast combinations that separates unique errors from the common errors. I call the proposed algorithm Factor Graphical Model (FGM) and show that it overcomes the challenge of recovering the structure of precision matrix under the factor structure. I prove consistency of forecast combination weights and the Mean Squared Forecast Error estimated using FGM. An empirical application to forecasting macroeconomic series in big data environment demonstrates the merits of FGM.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

This dissertation provides theoretical and practical guidance for the use of graphical models, a tool from machine learning and network theory, in financial econometrics and macroeconomic forecasting.

One can model a system as a network of interactions between its entities. The challenge in such modeling is the presence of unknown underlying structure of the variables within the system. We can make use of the data (system observations) to extract information on the interactions between variables. This is known as network inference or graphical model selection ([134]). Graphical modeling has gained popularity due to the availability of increasing number of variables and observations. Nonetheless, strict sparsity assumptions inherent to structure estimation using graphical models render such approach impractical for many economic problems, including asset allocation and forecast combinations. The goal of this dissertation it to build a bridge between graphical models and latent variable network inference.

Chapter 2 develops a new precision matrix estimator for portfolio allocation problem, Factor Graphical Lasso (FGL), that integrates two competing streams of literature, graphical models and factor models. The root cause why factor models and graphical models are treated separately is the sparsity assumption on the precision matrix made in the latter. However, when asset returns have common factors, the precision matrix cannot be sparse because all pairs of assets are partially correlated conditional on other assets through the common factors ([87]). FGL approach developed in Chapter 2 provides a framework that allows to use graphical models under the factor structure. In addition, I extend the theoretical results of POET ([53]) to allow the number of factors to grow with the number of assets. FGL-based portfolios are shown to consistently estimate precision matrix, portfolio weights and risk exposure. An empirical application uses daily and monthly data for the constituents of the S&P500 to demonstrate that FGL outperforms equal-weighted portfolio, index portfolio and several other prominent covariance and precision estimators.

Chapter 3 develops a methodology to construct sparse portfolios in high dimensions. Constructing non-sparse portfolios in high dimensions has been the main focus of the existing research on asset management for a long time. In particular, many papers focus on developing an improved covariance or precision estimator to achieve desirable statistical properties of portfolio weights. In contrast, the literature on constructing sparse portfolios is scarce: it is limited to a low-dimensional framework and lacks theoretical analysis of the resulting sparse allocations. This chapter fills this gap and proposes a new approach to construct sparse portfolios in high dimensions for three formulations of the optimal portfolio allocation. From the theoretical perspective, I establish consistency of sparse weight estimators and provide guidance regarding their distribution. From the empirical perspective, I examine the merit of sparse portfolios during different market scenarios. I find that in contrast to non-sparse counterparts, my strategy is robust recessions and can be used as a hedging vehicle during such times.

Chapter 4 develops a new approach for estimating optimal forecast combination weights when forecast errors admit approximate factor structure. The latter is motivated by the fact that the forecasters use the same set of public information to make forecasts, hence, they tend to make common mistakes. For example, in the European Central Bank's Survey of Professional forecasters of Euro-area real GDP growth, the forecasters tend to jointly understate or overstate GDP growth. I provide a simple framework, Factor Graphical Model (FGM), to learn from analyzing forecast errors: I separate unique errors from the common errors to improve the accuracy of the combined forecast. I demonstrate that FGM overcomes the challenge of recovering the structure of precision matrix under the factor structure. From the theoretical perspective, I prove consistency of forecast combination weights and the Mean Squared Forecast Error estimated using FGM. An empirical application to forecasting macroeconomic series in big data environment shows that incorporating the factor structure of the forecast errors into the graphical models improves the performance of a combined forecast over forecast combination using equal weights and graphical models without factors.

Chapter 2

Optimal Portfolio Using Factor Graphical Lasso

Abstract

[1] Graphical models are a powerful tool to estimate a high-dimensional inverse covariance (precision) matrix, which has been applied for a portfolio allocation problem. The assumption made by these models is a sparsity of the precision matrix. However, when stock returns are driven by common factors, such assumption does not hold. We address this limitation and develop a framework, Factor Graphical Lasso (FGL), which integrates graphical models with the factor structure in the context

[1] This paper is co-authored with Dr. Tae-Hwy Lee and is circulated under the name "Optimal Portfolio Using Factor Graphical Lasso".

of portfolio allocation by decomposing a precision matrix into low-rank and sparse components. Our theoretical results and simulations show that FGL consistently estimates the portfolio weights and risk exposure and also that FGL is robust to heavy-tailed distributions which makes our method suitable for financial applications. FGL-based portfolios are shown to exhibit superior performance over several prominent competitors including equal-weighted and Index portfolios in the empirical application for the S&P500 constituents.

2.1 Introduction

Estimating the inverse covariance matrix, or *precision* matrix, of excess stock returns is crucial for constructing weights of financial assets in the portfolio and estimating the out-of-sample Sharpe Ratio. In high-dimensional setting, when the number of assets, p, is greater than or equal to the sample size, T, using an estimator of *covariance* matrix for obtaining portfolio weights leads to the Markowitz' curse: a higher number of assets increases correlation between the investments, which calls for a more diversified portfolio, and yet unstable corner solutions for weights become more likely. The reason behind this curse is the need to invert a high-dimensional covariance matrix to obtain the optimal weights from the quadratic optimization problem: when $p \geq T$, the condition number of the covariance matrix (i.e., the absolute value of the ratio between maximal and minimal eigenvalues of the covariance matrix) is high. Hence, the inverted covariance matrix yields an unstable estimator of the precision matrix. To circumvent this issue one can estimate precision matrix directly, rather than inverting covariance matrix.

Graphical models were shown to provide consistent estimates of the precision matrix ([23, 65, 108]). [68] estimated a sparse precision matrix for portfolio hedging using graphical models. They found out that their portfolio achieves significant out-of-sample risk reduction and higher return, as compared to the portfolios based on equal weights, shrunk covariance matrix, industry factor models, and no-short-sale constraints. [4] used Graphical Lasso ([65]) to estimate a sparse covariance matrix for the Markowitz mean-variance portfolio problem to improve covariance estimation in terms of lower realized portfolio risk. [110] conducted an empirical study that applies Graphical Lasso for the estimation of covariance for the portfolio allocation. Their empirical findings suggest that portfolios that use Graphical Lasso for covariance estimation enjoy lower risk and higher returns compared to the empirical covariance matrix. They show that the results are robust to missing observations. [110] also construct a financial network using the estimated precision matrix to explore the relationship between the companies and show how the constructed network helps to make investment decisions. [24] use the nodewise-regression method of [108] to establish consistency of the estimated variance, weights and risk of high-dimensional financial portfolio. Their empirical application demonstrates that the precision matrix estimator based on the nodewise-regression outperforms the principal orthogonal complement thresholding estimator (POET) ([53]) and linear shrinkage ([91]). [20] use constrained ℓ_1-minimization for inverse matrix estimation (Clime) of the precision matrix ([23]) to develop a consistent estimator of the minimum variance for high-dimensional global minimum-variance portfolio. It is important to note that all the aforementioned methods impose some sparsity assumption on the precision matrix of excess returns.

An alternative strategy to handle high-dimensional setting uses factor models to acknowledge common variation in the stock prices, which was documented in many empirical studies (see [26] among many others). A common approach decomposes covariance matrix of excess returns into low-rank and sparse parts, the latter is further regularized since, after the common factors are accounted for, the remaining covariance matrix of the idiosyncratic components is still high-dimensional ([52, 53, 56]). This stream of literature, however, focuses on the estimation of a covariance matrix. The accuracy of precision matrices obtained from inverting the factor-based covariance matrix was investigated by [1], but they did not study a high-dimensional case. *Factor models are generally treated as competitors to graphical models*: as an example, [24] find evidence of superior performance of nodewise-regression estimator of precision matrix over a factor-based estimator POET ([53]) in terms of the out-of-sample Sharpe Ratio and risk of financial portfolio. The root cause why factor models and graphical models are treated separately is the sparsity assumption on the precision matrix made in the latter. Specifically, as pointed out in [87], *when asset returns have common factors, the precision matrix cannot be sparse because all pairs of assets are partially correlated conditional on other assets through the common factors.* One attempt to integrate factor modeling and high-dimensional precision estimation was made by [56] (Section 5.2): the authors referred to such class of models as "conditional graphical models". However, this was not the main focus of their paper which concentrated on covariance estimation through elliptical factor models. As [56] pointed out, "*though substantial amount of efforts have been made to understand the graphical model, little has been done for estimating conditional graphical model, which is more general and realistic*".

Concretely, to the best of our knowledge there are no studies that examine theoretical and empirical performance of graphical models integrated with the factor structure in the context of portfolio allocation.

In this paper we fill this gap and develop a new conditional precision matrix estimator for the excess returns under the approximate factor model that combines the benefits of graphical models and factor structure. We call our algorithm the *Factor Graphical Lasso (FGL)*. We use a factor model to remove the co-movements induced by the factors, and then we apply the Weighted Graphical Lasso for the estimation of the precision matrix of the idiosyncratic terms. We prove consistency of FGL in the spectral and ℓ_1 matrix norms. In addition, we prove consistency of the estimated portfolio weights and risk exposure for three formulations of the optimal portfolio allocation.

Our empirical application uses daily and monthly data for the constituents of the S&P500: we demonstrate that FGL outperforms equal-weighted portfolio, index portfolio, portfolios based on other estimators of precision matrix (Clime, [23]) and covariance matrix, including POET ([53]) and the shrinkage estimators adjusted to allow for the factor structure ([91], [94]), in terms of the out-of-sample Sharpe Ratio. Furthermore, we find strong empirical evidence that relaxing the constraint that portfolio weights sum up to one leads to a large increase in the out-of-sample Sharpe Ratio, which, to the best of our knowledge, has not been previously well-studied in the empirical finance literature.

From the theoretical perspective, our paper makes several important contributions to the existing literature on graphical models and factor models. First, to the best of out knowledge, there are no equivalent theoretical results that establish consistency of the portfolio weights and risk exposure in a high-dimensional setting *without assuming sparsity on the covariance or precision matrix of stock returns.* Second, we extend the theoretical results of POET ([53]) to allow the number of factors to grow with the number of assets. Concretely, we establish uniform consistency for the factors and factor loadings estimated using PCA. Third, we are not aware of any other papers that provide convergence results for estimating a high-dimensional precision matrix using the Weighted Graphical Lasso under the approximate factor model with unobserved factors. Furthermore, all theoretical results established in this paper hold for a wide range of distributions: Sub-Gaussian family (including Gaussian) and elliptical family. Our simulations demonstrate that FGL is robust to very heavy-tailed distributions, which makes our method suitable for the financial applications. Finally, we demonstrate that in contrast to POET, the success of the proposed method does not heavily depend on the factor pervasiveness assumption: FGL is robust to the scenarios when the gap between the diverging and bounded eigenvalues decreases.

This paper is organized as follows: Section 2 reviews the basics of the Markowitz mean-variance portfolio theory. Section 3 provides a brief summary of the graphical models and introduces the Factor Graphical Lasso. Section 4 contains theoretical results and Section 5 validates these results using simulations. Section 6 provides empirical application. Section 7 concludes.

Notation

For the convenience of the reader, we summarize the notation to be used throughout the paper. Let $\mathcal{S}_p$ denote the set of all $p \times p$ symmetric matrices, and $\mathcal{S}_p^{++}$ denotes the set of all $p \times p$ positive definite matrices. For any matrix $\mathbf{C}$, its (i, j)-th element is denoted as

c_{ij}. Given a vector $\mathbf{u} \in \mathbb{R}^d$ and parameter $a \in [1, \infty)$, let $\|\mathbf{u}\|_a$ denote ℓ_a-norm. Given a matrix $\mathbf{U} \in \mathcal{S}_p$, let $\Lambda_{\max}(\mathbf{U}) \equiv \Lambda_1(\mathbf{U}) \geq \Lambda_2(\mathbf{U}) \geq \ldots \geq \Lambda_{\min}(\mathbf{U}) \equiv \Lambda_p(\mathbf{U})$ be the eigenvalues of $\mathbf{U}$, and $\text{eig}_K(\mathbf{U}) \in \mathbb{R}^{K \times p}$ denote the first $K \leq p$ normalized eigenvectors corresponding to $\Lambda_1(\mathbf{U}), \ldots, \Lambda_K(\mathbf{U})$. Given parameters $a, b \in [1, \infty)$, let $|||\mathbf{U}|||_{a,b} \equiv \max_{\|\mathbf{y}\|_a = 1} \|\mathbf{U}\mathbf{y}\|_b$ denote the induced matrix-operator norm. The special cases are $|||\mathbf{U}|||_1 \equiv \max_{1 \leq j \leq N} \sum_{i=1}^{N} |u_{i,j}|$ for the ℓ_1/ℓ_1-operator norm; the operator norm (ℓ_2-matrix norm) $|||\mathbf{U}|||_2^2 \equiv \Lambda_{\max}(\mathbf{U}\mathbf{U}')$ is equal to the maximal singular value of $\mathbf{U}$; $|||\mathbf{U}|||_\infty \equiv \max_{1 \leq j \leq N} \sum_{i=1}^{N} |u_{j,i}|$ for the ℓ_∞/ℓ_∞-operator norm. Finally, $\|\mathbf{U}\|_{\max} \equiv \max_{i,j} |u_{i,j}|$ denotes the element-wise maximum, and $|||\mathbf{U}|||_F^2 \equiv \sum_{i,j} u_{i,j}^2$ denotes the Frobenius matrix norm.

2.2 Optimal Portfolio Allocation

The importance of the minimum-variance portfolio introduced by [105] as a risk-management tool has been studied by many researchers. In this section we review the basics of Markowitz mean-variance portfolio theory and provide several formulations of the optimal portfolio allocation.

Suppose we observe p assets (indexed by i) over T period of time (indexed by t). Let $\mathbf{r}_t = (r_{1t}, r_{2t}, \ldots, r_{pt})' \sim \mathcal{D}(\mathbf{m}, \mathbf{\Sigma})$ be a $p \times 1$ vector of *excess* returns drawn from a distribution $\mathcal{D}$, where $\mathbf{m}$ and $\mathbf{\Sigma}$ are the unconditional mean and covariance matrix of the returns. The goal of the Markowitz theory is to choose asset weights in a portfolio *optimally.* We will study two optimization problems: the well-known Markowitz weight-constrained (MWC) optimization problem, and the Markowitz risk-constrained (MRC) optimization with relaxing the constraint on portfolio weights.

The first optimization problem searches for asset weights such that the portfolio achieves a desired expected rate of return with minimum risk, under the restriction that all weights sum up to one. This can be formulated as the following quadratic optimization problem:

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}'\mathbf{\Sigma}\mathbf{w}, \text{ s.t. } \mathbf{w}'\boldsymbol{\iota} = 1 \text{ and } \mathbf{m}'\mathbf{w} \geq \mu \tag{2.1}$$

where $\mathbf{w}$ is a $p \times 1$ vector of asset weights in the portfolio, $\boldsymbol{\iota}$ is a $p \times 1$ vector of ones, and μ is a desired expected rate of portfolio return. Let $\mathbf{\Theta} \equiv \mathbf{\Sigma}^{-1}$ be the *precision matrix.*

If $\mathbf{m}'\mathbf{w} > \mu$, then the solution to (2.1) yields the *global minimum-variance (GMV) portfolio* weights $\mathbf{w}_{GMV}$:

$$\mathbf{w}_{GMV} = (\boldsymbol{\iota}'\mathbf{\Theta}\boldsymbol{\iota})^{-1}\mathbf{\Theta}\boldsymbol{\iota}. \tag{2.2}$$

If $\mathbf{m}'\mathbf{w} = \mu$, the solution to (2.1) is a well-known two-fund separation theorem introduced by [133]:

$$\mathbf{w}_{MWC} = (1 - a_1)\mathbf{w}_{GMV} + a_1\mathbf{w}_M, \tag{2.3}$$

$$\mathbf{w}_M = (\boldsymbol{\iota}'\mathbf{\Theta}\mathbf{m})^{-1}\mathbf{\Theta}\mathbf{m}, \tag{2.4}$$

$$a_1 = \frac{\mu(\mathbf{m}'\mathbf{\Theta}\boldsymbol{\iota})(\boldsymbol{\iota}'\mathbf{\Theta}\boldsymbol{\iota}) - (\mathbf{m}'\mathbf{\Theta}\boldsymbol{\iota})^2}{(\mathbf{m}'\mathbf{\Theta}\mathbf{m})(\boldsymbol{\iota}'\mathbf{\Theta}\boldsymbol{\iota}) - (\mathbf{m}'\mathbf{\Theta}\boldsymbol{\iota})^2}, \tag{2.5}$$

where $\mathbf{w}_{MWC}$ denotes the portfolio allocation with the constraint that the weights need to sum up to one and $\mathbf{w}_M$ captures all mean-related market information.

The MRC problem has the same objective as in (2.1), but portfolio weights are not required to sum up to one:

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}'\mathbf{\Sigma}\mathbf{w}, \text{ s.t. } \mathbf{m}'\mathbf{w} \geq \mu. \tag{2.6}$$

It can be easily shown that the solution to (2.6) is:

$$\mathbf{w}_1^* = \frac{\mu\mathbf{\Theta}\mathbf{m}}{\mathbf{m}'\mathbf{\Theta}\mathbf{m}}. \tag{2.7}$$

Alternatively, instead of searching for a portfolio with a specified desired expected rate of return, one can maximize expected portfolio return given a maximum risk-tolerance level:

$$\max_{\mathbf{w}} \mathbf{w}'\mathbf{m}, \text{ s.t. } \mathbf{w}'\mathbf{\Sigma}\mathbf{w} \leq \sigma^2. \tag{2.8}$$

In this case, the solution to (2.8) yields:

$$\mathbf{w}_2^* = \frac{\sigma^2}{\mathbf{w}'\mathbf{m}}\mathbf{\Theta}\mathbf{m} = \frac{\sigma^2}{\mu}\mathbf{\Theta}\mathbf{m}. \tag{2.9}$$

To get the second equality in (2.9) we use the definition of μ from (2.6). It follows that if $\mu = \sigma\sqrt{\theta}$, where $\theta \equiv \mathbf{m}'\mathbf{\Theta}\mathbf{m}$ is the squared Sharpe Ratio of the portfolio, then the solution to (2.6) and (2.8) admits the following expression:

$$\mathbf{w}_{MRC} = \frac{\sigma}{\sqrt{\mathbf{m}'\mathbf{\Theta}\mathbf{m}}}\mathbf{\Theta}\mathbf{m} = \frac{\sigma}{\sqrt{\theta}}\boldsymbol{\alpha}, \tag{2.10}$$

where $\boldsymbol{\alpha} \equiv \boldsymbol{\Theta}\mathbf{m}$. Equation (2.10) tells us that once an investor specifies the desired return, μ, and maximum risk-tolerance level, σ, this pins down the Sharpe Ratio of the portfolio which makes the optimization problems of minimizing risk in (2.6) and maximizing expected return of the portfolio in (2.8) identical.

This brings us to three alternative portfolio allocations commonly used in the existing literature: Global Minimum-Variance portfolio in (2.2), Markowitz Weight-Constrained portfolio in (2.3) and Markowitz Maximum-Risk-Constrained portfolio in (2.10). It is clear that all formulations require an estimate of the precision matrix $\boldsymbol{\Theta}$.

2.3 Factor Graphical Lasso

In this section we introduce a framework for estimating precision matrix for the aforementioned financial portfolios which accounts for the fact that the returns follow approximate factor structure.

The arbitrage pricing theory (APT), developed by [119], postulates that the expected returns on securities should be related to their covariance with the common components or factors only. The goal of the APT is to model the tendency of asset returns to move together via factor decomposition. Assume that the return generating process ($\mathbf{r}_t$) follows a K-factor model:

$$\underbrace{\mathbf{r}_t}_{p\times 1} = \mathbf{B} \underbrace{\mathbf{f}_t}_{K\times 1} + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T \tag{2.11}$$

where $\mathbf{f}_t = (f_{1t}, \ldots, f_{Kt})'$ are the factors, $\mathbf{B}$ is a $p \times K$ matrix of factor loadings, and $\boldsymbol{\varepsilon}_t$ is the idiosyncratic component that cannot be explained by the common factors. Factors in (2.11) can be either observable, such as in [46, 47], or can be estimated using statistical factor models. Unobservable factors and loadings are usually estimated by the principal component analysis (PCA), as studied in [5, 6, 35, 126]. Strict factor structure assumes that the idiosyncratic disturbances, $\boldsymbol{\varepsilon}_t$, are uncorrelated with each other, whereas approximate factor structure allows correlation of the idiosyncratic disturbances (see [5, 28] among others).

In this subsection we examine how to solve the Markowitz mean-variance portfolio allocation problems using factor structure in the returns. We also develop *Factor Graphical Lasso* that uses the estimated common factors to obtain a sparse precision matrix of the idiosyncratic component. The resulting estimator is used to obtain the precision of the asset returns necessary to form portfolio weights. In this paper our main interest lies in establishing asymptotic properties of the estimators of precision matrix, portfolio weights and risk-exposure for the high-dimensional case. We assume that the number of common factors, $K = K_{p,T} \to \infty$ as $p \to \infty$, or $T \to \infty$, or both $p, T \to \infty$, but we require that $\max\{K/p, K/T\} \to 0$ as $p, T \to \infty$.

Our setup is similar to the one studied in [53]: we consider a spiked covariance model when the first K principal eigenvalues of $\boldsymbol{\Sigma}$ are growing with p, while the remaining $p - K$ eigenvalues are bounded and grow slower than p.

Rewrite equation (2.11) in matrix form:

$$\underbrace{\mathbf{R}}_{p\times T} = \underbrace{\mathbf{B}}_{p\times K}\mathbf{F} + \mathbf{E}. \tag{2.12}$$

Recall that the factors and loadings in (2.12) are estimated by solving the following minimization problem: $(\widehat{\mathbf{B}},\widehat{\mathbf{F}}) = \text{argmin}_{\mathbf{B},\mathbf{F}}\|\mathbf{R}-\mathbf{B}\mathbf{F}\|_F^2$ s.t. $\frac{1}{T}\mathbf{F}\mathbf{F}' = \mathbf{I}_K$, $\mathbf{B}'\mathbf{B}$ is diagonal. The constraints are needed to identify the factors ([56]). It was shown ([126]) that $\widehat{\mathbf{F}} = \sqrt{T}\text{eig}_K(\mathbf{R}'\mathbf{R})$ and $\widehat{\mathbf{B}} = T^{-1}\mathbf{R}\widehat{\mathbf{F}}'$. Given $\widehat{\mathbf{F}}, \widehat{\mathbf{B}}$, define $\widehat{\mathbf{E}} = \mathbf{R} - \widehat{\mathbf{B}}\widehat{\mathbf{F}}$. Let $\mathbf{\Sigma}_\varepsilon = T^{-1}\mathbf{E}\mathbf{E}'$ and $\mathbf{\Sigma}_f = T^{-1}\mathbf{F}\mathbf{F}'$ be covariance matrices of the idiosyncratic components and factors, and let $\mathbf{\Theta}_\varepsilon = \mathbf{\Sigma}_\varepsilon^{-1}$ and $\mathbf{\Theta}_f = \mathbf{\Sigma}_f^{-1}$ be their inverses. Given a sample of the estimated residuals $\{\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{r}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t\}_{t=1}^T$ and the estimated factors $\{\widehat{\mathbf{f}}_t\}_{t=1}^T$, let $\widehat{\mathbf{\Sigma}}_\varepsilon = (1/T)\sum_{t=1}^T \widehat{\boldsymbol{\varepsilon}}_t\widehat{\boldsymbol{\varepsilon}}_t'$ and $\widehat{\mathbf{\Sigma}}_f = (1/T)\sum_{t=1}^T \widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t'$ be the sample counterparts of the covariance matrices.

Since our interest is in constructing portfolio weights, our goal is to estimate a precision matrix of the excess returns. We impose a sparsity assumption on the precision matrix of the idiosyncratic errors, $\mathbf{\Theta}_\varepsilon$, which is obtained using the estimated residuals after removing the co-movements induced by the factors (see [11, 18, 87]).

Let $\mathbf{W}_\varepsilon$ be an estimate of $\mathbf{\Sigma}_\varepsilon$. Also, let $\widehat{\mathbf{D}}_\varepsilon^2 \equiv \text{diag}(\mathbf{W}_\varepsilon)$. To induce sparsity in the estimation of precision matrix of the idiosyncratic errors $\mathbf{\Theta}_\varepsilon$, we use the following penalized Bregman divergence with the Weighted Graphical Lasso penalty:

$$\widehat{\mathbf{\Theta}}_{\varepsilon,\lambda} = \arg\min_{\mathbf{\Theta}\in\mathcal{S}_p^{++}} \text{trace}(\mathbf{W}_\varepsilon\mathbf{\Theta}_\varepsilon) - \log\det(\mathbf{\Theta}_\varepsilon) + \lambda\sum_{i\neq j}\widehat{d}_{\varepsilon,ii}\widehat{d}_{\varepsilon,jj}|\theta_{\varepsilon,ij}|. \tag{2.13}$$

The subscript λ in $\widehat{\mathbf{\Theta}}_{\varepsilon,\lambda}$ means that the solution of the optimization problem in (2.13) will depend upon the choice of the tuning parameter. More details are provided in Section 4 that establishes sparsity requirements that guarantee convergence of (2.13), and Section 5 that describes how to choose the shrinkage intensity in practice. In order to simplify notation, we will omit the subscript λ. To solve (2.13) we use the procedure based on the weighted Graphical Lasso which was first proposed in [65] and further studied in [106] and [76] among others. Define the following partitions of $\mathbf{W}_\varepsilon$, $\widehat{\mathbf{\Sigma}}_\varepsilon$ and $\mathbf{\Theta}_\varepsilon$:

$$\mathbf{W}_\varepsilon = \begin{pmatrix} \underbrace{\mathbf{W}_{\varepsilon,11}}_{(p-1)\times(p-1)} & \underbrace{\mathbf{w}_{\varepsilon,12}}_{(p-1)\times 1} \\ \mathbf{w}'_{\varepsilon,12} & w_{\varepsilon,22} \end{pmatrix}, \widehat{\mathbf{\Sigma}}_\varepsilon = \begin{pmatrix} \underbrace{\widehat{\mathbf{\Sigma}}_{\varepsilon,11}}_{(p-1)\times(p-1)} & \underbrace{\widehat{\boldsymbol{\sigma}}_{\varepsilon,12}}_{(p-1)\times 1} \\ \widehat{\boldsymbol{\sigma}}'_{\varepsilon,12} & \widehat{\sigma}_{\varepsilon,22} \end{pmatrix}, \mathbf{\Theta}_\varepsilon = \begin{pmatrix} \underbrace{\mathbf{\Theta}_{\varepsilon,11}}_{(p-1)\times(p-1)} & \underbrace{\boldsymbol{\theta}_{\varepsilon,12}}_{(p-1)\times 1} \\ \boldsymbol{\theta}'_{\varepsilon,12} & \theta_{\varepsilon,22} \end{pmatrix}. \tag{2.14}$$

Let $\boldsymbol{\beta} \equiv -\boldsymbol{\theta}_{\varepsilon,12}/\theta_{\varepsilon,22}$. The idea of GLASSO is to set $\mathbf{W}_\varepsilon = \widehat{\mathbf{\Sigma}}_\varepsilon + \lambda\mathbf{I}$ in (2.13) and combine the gradient of (2.13) with the formula for partitioned inverses to obtain the following ℓ_1-regularized quadratic program

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \Big\{ \frac{1}{2}\boldsymbol{\beta}'\mathbf{W}_{\varepsilon,11}\boldsymbol{\beta} - \boldsymbol{\beta}'\widehat{\boldsymbol{\sigma}}_{\varepsilon,12} + \lambda\|\boldsymbol{\beta}\|_1 \Big\}. \tag{2.15}$$

As shown by [65], (2.15) can be viewed as a LASSO regression, where the LASSO estimates are functions of the inner products of $\mathbf{W}_{\varepsilon,11}$ and $\widehat{\sigma}_{\varepsilon,12}$. Hence, (2.13) is equivalent to p coupled LASSO problems. Once we obtain $\widehat{\boldsymbol{\beta}}$, we can estimate the entries $\mathbf{\Theta}_\varepsilon$ using the formula for partitioned inverses. The procedure to obtain sparse $\mathbf{\Theta}_\varepsilon$ is summarized in Algorithm 1.

Algorithm 1: Graphical Lasso [65], adapted

```
Initialize $\mathbf{W}_\varepsilon = \widehat{\boldsymbol{\Sigma}}_\varepsilon + \lambda\mathbf{I}$. The diagonal of $\mathbf{W}_\varepsilon$ remains the same in what follows.
Repeat for $j = 1, \dots, p, 1, \dots, p, \dots$ until convergence:
    • Partition $\mathbf{W}_\varepsilon$ into part 1: all but the $j$-th row and column, and part 2: the $j$-th row and column.
    • Solve the score equations using the cyclical coordinate descent: $\mathbf{W}_{\varepsilon,11}\boldsymbol{\beta} - \widehat{\boldsymbol{\sigma}}_{\varepsilon,12} + \lambda \cdot \text{Sign}(\boldsymbol{\beta}) = \mathbf{0}$. This gives a $(p-1) \times 1$ vector solution $\widehat{\boldsymbol{\beta}}$.
    • Update $\widehat{\mathbf{w}}_{\varepsilon,12} = \mathbf{W}_{\varepsilon,11}\widehat{\boldsymbol{\beta}}$.
In the final cycle (for $i = 1, \dots, p$) solve for $\frac{1}{\widehat{\theta}_{22}} = w_{\varepsilon,22} - \widehat{\boldsymbol{\beta}}'\widehat{\mathbf{w}}_{\varepsilon,12}$ and $\widehat{\boldsymbol{\theta}}_{12} = -\widehat{\theta}_{22}\widehat{\boldsymbol{\beta}}$.
```

As was shown in [65] and the follow-up paper by [106], the estimator produced by Graphical Lasso is guaranteed to be positive definite. Note that the original algorithm developed by [65] is not suitable under the factor structure, therefore, a separate treatment of the statistical properties of the precision matrix estimator in Algorithm 1 is provided in Section 4. Algorithm 1 involves the tuning parameter λ, the procedure on how to choose the shrinkage intensity coefficient is described in more detail in Subsection 5.1.

Having estimated factors, factor loadings and precision matrix of the idiosyncratic components, we combine them using Sherman-Morrison-Woodbury formula to estimate the final precision matrix of excess returns:

$$\widehat{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Theta}}_\varepsilon - \widehat{\boldsymbol{\Theta}}_\varepsilon\widehat{\mathbf{B}}[\widehat{\boldsymbol{\Theta}}_f + \widehat{\mathbf{B}}'\widehat{\boldsymbol{\Theta}}_\varepsilon\widehat{\mathbf{B}}]^{-1}\widehat{\mathbf{B}}'\widehat{\boldsymbol{\Theta}}_\varepsilon. \tag{2.16}$$

We call the procedure described above Factor Graphical Lasso (FGL), and summarize it in Algorithm 2.

```
Algorithm 2: Factor Graphical Lasso
1: (FM) Estimate $\widehat{\mathbf{f}}_t$ and $\widehat{\mathbf{b}}_i$ (Theorem 1). Get $\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{r}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t$, $\widehat{\boldsymbol{\Sigma}}_{\varepsilon}$, $\widehat{\boldsymbol{\Sigma}}_f$ and $\widehat{\boldsymbol{\Theta}}_f = \widehat{\boldsymbol{\Sigma}}_f^{-1}$.
2: (GL) Use Algorithm 1 to get $\widehat{\boldsymbol{\Theta}}_{\varepsilon}$. (Theorem 2)
3: (FGL) Use $\widehat{\boldsymbol{\Theta}}_{\varepsilon}$, $\widehat{\boldsymbol{\Theta}}_f$ and $\widehat{\mathbf{b}}_i$ from Steps 1-2 to get $\widehat{\boldsymbol{\Theta}}$ in Equation (2.16). (Theorem 3)
4: Use $\widehat{\boldsymbol{\Theta}}$ to get $\widehat{\mathbf{w}}_{\xi}$, $\xi \in \{\text{GMV, MWC, MRC}\}$. (Theorem 4)
5: Use $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Theta}}^{-1}$ and $\widehat{\mathbf{w}}_{\xi}$ to get portfolio exposure $\widehat{\mathbf{w}}_{\xi}'\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}}_{\xi}$. (Theorem 5)
```

As we pointed out when discussing Algorithm 1, the estimator produced by Graphical Lasso in general and FGL in particular is guaranteed to be positive definite. We have verified it in the simulations and the empirical application. In Section 4, consistency properties of estimators are established for the factors and loadings (Theorem 1), the precision matrix of $\boldsymbol{\varepsilon}$ (Theorem 2), the precision matrix $\boldsymbol{\Theta}$ (Theorem 3), portfolio weights (Theorem 4), and the portfolio risk exposure (Theorem 5) as indicated in Algorithm 2. We can use $\widehat{\boldsymbol{\Theta}}$ obtained from (2.16) using Step 4 of Algorithm 2 to estimate portfolio weights in (2.2), (2.3) and (2.10):

Remark 1 *In practice, the number of common factors, K, is unknown and needs to be estimated. One of the standard and commonly used approaches is to determine K in a data-driven way ([6, 85]). As an example, in their paper [53] adopt the approach from [6]. However, all of the aforementioned papers deal with a fixed number of factors. Therefore, we need to adopt a different criteria since K is allowed to grow in our setup. For this reason,* *we use the methodology by [99]: let* $\mathbf{b}_{i,K}$ *and* $\mathbf{f}_{t,K}$ *denote* $K \times 1$ *vectors of loadings and factors when* K *needs to be estimated, and* $\mathbf{B}_K$ *is a* $p \times K$ *matrix of stacked* $\mathbf{b}_{i,K}$*. Define*

$$V(K) = \min_{\mathbf{B}_K, \mathbf{F}_K} \frac{1}{pT} \sum_{i=1}^{p} \sum_{t=1}^{T} \Big(r_{it} - \frac{1}{\sqrt{K}} \mathbf{b}'_{i,K} \mathbf{f}_{t,K} \Big)^2, \tag{2.17}$$

where the minimum is taken over $1 \leq K \leq K_{\max}$*, subject to normalization* $\mathbf{B}'_K \mathbf{B}_K / p = \mathbf{I}_K$*. Hence,* $\bar{\mathbf{F}}'_K = \sqrt{K} \mathbf{R}' \mathbf{B}_K / p$*. Define* $\widehat{\mathbf{F}}'_K = \bar{\mathbf{F}}'_K (\bar{\mathbf{F}}_K \bar{\mathbf{F}}'_K / T)^{1/2}$*, which is a rescaled estimator of the factors that is used to determine the number of factors when* K *grows with the sample size. We then apply the following procedure described in [99] to estimate* K*:*

$$\widehat{K} = \arg \min_{1 \leq K \leq K_{\max}} \ln(V(K, \hat{\mathbf{F}}_K)) + K g(p, T), \tag{2.18}$$

where $1 \leq K \leq K_{\max} = o(\min\{p^{1/17}, T^{1/16}\})$ *and* $g(p,T)$ *is a penalty function of* (p,T) *such that (i)* $K_{\max} \cdot g(p,T) \to 0$ *and (ii)* $C^{-1}_{p,T,K_{\max}} \cdot g(p,T) \to \infty$ *with* $C_{p,T,K_{\max}} = \mathcal{O}_P\Big(\max \Big[\frac{K^3_{\max}}{\sqrt{p}}, \frac{K^{5/2}_{\max}}{\sqrt{T}} \Big] \Big)$*. The choice of the penalty function is similar to [6]. Throughout the paper we let* $\widehat{K}$ *be the solution to* (2.18)*.*

2.4 Asymptotic Properties

In this section we first provide a brief review of the terminology used in the literature on graphical models and the approaches to estimate a precision matrix. After that we establish consistency of the Factor Graphical Lasso in Algorithm 2. We also study consistency of the estimators of weights in (2.2), (2.3) and (2.10) and the implications on the out-of sample Sharpe Ratio.

The review of the Gaussian graphical models is based on [71] and [16]. A *graph* consists of a set of *vertices* (nodes) and a set of *edges* (arcs) that join some pairs of the vertices. In graphical models, each vertex represents a random variable, and the graph visualizes the joint distribution of the entire set of random variables. The edges in a graph are parameterized by *potentials* (values) that encode the strength of the conditional dependence between the random variables at the corresponding vertices. *Sparse graphs* have a relatively small number of edges. Among the main challenges in working with the graphical models are choosing the structure of the graph (*model selection*) and estimation of the edge parameters from the data.

Let $A \in \mathcal{S}_p$. Define the following set for $j = 1, \ldots, p$:

$$D_j(A) \equiv \{i : A_{ij} \neq 0,\ i \neq j\}, \quad d_j(A) \equiv \mathrm{card}(D_j(A)), \quad d(A) \equiv \max_{j=1,\ldots,p} d_j(A), \tag{2.19}$$

where $d_j(A)$ is the number of edges adjacent to the vertex j (i.e., the *degree* of vertex j), and $d(A)$ measures the maximum vertex degree. Define $S(A) \equiv \bigcup_{j=1}^{p} D_j(A)$ to be the overall off-diagonal sparsity pattern, and $s(A) \equiv \sum_{j=1}^{p} d_j(A)$ is the overall number of edges contained in the graph. Note that $\mathrm{card}(S(A)) \leq s(A)$: when $s(A) = p(p-1)/2$ this would give a fully connected graph.

2.4.1 Assumptions

We now list the assumptions on the model (2.11):

(A.1) (Spiked covariance model) As $p \to \infty$, $\Lambda_1(\mathbf{\Sigma}) > \Lambda_2(\mathbf{\Sigma}) > \ldots > \Lambda_K(\mathbf{\Sigma}) \gg \Lambda_{K+1}(\mathbf{\Sigma}) \geq \ldots \geq \Lambda_p(\mathbf{\Sigma}) \geq 0$, where $\Lambda_j(\mathbf{\Sigma}) = \mathcal{O}(p)$ for $j \leq K$, while the non-spiked eigenvalues are bounded, $\Lambda_j(\mathbf{\Sigma}) = o(p)$ for $j > K$.

(A.2) (Pervasive factors) There exists a positive definite $K \times K$ matrix $\breve{\mathbf{B}}$ such that $\left|\!\left|\!\left| p^{-1}\mathbf{B}'\mathbf{B} - \breve{\mathbf{B}} \right|\!\right|\!\right|_2 \to 0$ and $\Lambda_{\min}(\breve{\mathbf{B}})^{-1} = \mathcal{O}(1)$ as $p \to \infty$.

(A.3) (a) $\{\boldsymbol{\varepsilon}_t, \mathbf{f}_t\}_{t\geq 1}$ is strictly stationary. Also, $\mathbb{E}[\varepsilon_{it}] = \mathbb{E}[\varepsilon_{it} f_{it}] = 0$ $\forall i \leq p$, $j \leq K$ and $t \leq T$.

(b) There are constants $c_1, c_2 > 0$ such that $\Lambda_{\min}(\mathbf{\Sigma}_\varepsilon) > c_1$, $|||\mathbf{\Sigma}_\varepsilon|||_1 < c_2$ and $\min_{i\leq p, j\leq p} \text{var}(\varepsilon_{it}\varepsilon_{jt}) > c_1$.

(c) There are $r_1, r_2 > 0$ and $b_1, b_2 > 0$ such that for any $s > 0$, $i \leq p$, $j \leq K$,

$$\Pr\left(|\varepsilon_{it}| > s\right) \leq \exp\{-(s/b_1)^{r_1}\},\ \Pr\left(|f_{jt}| > s\right) \leq \exp\{-(s/b_2)^{r_2}\}.$$

We also impose the strong mixing condition. Let $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_T^\infty$ denote the σ-algebras that are generated by $\{(\mathbf{f}_t, \boldsymbol{\varepsilon}_t) : t \leq 0\}$ and $\{(\mathbf{f}_t, \boldsymbol{\varepsilon}_t) : t \geq T\}$ respectively. Define the mixing coefficient

$$\alpha(T) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_T^\infty} |\Pr A \Pr B - \Pr AB|. \tag{2.20}$$

(A.4) (Strong mixing) There exists $r_3 > 0$ such that $3r_1^{-1} + 1.5r_2^{-1} + 3r_3^{-1} > 1$, and $C > 0$ satisfying, for all $T \in \mathbb{Z}^+$, $\alpha(T) \leq \exp(-CT^{r_3})$.

(A.5) (Regularity conditions) There exists $M > 0$ such that, for all $i \leq p$, $t \leq T$ and $s \leq T$, such that:

(a) $\|\mathbf{b}_i\|_{\max} < M$

(b) $\mathbb{E}\left[p^{-1/2}\{\boldsymbol{\varepsilon}_s'\boldsymbol{\varepsilon}_t - \mathbb{E}[\boldsymbol{\varepsilon}_s'\boldsymbol{\varepsilon}_t]\}\right]^4 < M$ and

(c) $\mathbb{E}\left[\left\|p^{-1/2}\sum_{i=1}^{p}\mathbf{b}_i\varepsilon_{it}\right\|^4\right] < K^2M$.

Some comments regarding the aforementioned assumptions are in order. Assumptions **(A.1)**-**(A.4)** are the same as in [53], and assumption **(A.5)** is modified to account for the increasing number of factors. Assumption **(A.1)** divides the eigenvalues into the diverging and bounded ones. Without loss of generality, we assume that K largest eigenvalues have multiplicity of 1. The assumption of a spiked covariance model is common in the literature on approximate factor models. However, we note that the model studied in this paper can be characterized as a "very spiked model". In other words, the gap between the first K eigenvalues and the rest is increasing with p. As pointed out by [56], **(A.1)** is typically satisfied by the factor model with pervasive factors, which brings us to Assumption **(A.2)**: the factors impact a non-vanishing proportion of individual time-series. At the end of section 5 we explore the sensitivity of portfolios constructed using FGL when the pervasiveness assumption is relaxed, that is, when the gap between the diverging and bounded eigenvalues decreases. Assumption **(A.3)**(a) is slightly stronger than in [5], since it requires strict stationarity and non-correlation between $\{\boldsymbol{\varepsilon}_t\}$ and $\{\mathbf{f}_t\}$ to simplify technical calculations. In **(A.3)(b)** we require $\|\!|\boldsymbol{\Sigma}_\varepsilon|\!\|_1 < c_2$ instead of $\lambda_{max}(\boldsymbol{\Sigma}_\varepsilon) = \mathcal{O}(1)$ to estimate

K consistently. When K is known, as in [52,87], this condition can be relaxed. **(A.3)**(c) requires exponential-type tails to apply the large deviation theory to $(1/T)\sum_{t=1}^{T}\varepsilon_{it}\varepsilon_{jt}-\sigma_{u,ij}$ and $(1/T)\sum_{t=1}^{T}f_{jt}u_{it}$. However, in Subsection 4.6 we discuss the extension of our results to the setting with elliptical distribution family which is more appropriate for financial applications. Specifically, we discuss the appropriate modifications to the initial estimator of the covariance matrix of returns such that the bounds derived in this paper continue to hold. **(A.4)-(A.5)** are technical conditions which are needed to consistently estimate the common factors and loadings. The conditions **(A.5)**(a-b) are weaker than those in [5] since our goal is to estimate a precision matrix, and **(A.5)**(c) differs from [5] and [7] in that the number of factors is assumed to slowly grow with p.

In addition, the following structural assumption on the population quantities is imposed:

(B.1) $\|\mathbf{\Sigma}\|_{\max}=\mathcal{O}(1)$, $\|\mathbf{B}\|_{\max}=\mathcal{O}(1)$, and $\|\mathbf{m}\|_{\infty}=\mathcal{O}(1)$.

The sparsity of $\mathbf{\Theta}_{\varepsilon}$ is controlled by the deterministic sequences s_T and d_T: $s(\mathbf{\Theta}_{\varepsilon})=\mathcal{O}_p(s_T)$ for some sequence $s_T\in(0,\infty)$, $T=1,2,\ldots$, and $d(\mathbf{\Theta}_{\varepsilon})=\mathcal{O}_p(d_T)$ for some sequence $d_T\in(0,\infty)$, $T=1,2,\ldots$. We will impose restrictions on the growth rates of s_T and d_T. Note that assumptions on d_T are weaker since they are always satisfied when $s_T=d_T$. However, d_T can generally be smaller than s_T. In contrast to [53] we do not impose sparsity on the covariance matrix of the idiosyncratic component, instead, it is more realistic and relevant for error quantification in portfolio analysis to impose conditional sparsity on the precision matrix after the common factors are accounted for.

2.4.2 The FGL Procedure

Recall the definition of the Weighted Graphical Lasso estimator in (2.13) for the precision matrix of the idiosyncratic components. Also, recall that to estimate $\mathbf{\Theta}$ we used equation (2.16). Therefore, in order to obtain the FGL estimator $\widehat{\mathbf{\Theta}}$ we take the following steps: **(1):** estimate unknown factors and factor loadings to get an estimator of $\mathbf{\Sigma}_\varepsilon$. **(2):** use $\widehat{\mathbf{\Sigma}}_\varepsilon$ to get an estimator of $\mathbf{\Theta}_\varepsilon$ in (2.13). **(3):** use $\widehat{\mathbf{\Theta}}_\varepsilon$ together with the estimators of factors and factor loadings from Step 1 to obtain the final precision matrix estimator $\widehat{\mathbf{\Theta}}$, portfolio weight estimator $\widehat{\mathbf{w}}_\xi$, and risk exposure estimator $\widehat{\Phi}_\xi = \widehat{\mathbf{w}}_\xi' \widehat{\mathbf{\Theta}}^{-1} \widehat{\mathbf{w}}_\xi$ where $\xi \in \{\text{GMV, MWC, MRC}\}$.

Subsection 4.3 examines the theoretical foundations of the first step, and Subsections 4.4-4.5 are devoted to steps 2 and 3.

2.4.3 Convergence of Unknown Factors and Loadings

As pointed out in [5] and [53], $K \times 1$-dimensional factor loadings $\{\mathbf{b}_i\}_{i=1}^p$, which are the rows of the factor loadings matrix $\mathbf{B}$, and $K \times 1$-dimensional common factors $\{\mathbf{f}_t\}_{t=1}^T$, which are the columns of $\mathbf{F}$, are not separately identifiable. Concretely, for any $K \times K$ matrix $\mathbf{H}$ such that $\mathbf{H}'\mathbf{H} = \mathbf{I}_K$, $\mathbf{B}\mathbf{f}_t = \mathbf{B}\mathbf{H}'\mathbf{H}\mathbf{f}_t$, therefore, we cannot identify the tuple $(\mathbf{B}, \mathbf{f}_t)$ from $(\mathbf{B}\mathbf{H}', \mathbf{H}\mathbf{f}_t)$. Let $\widehat{K} \in \{1, \ldots, K_{\max}\}$ denote the estimated number of factors, where $K_{\max}$ is allowed to increase at a slower speed than $\min\{p, T\}$ such that $K_{\max} = o(\min\{p^{1/3}, T\})$ (see [99] for the discussion about the rate).

Define $\mathbf{V}$ to be a $\widehat{K} \times \widehat{K}$ diagonal matrix of the first $\widehat{K}$ largest eigenvalues of the sample covariance matrix in decreasing order. Further, define a $\widehat{K} \times \widehat{K}$ matrix $\mathbf{H} =$

$(1/T)\mathbf{V}^{-1}\widehat{\mathbf{F}}'\mathbf{F}\mathbf{B}'\mathbf{B}$. For $t \leq T$, $\mathbf{Hf}_t = T^{-1}\mathbf{V}^{-1}\widehat{\mathbf{F}}'(\mathbf{Bf}_1, \ldots, \mathbf{Bf}_T)'\mathbf{Bf}_t$, which depends only on the data $\mathbf{V}^{-1}\widehat{\mathbf{F}}'$ and an identifiable part of parameters $\{\mathbf{Bf}_t\}_{t=1}^{T}$. Hence, $\mathbf{Hf}_t$ does not have an identifiability problem regardless of the imposed identifiability condition.

Let $\gamma^{-1} = 3r_1^{-1} + 1.5r_2^{-1} + r_3^{-1} + 1$. The following theorem is an extension of the results in [53] for the case when the number of factors is unknown and is allowed to grow. Proofs of all the theorems are in section 2.A.

Theorem 1 *Suppose that* $K_{\max} = o(\min\{p^{1/3}, T\})$, $K^3 \log p = o(T^{\gamma/6})$, $KT = o(p^2)$ *and Assumptions **(A.1)-(A.5)** and **(B.1)** hold. Let* $\omega_{1T} \equiv K^{3/2}\sqrt{\log p/T} + K/\sqrt{p}$ *and* $\omega_{2T} \equiv K/\sqrt{T} + KT^{1/4}/\sqrt{p}$. *Then* $\max_{i\leq p}\left\|\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\right\| = \mathcal{O}_P(\omega_{1T})$ *and* $\max_{t\leq T}\left\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\right\| = \mathcal{O}_P(\omega_{2T})$.

The conditions $K^3 \log p = o(T^{\gamma/6})$, $KT = o(p^2)$ are similar to [53], the difference arises due to the fact that we do not fix K, hence, in addition to the factor loadings, there are KT factors to estimate. Therefore, the number of parameters introduced by the unknown growing factors should not be "too large", such that we can consistently estimate them uniformly. The growth rate of the number of factors is controlled by $K_{\max} = o(\min\{p^{1/3}, T\})$.

The bounds derived in Theorem 1 help us establish the convergence properties of the estimated idiosyncratic covariance, $\widehat{\mathbf{\Sigma}}_\varepsilon$, and precision matrix $\widehat{\mathbf{\Theta}}_\varepsilon$ which are presented in the next theorem:

Theorem 2 *Let* $\omega_{3T} \equiv K^2\sqrt{\log p/T} + K^3/\sqrt{p}$. *Under the assumptions of Theorem 1 and with* $\lambda \asymp \omega_{3T}$ *(where* λ *is the tuning parameter in* (2.13)*), the estimator* $\widehat{\mathbf{\Sigma}}_\varepsilon$ *obtained by estimating factor model in* (2.12) *satisfies* $\left\|\widehat{\mathbf{\Sigma}}_\varepsilon - \mathbf{\Sigma}_\varepsilon\right\|_{\max} = \mathcal{O}_P(\omega_{3T})$. *Let* ϱ_T *be a sequence of positive-valued random variables such that* $\varrho_T^{-1}\omega_{3T} \xrightarrow{p} 0$. *If* $s_T\varrho_T \xrightarrow{p} 0$, *then* $\left|\left|\left|\widehat{\mathbf{\Theta}}_\varepsilon - \mathbf{\Theta}_\varepsilon\right|\right|\right|_l = \mathcal{O}_P(\varrho_T s_T)$ *as* $T \to \infty$ *for any* $l \in [1, \infty]$.

Note that the term containing $K^3/\sqrt{p}$ arises due to the need to estimate unknown factors: [52] obtained a similar rate but for the case when factors are observable (in their work, $\omega_{3T} = K^{1/2}\sqrt{\log p/T}$). The second part of Theorem 2 is based on the relationship between the convergence rates of the estimated covariance and precision matrices established in [76] (Theorem 14.1.3). [87] obtained the convergence rate when factors are observable: the rate obtained in our paper is slower due to the fact that factors need to be estimated (concretely, the rate under observable factors would satisfy $\varrho_T^{-1}\sqrt{K\log p/T} \xrightarrow{p} 0$). We now comment on the optimality of the rate in Theorem 2: as pointed out in [87], in the standard Gaussian setting without factor structure, the minimax optimal rate is $d(\mathbf{\Theta}_\varepsilon)\sqrt{\log p/T}$, which can be faster than the rate obtained in Theorem 2 if $d(\mathbf{\Theta}_\varepsilon) < s_T$. Using penalized nodewise regression could help achieve this faster rate. However, our empirical application to the monthly stock returns demonstrated superior performance of the Weighted Graphical Lasso compared to the nodewise regression in terms of the out-of-sample Sharpe Ratio and portfolio risk. Hence, in order not to divert the focus of this paper, we leave the theoretical properties of the nodewise regression for future research.

2.4.4 Convergence of Precision Matrix Estimator and Portfolio Weights by FGL

Having established the convergence properties of $\widehat{\mathbf{\Sigma}}_\varepsilon$ and $\widehat{\mathbf{\Theta}}_\varepsilon$, we now move to the estimation of the precision matrix of the factor-adjusted returns in equation (2.16).

Theorem 3 *Under the assumptions of Theorem 2, if* $d_T s_T \varrho_T \xrightarrow{p} 0$, *then* $\left|\left|\left|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\right|\right|\right|_2 = \mathcal{O}_P(\varrho_T s_T)$ *and* $\left|\left|\left|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\right|\right|\right|_1 = \mathcal{O}_P(\varrho_T d_T K^{3/2} s_T)$.

Note that since, by construction, the precision matrix obtained using the Factor Graphical Lasso is symmetric, $\left|\left|\left|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\right|\right|\right|_\infty$ can be trivially obtained from the above theorem.

Using Theorem 3, we can then establish the consistency of the estimated weights of portfolios based on the Factor Graphical Lasso.

Theorem 4 *Under the assumptions of Theorem 3, we additionally assume* $|||\mathbf{\Theta}|||_2 = \mathcal{O}(1)$ *(this additional requirement essentially imposes* $\Lambda_p(\mathbf{\Sigma}) > 0$ *in **(A.1)**), and* $\varrho_T d_T^2 s_T = o(1)$. *Algorithm 2 consistently estimates portfolio weights in* (2.2), (2.3) *and* (2.10)*:* $\|\widehat{\mathbf{w}}_{GMV} - \mathbf{w}_{GMV}\|_1 = \mathcal{O}_P\left(\varrho_T d_T^2 K^3 s_T\right) = o_P(1)$, $\|\widehat{\mathbf{w}}_{MWC} - \mathbf{w}_{MWC}\|_1 = \mathcal{O}_P(\varrho_T d_T^2 K^3 s_T) = o_P(1)$, *and* $\|\widehat{\mathbf{w}}_{MRC} - \mathbf{w}_{MRC}\|_1 = \mathcal{O}_P\left(d_T^{3/2} K^3 \cdot [\varrho_T s_T]^{1/2}\right) = o_P(1)$.

We now comment on the rates in Theorem 4: first, the rates obtained by [24] for GMV and MWC formulations, when no factor structure of stock returns is assumed, require $s(\mathbf{\Theta})^{3/2}\sqrt{\log p/T} = o_P(1)$, where the authors imposed sparsity on the precision matrix of stock returns, $\mathbf{\Theta}$. Therefore, if the precision matrix of stock returns is not sparse, portfolio weights can be consistently estimated only if p is less than $T^{1/3}$ (since $(p-1)^{3/2}\sqrt{\log p/T} = o(1)$ is required to ensure consistent estimation of portfolio weights). Our result in Theorem

4 improves this rate and shows that as long as $d_T^2 s_T K^3 \sqrt{\log p/T} = o_P(1)$ we can consistently estimate weights of the financial portfolio. Specifically, when the precision of the factor-adjusted returns is sparse, we can consistently estimate portfolio weights when $p > T$ *without* assuming sparsity on $\mathbf{\Sigma}$ or $\mathbf{\Theta}$. Second, note that GMV and MWC weights converge slightly slower than MRC weight. This result is further supported by our simulations presented in the next section.

2.4.5 Implications on Portfolio Risk Exposure

Having examined the properties of portfolio weights, it is natural to comment on the portfolio variance estimation error. It is determined by the errors in two components: the estimated covariance matrix and the estimated portfolio weights. Define $a = \boldsymbol{\iota}_p'\mathbf{\Theta}\boldsymbol{\iota}_p/p$, $b = \boldsymbol{\iota}_p'\mathbf{\Theta}\mathbf{m}/p$, $d = \mathbf{m}'\mathbf{\Theta}\mathbf{m}/p$, $g = \sqrt{\mathbf{m}'\mathbf{\Theta}\mathbf{m}}/p$ and $\widehat{a} = \boldsymbol{\iota}_p'\widehat{\mathbf{\Theta}}\boldsymbol{\iota}_p/p$, $\widehat{b} = \boldsymbol{\iota}_p'\widehat{\mathbf{\Theta}}\widehat{\mathbf{m}}/p$, $\widehat{d} = \widehat{\mathbf{m}}'\widehat{\mathbf{\Theta}}\widehat{\mathbf{m}}/p$, $\widehat{g} = \sqrt{\widehat{\mathbf{m}}'\widehat{\mathbf{\Theta}}\widehat{\mathbf{m}}}/p$. Define $\Phi_{\text{GMV}} = \mathbf{w}_{GMV}'\mathbf{\Sigma}\mathbf{w}_{GMV} = (pa)^{-1}$ to be the global minimum variance, $\Phi_{\text{MWC}} = \mathbf{w}_{MWC}'\mathbf{\Sigma}\mathbf{w}_{MWC} = p^{-1}\left[\frac{a\mu^2-2b\mu+d}{ad-b^2}\right]$ is the MWC portfolio variance, and $\Phi_{\text{MRC}} = \mathbf{w}_{MRC}'\mathbf{\Sigma}\mathbf{w}_{MRC} = \sigma^2(pg)$ is the MRC portfolio variance. We use the terms variance and risk exposure interchangeably. Let $\widehat{\Phi}_{\text{GMV}}$, $\widehat{\Phi}_{\text{MWC}}$, and $\widehat{\Phi}_{\text{MRC}}$ be the sample counterparts of the respective portfolio variances. The expressions for Φ_{GMV} and Φ_{MWC} were derived in [48] and [24]. Theorem 5 establishes the consistency of a large portfolio's variance estimator.

Theorem 5 *Under the assumptions of Theorem 3, FGL consistently estimates GMV, MWC, and MRC portfolio variance:*

$$\left|\widehat{\Phi}_{GMV}/\Phi_{GMV}-1\right| = \mathcal{O}_P(\varrho_T d_T s_T K^{3/2}) = o_P(1),$$
$$\left|\widehat{\Phi}_{MWC}/\Phi_{MWC}-1\right| = \mathcal{O}_P(\varrho_T d_T s_T K^{3/2}) = o_P(1),$$
$$\left|\widehat{\Phi}_{MRC}/\Phi_{MRC}-1\right| = \mathcal{O}_P\Big([\varrho_T d_T s_T K^{3/2}]^{1/2}\Big) = o_P(1).$$

[24] derived a similar result for Φ_{GMV} and Φ_{MWC} under the assumption that precision matrix of stock returns is sparse. Also, [43] derived the bounds for Φ_{GMV} under the factor structure assuming sparse covariance matrix of idiosyncratic components and gross exposure constraint on portfolio weights which limits negative positions.

The empirical application in Section 6 reveals that the portfolios constructed using MRC formulation have higher risk compared with GMV and MWC alternatives: using monthly and daily returns of the components of S&P500 index, MRC portfolios exhibit higher out-of-sample risk and return compared to the alternative formulations. Furthermore, the empirical exercise demonstrates that the higher return of MRC portfolios outweighs higher risk for the monthly data which is evidenced by the increased out-of-sample Sharpe Ratio.

2.4.6 Generalization: Sub-Gaussian and Elliptical Distributions

So far the consistency of the Factor Graphical Lasso in Theorem 4 relied on the assumption of the exponential-type tails in **(A.3)**(c). Since this tail-behavior may be too restrictive for financial portfolio, we comment on the possibility to relax it. First, recall where **(A.3)**(c) was used before: we required this assumption in order to establish convergence of unknown factors and loadings in Theorem 1, which was further used to obtain the convergence properties of $\widehat{\mathbf{\Sigma}}_{\varepsilon}$ in Theorem 2. Hence, when Assumption **(A.3)**(c) is relaxed, one needs to find another way to consistently estimate $\mathbf{\Sigma}_{\varepsilon}$. We achieve it using the tools developed in [56]. Specifically, let $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^{'}$, where $\mathbf{\Sigma}$ is the covariance matrix of returns that follow a factor structure described in equation (2.11). Define $\widehat{\mathbf{\Sigma}}, \widehat{\mathbf{\Lambda}}_K, \widehat{\mathbf{\Gamma}}_K$ to be the estimators of $\mathbf{\Sigma}, \mathbf{\Lambda}, \mathbf{\Gamma}$. We further let $\widehat{\mathbf{\Lambda}}_K = \mathrm{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_K)$ and $\widehat{\mathbf{\Gamma}}_K = (\hat{v}_1, \dots, \hat{v}_K)$ to be constructed by the first K leading empirical eigenvalues and the corresponding eigenvectors of $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{B}}\widehat{\mathbf{B}}' = \widehat{\mathbf{\Gamma}}_K \widehat{\mathbf{\Lambda}}_K \widehat{\mathbf{\Gamma}}_K^{'}$. Similarly to [56], we require the following bounds on the componentwise maximums of the estimators:

(C.1) $\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_{\max} = \mathcal{O}_P(\sqrt{\log p/T})$,

(C.2) $\left\|(\widehat{\mathbf{\Lambda}}_K - \mathbf{\Lambda})\mathbf{\Lambda}^{-1}\right\|_{\max} = \mathcal{O}_P(K\sqrt{\log p/T})$,

(C.3) $\left\|\widehat{\mathbf{\Gamma}}_K - \mathbf{\Gamma}\right\|_{\max} = \mathcal{O}_P(K^{1/2}\sqrt{\log p/(Tp)})$.

Let $\widehat{\mathbf{\Sigma}}^{SG}$ be the sample covariance matrix, with $\widehat{\mathbf{\Lambda}}_K^{SG}$ and $\widehat{\mathbf{\Gamma}}_K^{SG}$ constructed with the first K leading empirical eigenvalues and eigenvectors of $\widehat{\mathbf{\Sigma}}^{SG}$ respectively. Also, let $\widehat{\mathbf{\Sigma}}^{EL1} = \widehat{\mathbf{D}}\widehat{\mathbf{R}}_1\widehat{\mathbf{D}}$, where $\widehat{\mathbf{R}}_1$ is obtained using the Kendall's tau correlation coefficients and $\widehat{\mathbf{D}}$ is a robust estimator of variances constructed using the Huber loss. Furthermore, let $\widehat{\mathbf{\Sigma}}^{EL2} = \widehat{\mathbf{D}}\widehat{\mathbf{R}}_2\widehat{\mathbf{D}}$, where $\widehat{\mathbf{R}}_2$ is obtained using the spatial Kendall's tau estimator. Define $\widehat{\mathbf{\Lambda}}_K^{EL}$ to be the matrix of the first K leading empirical eigenvalues of $\widehat{\mathbf{\Sigma}}^{EL1}$, and $\widehat{\mathbf{\Gamma}}_K^{EL}$ is the matrix of the first K leading empirical eigenvectors of $\widehat{\mathbf{\Sigma}}^{EL2}$. For more details regarding constructing $\widehat{\mathbf{\Sigma}}^{SG}$, $\widehat{\mathbf{\Sigma}}^{EL1}$ and $\widehat{\mathbf{\Sigma}}^{EL2}$ see [56], Sections 3 and 4.

Proposition 1 *For sub-Gaussian distributions,* $\widehat{\mathbf{\Sigma}}^{SG}$*,* $\widehat{\mathbf{\Lambda}}_K^{SG}$ *and* $\widehat{\mathbf{\Gamma}}_K^{SG}$ *satisfy* ***(C.1)-(C.3)****. For elliptical distributions,* $\widehat{\mathbf{\Sigma}}^{EL1}$*,* $\widehat{\mathbf{\Lambda}}_K^{EL}$ *and* $\widehat{\mathbf{\Gamma}}_K^{EL}$ *satisfy* ***(C.1)-(C.3)****.*

When ***(C.1)-(C.3)*** *are satisfied, the bounds obtained in Theorems 2-5 continue to hold.*

Proposition 1 is essentially a rephrasing of the results obtained in [56], Sections 3 and 4. The difference arises due to the fact that we allow K to increase, which is reflected in the modified rates in **(C.2)-(C.3)**. As evidenced from the above Proposition, $\widehat{\mathbf{\Sigma}}^{EL2}$ is only used for estimating the eigenvectors. This is necessary due to the fact that, in contrast with $\widehat{\mathbf{\Sigma}}^{EL2}$, the theoretical properties of the eigenvectors of $\widehat{\mathbf{\Sigma}}^{EL}$ are mathematically involved because of the sin function. The FGL for the elliptical distributions will be called the *Robust FGL.*

2.5 Monte Carlo

In order to validate our theoretical results, we perform several simulation studies which are divided into four parts. The first set of results computes the empirical convergence rates and compares them with the theoretical expressions derived in Theorems 3-5. The second set of results compares the performance of the FGL with several alternative models for estimating covariance and precision matrix. To highlight the benefit of using the information about factor structure as opposed to standard graphical models, we include Graphical Lasso by [65] (GL) that does not account for the factor structure. To explore the benefits of using FGL for error quantification in (2.16), we consider several alternative estimators of covariance/precision matrix of the idiosyncratic component in (2.16): (1) linear shrinkage estimator of covariance developed by [91] further referred to as Factor LW or

FLW; (2) nonlinear shrinkage estimator of covariance by [94] (Factor NLW or FNLW); (3) POET ([53]); (4) constrained ℓ_1-minimization for inverse matrix estimator, Clime ([23]) (Factor Clime or FClime). Furthermore, we discovered that in certain setups the estimator of covariance produced by POET is not positive definite. In such cases we use the matrix symmetrization procedure as in [56] and then use eigenvalue cleaning as in [25] and [72]. This estimator is referred to as Projected POET; it coincides with POET when the covariance estimator produced by the latter is positive definite. The third set of results examines the performance of FGL and Robust FGL (described in Subsection 4.6) when the dependent variable follows elliptical distribution. The fourth set of results explores the sensitivity of portfolios constructed using different covariance and precision estimators of interest when the pervasiveness assumption **(A.2)** is relaxed, that is, when the gap between the diverging and bounded eigenvalues decreases. All exercises in this section use 100 Monte Carlo simulations.

We first discuss the choice of the tuning parameter λ in (2.13) used in Algorithm 1. Let $\widehat{\mathbf{\Theta}}_{\varepsilon,\lambda}$ be the solution to (2.13) for a fixed λ. Following [87], we minimize the following Bayesian Information Criterion (BIC) using grid search:

$$\text{BIC}(\lambda) \equiv T\Big[\text{trace}(\widehat{\mathbf{\Theta}}_{\varepsilon,\lambda}\widehat{\mathbf{\Sigma}}_{\varepsilon}) - \log\det(\widehat{\mathbf{\Theta}}_{\varepsilon,\lambda})\Big] + (\log T)\sum_{i\leq j} \mathbb{1}\left[\widehat{\theta}_{\varepsilon,\lambda,ij} \neq 0\right]. \tag{2.21}$$

The grid $\mathcal{G} \equiv \{\lambda_1, \ldots, \lambda_m\}$ is constructed as follows: the maximum value in the grid, λ_m, is set to be the smallest value for which all the off-diagonal entries of $\widehat{\mathbf{\Theta}}_{\varepsilon,\lambda_m}$ are zero, that is, the maximum modulus of the off-diagonal entries of $\widehat{\mathbf{\Sigma}}_{\varepsilon}$. The smallest value of the grid, $\lambda_1 \in \mathcal{G}$, is determined as $\lambda_1 \equiv \vartheta\lambda_m$ for a constant $\vartheta > 0$. The remaining grid values

$\lambda_1, \dots, \lambda_m$ are constructed in the ascending order from λ_1 to λ_m on the log scale:

$$\lambda_i = \exp\Big(\log(\lambda_1) + \frac{i-1}{m-1}\log(\lambda_m/\lambda_1)\Big), \quad i = 2, \dots, m-1.$$

We use $\vartheta = \omega_{3T}$ and $m = 10$ in the simulations and the empirical exercise. We consider the following setup: let $p = T^{\delta}$, $\delta = 0.85$, $K = 2(\log T)^{0.5}$ and $T = [2^h]$, for $h = 7, 7.5, 8, \dots, 9.5$. A sparse precision matrix of the idiosyncratic components is constructed as follows: we first generate the adjacency matrix using a random graph structure. Define a $p \times p$ adjacency matrix $\mathbf{A}_\varepsilon$ which is used to represent the structure of the graph:

$$a_{\varepsilon,ij} = \begin{cases} 1, & \text{for } i \neq j \;\text{ with probability } q, \\ 0, & \text{otherwise.} \end{cases} \tag{2.22}$$

Let $a_{\varepsilon,ij}$ denote the i,j-th element of the adjacency matrix $\mathbf{A}_\varepsilon$. We set $a_{\varepsilon,ij} = a_{\varepsilon,ji} = 1$, for $i \neq j$ with probability q, and 0 otherwise. Such structure results in $s_T = p(p-1)q/2$ edges in the graph. To control sparsity, we set $q = 1/(pT^{0.8})$, which makes $s_T = \mathcal{O}(T^{0.05})$. The adjacency matrix has all diagonal elements equal to zero. Hence, to obtain a positive definite precision matrix we apply the procedure described in [142]: using their notation, $\mathbf{\Theta}_\varepsilon = \mathbf{A}_\varepsilon \cdot v + \mathbf{I}(|\tau| + 0.1 + u)$, where $u > 0$ is a positive number added to the diagonal of the precision matrix to control the magnitude of partial correlations, v controls the magnitude of partial correlations with u, and τ is the smallest eigenvalue of $\mathbf{A}_\varepsilon \cdot v$. In our simulations we use $u = 0.1$ and $v = 0.3$.

Factors are assumed to have the following structure:

$$\mathbf{f}_t = \phi_f \mathbf{f}_{t-1} + \boldsymbol{\zeta}_t \tag{2.23}$$

$$\underbrace{\mathbf{r}_t}_{p\times 1} = \mathbf{B} \underbrace{\mathbf{f}_t}_{K\times 1} + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T \tag{2.24}$$

where $\boldsymbol{\varepsilon}_t$ is a $p \times 1$ random vector of idiosyncratic errors following $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$, with sparse $\boldsymbol{\Theta}_\varepsilon$ that has a random graph structure described above, $\mathbf{f}_t$ is a $K \times 1$ vector of factors, ϕ_f is an autoregressive parameter in the factors which is a scalar for simplicity, $\mathbf{B}$ is a $p \times K$ matrix of factor loadings, $\boldsymbol{\zeta}_t$ is a $K \times 1$ random vector with each component independently following $\mathcal{N}(0, \sigma_\zeta^2)$. To create $\mathbf{B}$ in (2.24) we take the first K rows of an upper triangular matrix from a Cholesky decomposition of the $p \times p$ Toeplitz matrix parameterized by ρ. For the first set of results we set $\rho = 0.2$, $\phi_f = 0.2$ and $\sigma_\zeta^2 = 1$. The specification in (2.24) leads to the low-rank plus sparse decomposition of the covariance matrix of stock returns $\mathbf{r}_t$.

As a first exercise, we compare the empirical and theoretical convergence rates of the precision matrix, portfolio weights and exposure. A detailed description of the procedure and the simulation results is provided in Appendix 2.B.1. We confirm that the empirical rates and theoretical rates from Theorems 3-5 are matched.

As a second exercise, we compare the performance of FGL with the alternative models listed at the beginning of this section. We consider two cases: **Case 1** is the same as for the first set of simulations ($p < T$): $p = T^{\delta}$, $\delta = 0.85$, $K = 2(\log T)^{0.5}$, $s_T = \mathcal{O}(T^{0.05})$. **Case 2** captures the cases when $p > T$ with $p = 3 \cdot T^{\delta}$, $\delta = 0.85$, all else equal. The results for Case 2 are reported in Figure 2.1-2.3, and Case 1 is located in Appendix 2.B.2. FGL demonstrates superior performance for estimating precision matrix and portfolio weights in both cases, exhibiting consistency for both Case 1 and Case 2 settings. Also, FGL outperforms GL for estimating portfolio exposure and consistently estimates the latter, however, depending on the case under consideration some alternative models produce lower averaged error.

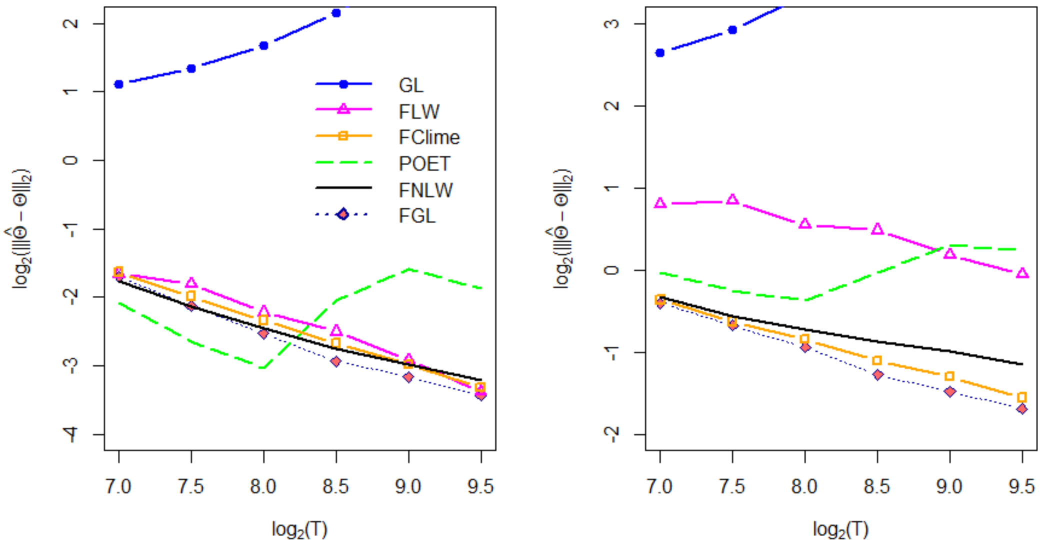

Figure 2.1: **Averaged errors of the estimators of $\mathbf{\Theta}$ for Case 2 on logarithmic scale:** $p = 3 \cdot T^{0.85}$**,** $K = 2(\log T)^{0.5}$**,** $s_T = \mathcal{O}(T^{0.05})$**.**

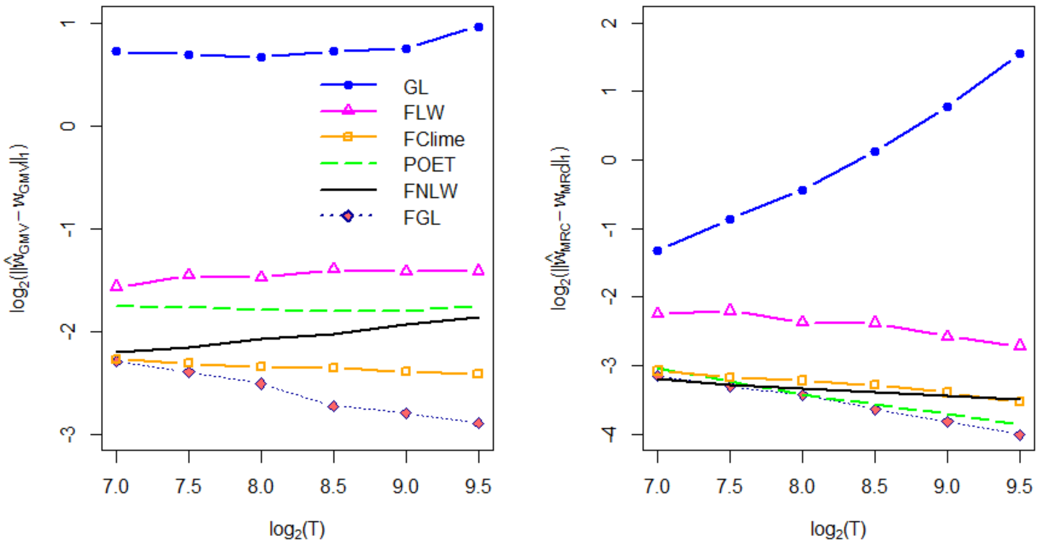

Figure 2.2: **Averaged errors of the estimators of $\mathbf{w}_{\mathbf{GMV}}$ (left) and $\mathbf{w}_{\mathbf{MRC}}$ (right) for Case 2 on logarithmic scale: $p = 3 \cdot T^{0.85}$, $K = 2(\log T)^{0.5}$, $s_T = \mathcal{O}(T^{0.05})$.**

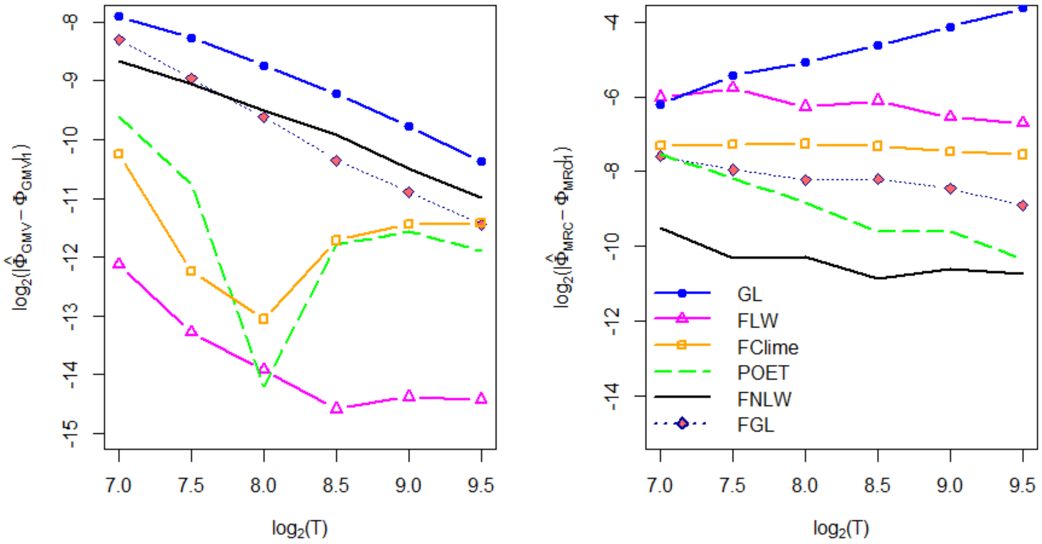

Figure 2.3: **Averaged errors of the estimators of $\Phi_{\mathbf{GMV}}$ (left) and $\Phi_{\mathbf{MRC}}$ (right) for Case 2 on logarithmic scale: $p = 3 \cdot T^{0.85}$, $K = 2(\log T)^{0.5}$, $s_T = \mathcal{O}(T^{0.05})$.**

As a third exercise, we examine the performance of FGL and Robust FGL (described in subsection 4.6) when the dependent variable follows elliptical distributions. A detailed description of the data generating process (DGP) and simulation results are provided in Appendix 2.B.3. We find that the performance of FGL for estimating the precision matrix is comparable with that of Robust FGL: this suggests that our FGL algorithm is robust to heavy-tailed distributions even without additional modifications.

As a final exercise, we explore the sensitivity of portfolios constructed using different covariance and precision estimators of interest when the pervasiveness assumption **(A.2)** is relaxed. A detailed description of the data generating process (DGP) and simulation results are provided in Appendix 2.B.4. We verify that FGL exhibits robust performance when the gap between the diverging and bounded eigenvalues decreases. In contrast, POET and Projected POET are most sensitive to relaxing pervasiveness assumption which is consistent with our empirical findings and also with the simulation results by [113].

2.6 Empirical Application

In this section we examine the performance of the Factor Graphical Lasso for constructing a financial portfolio using daily data. The description and empirical results for monthly data can be found in Appendix 2.C. We first describe the data and the estimation methodology, then we list four metrics commonly reported in the finance literature, and, finally, we present the results.

2.6.1 Data

We use daily returns of the components of the S&P500 index. The data on historical S&P500 constituents and stock returns is fetched from CRSP and Compustat using SAS interface. For the daily data the full sample size has 5040 observations on 420 stocks from January 20, 2000 - January 31, 2020. We use January 20, 2000 - January 24, 2002 (504 obs) as the first training (estimation) period and January 25, 2002 - January 31, 2020 (4536 obs) as the out-of-sample test period. We roll the estimation window (training periods) over the test sample to rebalance the portfolios monthly. At the end of each month, prior to portfolio construction, we remove stocks with less than 2 years of historical stock return data.

We examine the performance of Factor Graphical Lasso for three alternative portfolio allocations (2.2), (2.3) and (2.10) and compare it with the equal-weighted portfolio (EW), index portfolio (Index), FClime, FLW, FNLW (as in the simulations, we use alternative covariance and precision estimators that incorporate the factor structure through Sherman-Morrison inversion formula), POET and Projected POET. Index is the composite S&P500 index listed as ^GSPC. We take the risk-free rate and Fama/French factors from Kenneth R. French's data library.

2.6.2 Performance Measures

Similarly to [24], we consider four metrics commonly reported in the finance literature: the Sharpe Ratio, the portfolio turnover, the average return and the risk of a portfolio (which is defined as the square root of the out-of-sample variance of the portfolio). We consider two scenarios: with and without transaction costs. Let T denote the total number of observations, the training sample consists of $m = 504$ observations, and the test sample is $n = T - m$.

When transaction costs are not taken into account, the out-of-sample average portfolio return, variance and Sharpe Ratio (SR) are

$$\hat{\mu}_{\text{test}} = \frac{1}{n}\sum_{t=m}^{T-1} \widehat{\mathbf{w}}_t' \mathbf{r}_{t+1}, \;\; \hat{\sigma}^2_{\text{test}} = \frac{1}{n-1}\sum_{t=m}^{T-1}(\widehat{\mathbf{w}}_t'\mathbf{r}_{t+1} - \hat{\mu}_{\text{test}})^2, \;\; \text{SR} = \hat{\mu}_{\text{test}}/\hat{\sigma}_{\text{test}}. \tag{2.25}$$

When transaction costs are considered, we follow [9, 24, 39, 100] to account for the transaction costs, further denoted as tc. In line with the aforementioned papers, we set tc = 50bps. Define the excess portfolio at time $t+1$ with transaction costs (tc) as

$$r_{t+1,\text{portfolio}} = \; \widehat{\mathbf{w}}_t'\mathbf{r}_{t+1} - \text{tc}(1 + \widehat{\mathbf{w}}_t'\mathbf{r}_{t+1})\sum_{j=1}^{p}\left|\hat{w}_{t+1,j} - \hat{w}^+_{t,j}\right|, \tag{2.26}$$

where

$$\hat{w}^+_{t,j} = \hat{w}_{t,j}\frac{1 + r_{t+1,j} + r^f_{t+1}}{1 + r_{t+1,\text{portfolio}} + r^f_{t+1}}, \tag{2.27}$$

$r_{t+1,j} + r^f_{t+1}$ is sum of the excess return of the j-th asset and risk-free rate, and $r_{t+1,\text{portfolio}} + r^f_{t+1}$ is the sum of the excess return of the portfolio and risk-free rate. The out-of-sample average portfolio return, variance, Sharpe Ratio and turnover are defined accordingly:

$$\hat{\mu}_{\text{test,tc}} = \frac{1}{n}\sum_{t=m}^{T-1} r_{t,\text{portfolio}},\ \hat{\sigma}^2_{\text{test,tc}} = \frac{1}{n-1}\sum_{t=m}^{T-1}(r_{t,\text{portfolio}} - \hat{\mu}_{\text{test,tc}})^2,\ \text{SR}_{\text{tc}} = \hat{\mu}_{\text{test,tc}}/\hat{\sigma}_{\text{test,tc}}, \tag{2.28}$$

$$\text{Turnover} = \frac{1}{n}\sum_{t=m}^{T-1}\sum_{j=1}^{p}\left|\hat{w}_{t+1,j} - \hat{w}^{+}_{t,j}\right|. \tag{2.29}$$

2.6.3 Results

This section explores the performance of the Factor Graphical Lasso for the financial portfolio using daily data. We consider two scenarios, when the factors are unknown and estimated using the standard PCA (statistical factors), and when the factors are known. The number of statistical factors, $\hat{K}$, is estimated in accordance with Remark 1. For the scenario with known factors we include up to 5 Fama-French factors: FF1 includes the excess return on the market, FF3 includes FF1 plus size factor (Small Minus Big, SMB) and value factor (High Minus Low, HML), and FF5 includes FF3 plus profitability factor (Robust Minus Weak, RMW) and risk factor (Conservative Minus Agressive, CMA). In Table 2.1 and Appendix 2.C, we report the daily and monthly portfolio performance for three alternative portfolio allocations in (2.2), (2.3) and (2.10). Following [24], we set a return target $\mu = 0.0378\%$ which is equivalent to 10% yearly return when compounded. The target level of risk for the weight-constrained and risk-constrained Markowitz portfolio (MWC and MRC) is set at $\sigma = 0.013$ which is the standard deviation of the daily excess returns of the S&P500 index in the first training set. Following [24], transaction costs for each individual stock are set to be a constant 0.1%.

	Markowitz Risk-Constrained				Markowitz Weight-Constrained				Global Minimum-Variance			
	Return	**Risk**	**SR**	**Turnover**	**Return**	**Risk**	**SR**	**Turnover**	**Return**	**Risk**	**SR**	**Turnover**
Without TC												
EW	2.33E-04	1.90E-02	0.0123	-	2.33E-04	1.90E-02	0.0123	-	2.33E-04	1.90E-02	0.0123	-
Index	1.86E-04	1.17E-02	0.0159	-	1.86E-04	1.17E-02	0.0159	-	1.86E-04	1.17E-02	0.0159	-
FGL	8.12E-04	2.66E-02	0.0305	-	2.95E-04	8.21E-03	0.0360	-	2.94E-04	7.51E-03	0.0392	-
FClime	2.15E-03	8.46E-02	0.0254	-	2.02E-04	9.85E-03	0.0205	-	2.73E-04	1.07E-02	0.0255	-
FLW	4.34E-04	2.65E-02	0.0164	-	3.12E-04	9.96E-03	0.0313	-	3.10E-04	9.38E-03	0.0330	-
FNLW	4.91E-04	6.66E-02	0.0074	-	2.98E-04	1.24E-02	0.0241	-	3.06E-04	1.32E-02	0.0231	-
POET	NaN	NaN	NaN	-	-7.06E-04	2.74E-01	-0.0026	-	1.07E-03	2.71E-01	0.0039	-
Projected POET	1.20E-03	1.71E-01	0.0070	-	-8.06E-05	1.61E-02	-0.0050	-	-7.57E-05	1.93E-02	-0.0039	-
FGL (FF1)	7.96E-04	2.80E-02	0.0285	-	3.73E-04	8.73E-03	0.0427	-	3.52E-04	8.62E-03	0.0408	-
FGL (FF3)	6.51E-04	2.74E-02	0.0238	-	3.52E-04	8.96E-03	0.0393	-	3.39E-04	8.94E-03	0.0379	-
FGL (FF5)	5.87E-04	2.70E-02	0.0217	-	3.47E-04	9.38E-03	0.0370	-	3.36E-04	9.29E-03	0.0362	-
With TC												
EW	2.01E-04	1.90E-02	0.0106	0.0292	2.01E-04	1.90E-02	0.0106	0.0292	2.01E-04	1.90E-02	0.0106	0.0292
FGL	4.47E-04	2.66E-02	0.0168	0.3655	2.30E-04	8.22E-03	0.0280	0.0666	2.32E-04	7.52E-03	0.0309	0.0633
FClime	1.18E-03	8.48E-02	0.0139	1.0005	1.67E-04	9.86E-03	0.0170	0.0369	2.46E-04	1.07E-02	0.0230	0.0290
FLW	-5.54E-05	2.65E-02	-0.0021	0.4874	1.92E-04	9.98E-03	0.0193	0.1207	1.92E-04	9.39E-03	0.0204	0.1194
FNLW	-2.39E-03	7.03E-02	-0.0340	3.6370	5.50E-05	1.25E-02	0.0044	0.2441	6.08E-05	1.33E-02	0.0046	0.2457
POET	NaN	NaN	NaN	NaN	-2.28E-02	5.55E-01	-0.0411	113.3848	-2.81E-02	4.21E-01	-0.0666	132.8215
Projected POET	-1.59E-02	3.64E-01	-0.0437	35.9692	-1.03E-03	1.68E-02	-0.0616	0.9544	-1.37E-03	2.06E-02	-0.0666	1.2946
FGL (FF1)	3.86E-04	2.80E-02	0.0138	0.4068	2.82E-04	8.74E-03	0.0323	0.0903	2.63E-04	8.63E-03	0.0305	0.0887
FGL (FF3)	2.47E-04	2.74E-02	0.0090	0.4043	2.60E-04	8.98E-03	0.0290	0.0928	2.49E-04	8.96E-03	0.0278	0.0911
FGL (FF5)	1.83E-04	2.71E-02	0.0068	0.4032	2.53E-04	9.40E-03	0.0269	0.0952	2.43E-04	9.30E-03	0.0262	0.0937

Table 2.1: Daily portfolio returns, risk, Sharpe Ratio (SR) and turnover.

Let us summarize the results for daily data in Table 2.1: **(1)** MRC portfolios produce higher return and higher risk, compared to MWC and GMV. However, the out-of-sample Sharpe Ratio for MRC is lower than that of MWC and GMV, which implies that the higher risk of MRC portfolios is not fully compensated by the higher return. **(2)** FGL outperforms all the competitors, including EW and Index. Specifically, our method has the lowest risk and turnover (compared to FClime, FLW, FNLW and POET), and the highest out-of-sample Sharpe Ratio compared with all alternative methods. **(3)** The implementation of POET for MRC resulted in the erratic behavior of this method for estimating portfolio weights, concretely, many entries in the weight matrix had "NaN" entries. We elaborate on the reasons behind such performance below. **(4)** Using the observable Fama-French factors in the FGL, in general, produces portfolios with higher return and higher out-of-sample Sharpe Ratio compared to the portfolios based on statistical factors. Interestingly, this increase in return is not followed by higher risk. The results for monthly data are provided in Appendix 2.C: all the conclusions are similar to the ones for daily data.

We now examine possible reasons behind the observed puzzling behavior of POET and Projected POET. The erratic behavior of the former is caused by the fact that POET estimator of covariance matrix was not positive-definite which produced poor estimates of GMV and MWC weights and made it infeasible to compute MRC weights (recall, by construction MRC weight in (2.10) requires taking a square root). To explore deteriorated behavior of Projected POET, let us highlight two findings outlined by the existing closely related literature. First, [8] examined "pervasiveness" degree, or strength, of 146 factors commonly used in the empirical finance literature, and found that only the market factor was strong, while all other factors were semi-strong. This indicates that the factor pervasiveness assumption **(A.2)** might be unrealistic in practice. Second, as pointed out by [113], “the quality of POET dramatically deteriorates as the systematic-idiosyncratic eigenvalue gap becomes small”. Therefore, being guided by the two aforementioned findings, we attribute deteriorated performance of POET and Projected POET to the decreased gap between the diverging and bounded eigenvalues documented in the past studies on financial returns. High sensitivity of these two covariance estimators in such settings was further supported by our additional simulation study (Appendix 2.B.4) examining the robustness of portfolios constructed using different covariance and precision estimators.

Table 2.2 compares the performance of FGL and the alternative methods for the daily data for different time periods of interesting episodes in terms of the cumulative excess return (CER) and risk. To demonstrate the performance of all methods during the periods of recession and expansion, we chose four periods and recorded CER for the whole year in each period of interest. Two years, 2002 and 2008 correspond to the recession periods, which is why we we refer to them as “Downturns”. We note that the references to Argentine Great Depression and The Financial Crisis do not intend to limit these economic downturns to only one year. They merely provide the context for the recessions. The other two years, 2017 and 2019, correspond to the years which were relatively favorable to the stock market (“Booms”).

Table 2.2 reveals some interesting findings: **(1)** MRC portfolios yield higher CER and they are characterized by higher risk. **(2)** MRC is the only type of portfolio that produces positive CER during both recessions. Note that all models that used MWC and GMV during that time experienced large negative CER. **(3)** When EW and Index have positive CER (during Boom periods), all portfolio formulations also produce positive CER. However, the return accumulated by MRC is mostly higher than that by MWC and GMV portfolio formulations. **(4)** FGL mostly outperforms the competitors, including EW and Index in terms of CER and risk.

	Downturn #1 Argentine Great Depression (2002)		Downturn #2 Financial Crisis (2008)		Boom #1 (2017)		Boom #2 (2019)	
	CER	**Risk**	**CER**	**Risk**	**CER**	**Risk**	**CER**	**Risk**
Equal-Weighted and Index								
EW	-0.1633	0.0160	-0.5622	0.0310	0.0627	0.0218	0.1642	0.0185
Index	-0.2418	0.0168	-0.4746	0.0258	0.1752	0.0042	0.2934	0.0086
Markowitz Risk-Constrained (MRC)								
FGL	0.2909	0.0206	0.2938	0.0282	0.7267	0.0142	0.6872	0.0263
FClime	-0.0079	0.0348	-0.8912	0.1484	0.5331	0.0383	0.2346	0.0557
FLW	0.0308	0.0231	0.2885	0.0315	0.3164	0.0118	0.5520	0.0287
FNLW	0.0728	0.0213	0.2075	0.0392	0.5796	0.0497	0.6315	0.0355
Projected POET	-0.6178	0.0545	2.81E-05	0.1874	-0.7599	0.1197	1.8592	0.1177
Markowitz Weight-Constrained (MWC)								
FGL	-0.0138	0.0082	-0.1956	0.0135	0.1398	0.0044	0.3787	0.0072
FClime	-0.1045	0.0124	-0.3974	0.0204	0.1309	0.0041	0.2595	0.0078
FLW	-0.0158	0.0080	-0.2789	0.0126	0.1267	0.0037	0.3018	0.0085
FNLW	-0.0195	0.0078	-0.2811	0.0123	-0.0361	0.0087	0.4078	0.0098
POET	-0.2820	0.0324	-0.9989	0.1198	0.5720	0.0630	1.4756	0.0403
Projected POET	-0.0217	0.0130	-0.0842	0.0176	-0.0877	0.0089	0.5300	0.0176
Global Minimum-Variance Portfolio (GMV)								
FGL	-0.0044	0.0081	-0.2113	0.0138	0.1384	0.0045	0.3703	0.0072
FClime	-0.1061	0.0129	-0.4410	0.0241	0.1264	0.0041	0.2829	0.0081
FLW	-0.0151	0.0080	-0.2926	0.0128	0.1323	0.0037	0.2994	0.0084
FNLW	-0.0206	0.0078	-0.2959	0.0124	-0.0388	0.0090	0.3287	0.0097
POET	-0.3190	0.0330	-0.9928	0.0931	-1.0000	0.2414	1.6301	0.0318
Projected POET	-0.0662	0.0135	0.0829	0.0247	-0.1106	0.0115	0.6870	0.0186

Table 2.2: Cumulative excess return (CER) and risk of portfolios using daily data.

2.7 Conclusion

In this paper, we propose a new conditional precision matrix estimator for the excess returns under the approximate factor model with unobserved factors that combines the benefits of graphical models and factor structure. We established consistency of FGL in the spectral and ℓ_1 matrix norms. In addition, we proved consistency of the portfolio weights and risk exposure for three formulations of the optimal portfolio allocation without assuming sparsity on the covariance or precision matrix of stock returns. All theoretical results established in this paper hold for a wide range of distributions: sub-Gaussian family (including Gaussian) and elliptical family. Our simulations demonstrate that FGL is robust to very heavy-tailed distributions, which makes our method suitable for the financial applications. Furthermore, we demonstrate that in contrast to POET and Projected POET, the success of the proposed method does not heavily depend on the factor pervasiveness assumption: FGL is robust to the scenarios when the gap between the diverging and bounded eigenvalues decreases.

The empirical exercise uses the constituents of the S&P500 index and demonstrates superior performance of FGL compared to several alternative models for estimating precision (FClime) and covariance (FLW, FNLW, POET) matrices, Equal-Weighted (EW) portfolio and Index portfolio in terms of the out-of-sample Sharpe Ratio and risk. This result is robust to both monthly and daily data. We examine three different portfolio formulations and discover that the only portfolios that produce positive cumulative excess return (CER) during recessions are the ones that relax the constraint requiring portfolio weights sum up to one.

Appendices

This Appendix is structured as follows: Appendix 2.A contains proofs of the theorems and accompanying lemmas, Appendix 2.B provides additional simulations for Section 5, additional empirical results for Section 6 are located in Appendix 2.C.

2.A Proofs of the Theorems

2.A.1 Lemmas for Theorem 1

Lemma 1 *Under the assumptions of Theorem 1,*

(a) $\max_{i,j\le K}\Big|(1/T)\sum_{t=1}^{T} f_{it}f_{jt} - \mathbb{E}[f_{it}f_{jt}]\Big| = \mathcal{O}_P(\sqrt{1/T})$,

(b) $\max_{i,j\le p}\Big|(1/T)\sum_{t=1}^{T} \varepsilon_{it}\varepsilon_{jt} - \mathbb{E}[\varepsilon_{it}\varepsilon_{jt}]\Big| = \mathcal{O}_P(\sqrt{\log p/T})$,

(c) $\max_{i\le K, j\le p}\Big|(1/T)\sum_{t=1}^{T} f_{it}\varepsilon_{jt}\Big| = \mathcal{O}_P(\sqrt{\log p/T})$.

Proof. The proof of Lemma 1 can be found in Fan et al. (2011) (Lemma B.1). ■

Lemma 2 *Under Assumption **(A.4)**,* $\max_{t\le T}\sum_{s=1}^{K}|\mathbb{E}[\boldsymbol{\varepsilon}_s'\boldsymbol{\varepsilon}_t]|/p = \mathcal{O}(1)$.

Proof. The proof of Lemma 2 can be found in Fan et al. (2013) (Lemma A.6). ■

Lemma 3 *For $\widehat{K}$ defined in expression (3.6),*

$$\Pr\left(\widehat{K}=K\right)\to 1.$$

Proof. The proof of Lemma 3 can be found in Li et al. (2017) (Theorem 1 and Corollary 1). ■ Using the expressions (A.1) in Bai (2003) and (C.2) in Fan et al. (2013), we have the following identity:

$$\widehat{\mathbf{f}}_t-\mathbf{H}\mathbf{f}_t=\left(\frac{\mathbf{V}}{p}\right)^{-1}\left[\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s\frac{\mathbb{E}\left[\boldsymbol{\varepsilon}_s'\boldsymbol{\varepsilon}_t\right]}{p}+\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s\zeta_{st}+\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s\eta_{st}+\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s\xi_{st}\right],\tag{2.30}$$

where $\zeta_{st}=\boldsymbol{\varepsilon}_s'\boldsymbol{\varepsilon}_t/p-\mathbb{E}\left[\boldsymbol{\varepsilon}_s'\boldsymbol{\varepsilon}_t\right]/p$, $\eta_{st}=\mathbf{f}_s'\sum_{i=1}^{p}\mathbf{b}_i\varepsilon_{it}/p$ and $\xi_{st}=\mathbf{f}_t'\sum_{i=1}^{p}\mathbf{b}_i\varepsilon_{is}/p$.

Lemma 4 *For all $i\leq\widehat{K}$,*

(a) $(1/T)\sum_{t=1}^{T}\left[(1/T)\sum_{t=1}^{T}\hat{f}_{is}\mathbb{E}\left[\boldsymbol{\varepsilon}_s'\boldsymbol{\varepsilon}_t\right]/p\right]^2=\mathcal{O}_P(T^{-1})$,

(b) $(1/T)\sum_{t=1}^{T}\left[(1/T)\sum_{t=1}^{T}\hat{f}_{is}\zeta_{st}/p\right]^2=\mathcal{O}_P(p^{-1})$,

(c) $(1/T)\sum_{t=1}^{T}\left[(1/T)\sum_{t=1}^{T}\hat{f}_{is}\eta_{st}/p\right]^2=\mathcal{O}_P(K^2/p)$,

(d) $(1/T)\sum_{t=1}^{T}\left[(1/T)\sum_{t=1}^{T}\hat{f}_{is}\xi_{st}/p\right]^2=\mathcal{O}_P(K^2/p)$.

Proof. We only prove (c) and (d), the proof of (a) and (b) can be found in Fan et al. (2013) (Lemma 8).

(c) Recall, $\eta_{st}=\mathbf{f}_s'\sum_{i=1}^{p}\mathbf{b}_i\varepsilon_{it}/p$. Using Assumption **(A.5)**, we get $\mathbb{E}\left[(1/T)\times\sum_{t=1}^{T}\|\sum_{i=1}^{p}\mathbf{b}_i\varepsilon_{it}\|^2\right]=\mathbb{E}\left[\|\sum_{i=1}^{p}\mathbf{b}_i\varepsilon_{it}\|^2\right]=\mathcal{O}(pK)$. Therefore, by the Cauchy-Schwarz inequality and the facts that $(1/T)\sum_{t=1}^{T}\|\mathbf{f}_t\|^2=\mathcal{O}(K)$, and, $\forall i$,

$\sum_{s=1}^T \hat{f}_{is}^2 = T$,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T\Big(\frac{1}{T}\sum_{s=1}^T \hat{f}_{is}\eta_{st}\Big)^2 &\le \left\|\frac{1}{T}\sum_{s=1}^T\|\hat{f}_{is}\mathbf{f}_s'\|^2\frac{1}{T}\sum_{t=1}^T\frac{1}{p}\|\sum_{j=1}^p \mathbf{b}_i\varepsilon_{jt}\|\right\|^2 \\
&\le \frac{1}{Tp^2}\sum_{t=1}^T\left\|\sum_{j=1}^p \mathbf{b}_i\varepsilon_{jt}\right\|^2\left(\frac{1}{T}\sum_{s=1}^T\hat{f}_{is}^2\frac{1}{T}\sum_{s=1}^T\|\mathbf{f}_s\|^2\right) \\
&= \mathcal{O}_P\Big(\frac{K}{p}\cdot K\Big) = \mathcal{O}_P\Big(\frac{K^2}{p}\Big).
\end{aligned}$$

(d) Using a similar approach as in part (c):

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T\Big(\frac{1}{T}\sum_{s=1}^T \hat{f}_{is}\xi_{st}\Big)^2 &= \frac{1}{T}\sum_{t=1}^T\left|\frac{1}{T}\sum_{s=1}^T \mathbf{f}_t'\sum_{j=1}^p \varepsilon_{js}\frac{1}{p}\hat{f}_{is}\right|^2 \\
&\le \Big(\frac{1}{T}\sum_{t=1}^T\|\mathbf{f}_t\|^2\Big)\left\|\frac{1}{T}\sum_{s=1}^T\sum_{j=1}^p \mathbf{b}_j\varepsilon_{js}\frac{1}{p}\hat{f}_{is}\right\|^2 \\
&\le \Big(\frac{1}{T}\sum_{t=1}^T\|\mathbf{f}_t\|^2\Big)\frac{1}{T}\sum_{s=1}^T\left\|\sum_{j=1}^p \mathbf{b}_j\varepsilon_{js}\frac{1}{p}\right\|^2\Big(\frac{1}{T}\sum_{s=1}^T\hat{f}_{is}^2\Big) \\
&= \mathcal{O}_P\Big(K\cdot\frac{pK}{p^2}\cdot 1\Big) = \mathcal{O}_P\Big(\frac{K^2}{p}\Big)
\end{aligned}$$

■

Lemma 5

(a) $\max_{t\le T}\left\|(1/(Tp))\sum_{s=1}^T\widehat{\mathbf{f}}_s'\mathbb{E}[\boldsymbol{\varepsilon}_s'\boldsymbol{\varepsilon}_t]\right\| = \mathcal{O}_P(K/\sqrt{T})$.

(b) $\max_{t\le T}\left\|(1/(Tp))\sum_{s=1}^T\widehat{\mathbf{f}}_s'\zeta_{st}\right\| = \mathcal{O}_P(\sqrt{K}T^{1/4}/\sqrt{p})$.

(c) $\max_{t\le T}\left\|(1/(Tp))\sum_{s=1}^T\widehat{\mathbf{f}}_s'\eta_{st}\right\| = \mathcal{O}_P(KT^{1/4}/\sqrt{p})$.

(d) $\max_{t\le T}\left\|(1/(Tp))\sum_{s=1}^T\widehat{\mathbf{f}}_s'\xi_{st}\right\| = \mathcal{O}_P(KT^{1/4}/\sqrt{p})$.

Proof. Our proof is similar to the proof in Fan et al. (2013). However, we relax the assumptions of fixed K.

(a) Using the Cauchy-Schwarz inequality, Lemma 2, and the fact that $(1/T)\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t\|^2 = \mathcal{O}_P(K)$, we get

$$\begin{aligned}
&\max_{t\leq T}\left\|\frac{1}{Tp}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s'\mathbb{E}\left[\boldsymbol{\varepsilon}_s'\boldsymbol{\varepsilon}_t\right]\right\| \leq \max_{t\leq T}\left[\frac{1}{T}\sum_{s=1}^{T}\left\|\widehat{\mathbf{f}}_s\right\|\frac{1}{T}\sum_{s=1}^{T}\left(\frac{\mathbb{E}\left[\boldsymbol{\varepsilon}_s'\boldsymbol{\varepsilon}_t\right]}{p}\right)^2\right]^{1/2}\\
&\leq \mathcal{O}_P(K)\max_{t\leq T}\left[\frac{1}{T}\sum_{s=1}^{T}\left(\frac{\mathbb{E}\left[\boldsymbol{\varepsilon}_s'\boldsymbol{\varepsilon}_t\right]}{p}\right)^2\right]^{1/2}\\
&\leq \mathcal{O}_P(K)\max_{s,t}\sqrt{\left|\frac{\mathbb{E}\left[\boldsymbol{\varepsilon}_s'\boldsymbol{\varepsilon}_t\right]}{p}\right|}\max_{t\leq T}\left[\frac{1}{T}\sum_{s=1}^{T}\left|\frac{\mathbb{E}\left[\boldsymbol{\varepsilon}_s'\boldsymbol{\varepsilon}_t\right]}{p}\right|\right]^{1/2} = \mathcal{O}_P\Big(K\cdot 1\cdot\frac{1}{\sqrt{T}}\Big) = \mathcal{O}_P\Big(\frac{K}{\sqrt{T}}\Big).
\end{aligned}$$

(b) Using the Cauchy-Schwarz inequality,

$$\begin{aligned}
\max_{t\leq T}\left\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s'\zeta_{st}\right\| \leq \max_{t\leq T}\frac{1}{T}\left(\sum_{s=1}^{T}\left\|\widehat{\mathbf{f}}_s\right\|^2\sum_{s=1}^{T}\zeta_{st}^2\right)^{1/2} &\leq \left(\mathcal{O}_P(K)\max_{t}\frac{1}{T}\sum_{s=1}^{T}\zeta_{st}^2\right)^{1/2}\\
&= \mathcal{O}_P\Big(\sqrt{K}\cdot T^{1/4}/\sqrt{p}\cdot\Big).
\end{aligned}$$

To obtain the last inequality we used Assumption **(A.5)**(b) to get $\mathbb{E}\left[(1/T)\sum_{s=1}^{T}\zeta_{st}^2\right]^2 \leq \max_{s,t\leq T}\mathbb{E}\left[\zeta_{st}^4\right] = \mathcal{O}(1/p^2)$, and then applied the Chebyshev inequality and Bonferroni's method that yield $\max_t(1/T)\sum_{s=1}^{T}\zeta_{st}^2 = \mathcal{O}_P\Big(\sqrt{T}/p\Big)$.

(c) Using the definition of η_{st} we get

$$\max_{t\leq T}\left\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s'\eta_{st}\right\| \leq \left\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s\mathbf{f}_s'\right\|\max_{t}\left\|\frac{1}{p}\sum_{i=1}^{p}\mathbf{b}_i\varepsilon_{it}\right\| = \mathcal{O}_P\Big(K\cdot T^{1/4}/\sqrt{p}\Big).$$

To obtain the last rate we used Assumption **(A.5)**(c) together with the Chebyshev inequality and Bonferroni's method to get $\max_{t\leq T}\|\sum_{i=1}^{p}\mathbf{b}_i\varepsilon_{it}\| = \mathcal{O}_P\Big(T^{1/4}\sqrt{p}\Big)$.

(d) In the proof of Lemma 4 we showed that $\|(1/T)\times\sum_{t=1}^{T}\sum_{i=1}^{p}\mathbf{b}_i\varepsilon_{it}(1/p)\widehat{\mathbf{f}}_s\|^2 = \mathcal{O}\Big(\sqrt{K/p}\Big)$. Furthermore, Assumption **(A.3)** implies $\mathbb{E}\left[K^{-2}\mathbf{f}_t\right]^4 < M$, therefore, $\max_{t\leq T}\|\mathbf{f}_t\| = \mathcal{O}_P\Big(T^{1/4}\sqrt{K}\Big)$. Using these bounds we get

$$\max_{t\leq T}\left\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s'\xi_{st}\right\| \leq \max_{t\leq T}\|\mathbf{f}_t\|\cdot\left\|\sum_{s=1}^{T}\sum_{i=1}^{p}\mathbf{b}_i\varepsilon_{it}\frac{1}{p}\widehat{\mathbf{f}}_s\right\| = \mathcal{O}_P\Big(T^{1/4}\sqrt{K}\cdot\sqrt{K/p}\Big)$$
$$= \mathcal{O}_P\Big(T^{1/4}K/\sqrt{p}\Big).$$

■

Lemma 6

(a) $\max_{i\leq K}(1/T)\sum_{t=1}^{T}(\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t)_i^2 = \mathcal{O}_P(1/T + K^2/p)$.

(b) $(1/T)\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\|^2 = \mathcal{O}_P(K/T + K^3/p)$.

(c) $\max_{t\leq T}(1/T)\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\| = \mathcal{O}_P(K/\sqrt{T} + KT^{1/4}/\sqrt{p})$.

Proof. Similarly to Fan et al. (2013), we prove this lemma conditioning on the event $\hat{K} = K$. Since $\Pr(\hat{K} \neq K) = o(1)$, the unconditional arguments are implied.

(a) Using (2.30), for some constant $C > 0$,

$$\begin{aligned}
\max_{i\le K}(1/T)\sum_{t=1}^{T}(\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t)_i^2 &\le C\max_{i\le K}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}\frac{\mathbb{E}[\boldsymbol{\varepsilon}_s'\boldsymbol{\varepsilon}_t]}{p}\right)^2 \\
&+ C\max_{i\le K}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}\zeta_{st}\right)^2 \\
&+ C\max_{i\le K}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}\zeta_{st}\right)^2 \\
&+ C\max_{i\le K}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}\xi_{st}\right)^2 \\
&= \mathcal{O}_P\left(\frac{1}{T}+\frac{1}{p}+\frac{K^2}{p}+\frac{K^2}{p}\right) = \mathcal{O}_P(1/T + K^2/p).
\end{aligned}$$

(b) Part (b) follows from part (a) and

$$\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\|^2 \le K\max_{i\le K}\frac{1}{T}\sum_{t=1}^{T}(\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t)_i^2.$$

(c) Part (c) is a direct consequence of 2.30 and Lemma 5.

■

Lemma 7

(a) $\mathbf{H}\mathbf{H}' = \mathbf{I}_{\hat{K}} + \mathcal{O}_P(K^{5/2}/\sqrt{T} + K^{5/2}/\sqrt{p})$.

(b) $\mathbf{H}\mathbf{H}' = \mathbf{I}_{K} + \mathcal{O}_P(K^{5/2}/\sqrt{T} + K^{5/2}/\sqrt{p})$.

Proof. Similarly to Lemma 6, we first condition on $\hat{K} = K$.

(a) The key observation here is that, according to the definition of $\mathbf{H}$, its rank grows with K, that is, $\|\mathbf{H}\| = \mathcal{O}_P(K)$. Let $\widehat{\text{cov}}(\mathbf{Hf}_t) = (1/T)\sum_{t=1}^{T} \mathbf{Hf}_t(\mathbf{Hf}_t)'$. Using the triangular inequality we get

$$\left\|\mathbf{HH'} - \mathbf{I}_{\hat{K}}\right\|_F \leq \left\|\mathbf{HH'} - \widehat{\text{cov}}(\mathbf{Hf}_t)\right\|_F + \left\|\widehat{\text{cov}}(\mathbf{Hf}_t) - \mathbf{I}_{\hat{K}}\right\|_F. \tag{2.31}$$

To bound the first term in (2.31), we use Lemma 1:

$\|\mathbf{HH'} - \widehat{\text{cov}}(\mathbf{Hf}_t)\|_F \leq \|\mathbf{H}\|^2 \|\mathbf{I}_K - \widehat{\text{cov}}(\mathbf{Hf}_t)\|_F = \mathcal{O}_P(K^{5/2}/\sqrt{T})$.

To bound the second term in (2.31), we use the Cauchy-Schwarz inequality and Lemma 6:

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{Hf}_t(\mathbf{Hf}_t)' - \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t'\right\|_F \leq \left\|\frac{1}{T}\sum_{t=1}^{T}(\mathbf{Hf}_t - \widehat{\mathbf{f}}_t)(\mathbf{Hf}_t)'\right\|_F + \left\|\frac{1}{T}\sum_{t}\widehat{\mathbf{f}}_t(\widehat{\mathbf{f}}_t' - (\mathbf{Hf}_t)')\right\|_F$$
$$\leq \left(\frac{1}{T}\sum_{t=1}\left\|\mathbf{Hf}_t - \widehat{\mathbf{f}}_t\right\|^2 \frac{1}{T}\sum_{t=1}\|\mathbf{Hf}_t\|^2\right)^{1/2} + \left(\frac{1}{T}\sum_{t=1}\left\|\mathbf{Hf}_t - \widehat{\mathbf{f}}_t\right\|^2 \frac{1}{T}\sum_{t=1}\left\|\widehat{\mathbf{f}}_t\right\|^2\right)^{1/2}$$
$$= \mathcal{O}_P\left(\left(\frac{K}{T} + \frac{K^3}{p}\cdot K\right)^{1/2} + \left(\frac{K}{T} + \frac{K^3}{p}\cdot K^2\right)^{1/2}\right) = \mathcal{O}_P\left(\frac{K^{3/2}}{\sqrt{T}} + \frac{K^{5/2}}{\sqrt{p}}\right).$$

(b) The proof of (b) follows from $\Pr(\hat{K} = K) \to 1$ and the arguments made in Fan et al. (2013), (Lemma 11) for fixed K.

■

2.A.2 Proof of Theorem 1

The second part of Theorem 1 was proved in Lemma 6. We now proceed to the convergence rate of the first part. Using the following definitions: $\widehat{\mathbf{b}}_i = (1/T)\sum_{t=1}^{T} r_{it}\widehat{\mathbf{f}}_t$ and $(1/T)\sum_{t=1}^{T} \widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t' = \mathbf{I}_K$, we obtain

$$\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i = \frac{1}{T}\sum_{t=1}^{T}\mathbf{H}\mathbf{f}_t\varepsilon_{it} + \frac{1}{T}\sum_{t=1}^{T} r_{it}(\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t) + \mathbf{H}\Big(\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t' - \mathbf{I}_K\Big)\mathbf{b}_i. \tag{2.32}$$

Let us bound each term on the right-hand side of (2.32). The first term is

$$\begin{aligned}\max_{i\le p}\|\mathbf{H}\mathbf{f}_t\varepsilon_{it}\| \le \|\mathbf{H}\| \max_i \sqrt{\sum_{k=1}^{K}\left(\frac{1}{T}\sum_{t=1}^{T} f_{kt}\varepsilon_{it}\right)^2} &\le \|\mathbf{H}\|\sqrt{K} \max_{i\le p, j\le K}\left|\frac{1}{T}\sum_{t=1}^{T} f_{jt}\varepsilon_{it}\right| \\ &= \mathcal{O}_P\Big(K \cdot K^{1/2} \cdot \sqrt{\log p/T}\Big),\end{aligned}$$

where we used Lemmas 1 and 7 together with Bonferroni's method. For the second term,

$$\max_i\left\|\frac{1}{T}\sum_{t=1}^{T} r_{it}\Big(\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\Big)\right\| \le \max_i\left(\frac{1}{T}\sum_{t=1}^{T} r_{it}^2 \frac{1}{T}\sum_{t=1}^{T}\Big\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\Big\|^2\right)^{1/2} = \mathcal{O}_P\left(\frac{1}{T} + \frac{K^2}{p}\right)^{1/2},$$

where we used Lemma 6 and the fact that $\max_i T^{-1}\sum_{t=1}^{T} r_{it}^2 = \mathcal{O}_P(1)$ since $\mathbb{E}\left[r_{it}^2\right] = \mathcal{O}(1)$. Finally, the third term is $\mathcal{O}_P(K^2T^{-1/2})$ since $\|(1/T)\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t' - \mathbf{I}_K\| = \mathcal{O}_P\Big(KT^{-1/2}\Big)$, $\|\mathbf{H}\| = \mathcal{O}_P(K)$ and $\max_i\|\mathbf{b}\|_i = \mathcal{O}(1)$ by Assumption **(B.1)**.

2.A.3 Corollary 1

As a consequence of Theorem 1, we get the following corollary:

Corollary 1 *Under the assumptions of Theorem 1,*

$$\max_{i\leq p, t\leq T}\left\|\widehat{\mathbf{b}}_i'\widehat{\mathbf{f}}_t - \mathbf{b}_i'\mathbf{f}_t\right\| = \mathcal{O}_P(\log T^{1/r_2} K^2\sqrt{\log p/T} + K^2T^{1/4}/\sqrt{p}).$$

Proof. Using Assumption **(A.4)** and Bonferroni's method, we have $\max_{t\leq T}\|\mathbf{f}_t\| = \mathcal{O}_P(\sqrt{K}\log T^{1/r_2})$. By Theorem 1, uniformly in i and t:

$$\begin{aligned}
\left\|\widehat{\mathbf{b}}_i'\widehat{\mathbf{f}}_t - \mathbf{b}_i'\mathbf{f}_t\right\| &\leq \left\|\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\right\|\left\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\right\| + \|\mathbf{H}\mathbf{b}_i\|\left\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\right\| \\
&\quad + \left\|\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\right\|\|\mathbf{H}\mathbf{f}_t\| + \|\mathbf{b}_i\|\|\mathbf{f}_t\|\left\|\mathbf{H}'\mathbf{H} - \mathbf{I}_K\right\| \\
&= \mathcal{O}_P\Bigg(\Big(K^{3/2}\sqrt{\frac{\log p}{T}} + \frac{K}{\sqrt{p}}\Big)\cdot\Big(\frac{K}{\sqrt{T}} + \frac{KT^{1/4}}{\sqrt{p}}\Big)\Bigg) \\
&\quad + \mathcal{O}_P\Bigg(K\cdot\Big(\frac{K}{\sqrt{T}} + \frac{KT^{1/4}}{\sqrt{p}}\Big)\Bigg) \\
&\quad + \mathcal{O}_P\Bigg(\Big(K^{3/2}\sqrt{\frac{\log p}{T}} + \frac{K}{\sqrt{p}}\Big)\cdot\log T^{1/r_2}K^{1/2}\Bigg) \\
&\quad + \mathcal{O}_P\Bigg(\log T^{1/r_2}K^{1/2}\Big(\frac{K^{5/2}}{\sqrt{T}} + \frac{K^{5/2}}{\sqrt{p}}\Big)\Bigg) \\
&= \mathcal{O}_P\Big(\log T^{1/r_2}K^2\sqrt{\log p/T} + K^2T^{1/4}/\sqrt{p}\Big).
\end{aligned}$$

■

2.A.4 Proof of Theorem 2

Using the definition of the idiosyncratic components we have $\varepsilon_{it} - \hat{\varepsilon}_{it} = \mathbf{b}_i'\mathbf{H}'(\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t) + (\widehat{\mathbf{b}}_i' - \mathbf{b}_i'\mathbf{H}')\widehat{\mathbf{f}}_t + \mathbf{b}_i'(\mathbf{H}'\mathbf{H} - \mathbf{I}_K)\mathbf{f}_t$. We bound the maximum element-wise difference as follows:

$$\begin{aligned}
\max_{i\le p}\frac{1}{T}\sum_{t=1}^{T}(\varepsilon_{it}-\hat{\varepsilon}_{it})^2 &\le 4\max_i\left\|\mathbf{b}_i'\mathbf{H}'\right\|^2\frac{1}{T}\sum_{t=1}^{T}\left\|\widehat{\mathbf{f}}_t-\mathbf{H}\mathbf{f}_t\right\|^2+4\max_i\left\|\widehat{\mathbf{b}}_i'-\mathbf{b}_i'\mathbf{H}'\right\|^2\frac{1}{T}\sum_{t=1}^{T}\left\|\widehat{\mathbf{f}}_t\right\|^2\\
&+4\max_i\left\|\mathbf{b}_i'\right\|\frac{1}{T}\sum_{t=1}^{T}\left\|\mathbf{f}_t\right\|^2\left\|\mathbf{H}'\mathbf{H}-\mathbf{I}_K\right\|_F^2\\
&=\mathcal{O}\left(K^2\cdot\Big(\frac{K}{T}+\frac{K^3}{p}\Big)\right)+\mathcal{O}\left(\Big(\frac{K^3\log p}{T}+\frac{K^2}{p}\Big)\cdot K\right)\\
&+\mathcal{O}\left(K\cdot\Big(\frac{K^5}{T}+\frac{K^5}{p}\Big)\right)=\mathcal{O}\left(\frac{K^4\log p}{T}+\frac{K^6}{p}\right).
\end{aligned}$$

Let $\omega_{3T} \equiv K^2\sqrt{\log p/T} + K^3/\sqrt{p}$. Denote $\max_{i\le p}(1/T)\sum_{t=1}^{T}(\varepsilon_{it}-\hat{\varepsilon}_{it})^2 = \mathcal{O}_P(\omega_{3T}^2)$. Then, $\max_{i,t}|\varepsilon_{it}-\hat{\varepsilon}_{it}| = \mathcal{O}_P(\omega_{3T}) = o_P(1)$, where the last equality is implied by Corollary 1. As pointed out in the main text, the second part of Theorem 2 is based on the relationship between the convergence rates of the estimated covariance and precision matrices established in Janková and van de Geer (2018) (Theorem 14.1.3).

2.A.5 Lemmas for Theorem 3

Lemma 8 *Under the assumptions of Theorem 1, we have the following results:*

(a) $\|\mathbf{B}\| = \|\mathbf{B}\mathbf{H}'\| = \mathcal{O}(\sqrt{p})$.

(b) $\varrho_T^{-1}\max_{1\le i\le p}\left\|\widehat{\mathbf{b}}_i - \mathbf{H}'\mathbf{b}_i\right\| = o_P(1/\sqrt{K})$ *and* $\max_{1\le i\le p}\left\|\widehat{\mathbf{b}}_i\right\| = \mathcal{O}_P(\sqrt{K})$.

(c) $\varrho_T^{-1}\left\|\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}'\right\| = o_P\left(\sqrt{p/K}\right)$ *and* $\left\|\widehat{\mathbf{B}}\right\| = \mathcal{O}_P(\sqrt{p})$.

Proof. Part (c) is direct consequences of (a)-(b), therefore, we only prove the first two parts in what follows.

(a) Part (a) easily follows from **(B.1)**: $\text{tr}(\mathbf{\Sigma}-\mathbf{B}\mathbf{B}') = \text{tr}(\mathbf{\Sigma})-\|\mathbf{B}\|^2 \geq 0$, since $\text{tr}(\mathbf{\Sigma}) = \mathcal{O}(p)$ by **(B.1)**, we get $\|\mathbf{B}\|^2 = \mathcal{O}(p)$. Part (a) follows from the fact that the linear space spanned by the rows of $\mathbf{B}$ is the same as that by the rows of $\mathbf{B}\mathbf{H}'$, hence, in practice, it does not matter which one is used.

(b) From Theorem 1, we have $\max_{i\leq p}\left\|\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\right\| = \mathcal{O}_P(\omega_{1T})$. Using the definition of ϱ_T from Theorem 2, it follows that $\varrho_T^{-1}\omega_{1T} = o_P(\omega_{1T}\omega_{3T}^{-1})$. Let $\widetilde{z}_T \equiv \omega_{1T}\omega_{3T}^{-1}$. Consider $\varrho_T^{-1}\max_{1\leq i\leq p}\left\|\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\right\| = o_P(z_T)$. The latter holds for any $z_t \geq \widetilde{z}_T$, with the tightest bound obtained when $z_T = \widetilde{z}_T$. For the ease of representation, we use $z_T = 1/\sqrt{K}$ instead of $\widetilde{z}_T$.

The second result in Part (b) is obtained using the fact that $\max_{1\leq i\leq p}\left\|\widehat{\mathbf{b}}_i\right\| \leq \sqrt{K}\|\mathbf{B}\|_{\max}$, where $\|\mathbf{B}\|_{\max} = \mathcal{O}(1)$ by **(B.1)**.

■

Lemma 9 *Let* $\mathbf{\Pi} \equiv \left[\mathbf{\Theta}_f + (\mathbf{B}\mathbf{H}')'\mathbf{\Theta}_\varepsilon(\mathbf{B}\mathbf{H}')\right]^{-1}$, $\widehat{\mathbf{\Pi}} \equiv \left[\widehat{\mathbf{\Theta}}_f + \widehat{\mathbf{B}}'\widehat{\mathbf{\Theta}}_\varepsilon\widehat{\mathbf{B}}\right]^{-1}$. *Also, define* $\mathbf{\Sigma}_f = (1/T)\sum_{t=1}^{T}\mathbf{H}\mathbf{f}_t(\mathbf{H}\mathbf{f}_t)'$, $\mathbf{\Theta}_f = \mathbf{\Sigma}_f^{-1}$, $\widehat{\mathbf{\Sigma}}_f \equiv (1/T)\sum_{t=1}^{T}\widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t'$, *and* $\widehat{\mathbf{\Theta}}_f = \widehat{\mathbf{\Sigma}}_f^{-1}$. *Under the assumptions of Theorem 2, we have the following results:*

(a) $\Lambda_{min}(\mathbf{B}'\mathbf{B})^{-1} = \mathcal{O}(1/p)$.

(b) $\lvert\lvert\lvert\mathbf{\Pi}\rvert\rvert\rvert_2 = \mathcal{O}(1/p)$.

(c) $\varrho_T^{-1}\left\lvert\left\lvert\left\lvert\widehat{\mathbf{\Theta}}_f - \mathbf{\Theta}_f\right\rvert\right\rvert\right\rvert_2 = o_P\left(1/\sqrt{K}\right)$.

(d) $\varrho_T^{-1}\left\lvert\left\lvert\left\lvert\widehat{\mathbf{\Pi}} - \mathbf{\Pi}\right\rvert\right\rvert\right\rvert_2 = \mathcal{O}_P\left(s_T/p\right)$ *and* $\left\lvert\left\lvert\left\lvert\widehat{\mathbf{\Pi}}\right\rvert\right\rvert\right\rvert_2 = \mathcal{O}_P(1/p)$.

Proof.

(a) Using Assumption **(A.2)** we have $\left|\Lambda_{\min}(p^{-1}\mathbf{B}'\mathbf{B}) - \Lambda_{\min}(\breve{\mathbf{B}})\right| \leq \left|\!\left|\!\left| p^{-1}\mathbf{B}'\mathbf{B} - \breve{\mathbf{B}} \right|\!\right|\!\right|_2$, which implies Part (a).

(b) First, notice that $|||\mathbf{\Pi}|||_2 = \Lambda_{\min}(\mathbf{\Theta}_f + (\mathbf{BH}')'\mathbf{\Theta}_\varepsilon(\mathbf{BH}'))^{-1}$. Therefore, we get

$$\begin{aligned} |||\mathbf{\Pi}|||_2 &\leq \Lambda_{\min}((\mathbf{BH}')'\mathbf{\Theta}_\varepsilon(\mathbf{BH}'))^{-1} \leq \Lambda_{\min}(\mathbf{B}'\mathbf{B})^{-1}\Lambda_{\min}(\mathbf{\Theta}_\varepsilon)^{-1} \\ &= \Lambda_{\min}(\mathbf{B}'\mathbf{B})^{-1}\Lambda_{\max}(\mathbf{\Sigma}_\varepsilon), \end{aligned}$$

where the second inequality is due to the fact that the linear space spanned by the rows of $\mathbf{B}$ is the same as that by the rows of $\mathbf{BH}'$, hence, in practice, it does not matter which one is used. Therefore, the result in Part (b) follows from Part (a), Assumptions **(A.1)** and **(A.2)**.

(c) From Lemma 7 we obtained:

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{Hf}_t(\mathbf{Hf}_t)' - \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t'\right\|_F = \mathcal{O}_P\left(\frac{K^{3/2}}{\sqrt{T}} + \frac{K^{5/2}}{\sqrt{p}}\right).$$

Since $\left|\!\left|\!\left|\mathbf{\Theta}_f(\widehat{\mathbf{\Sigma}}_f - \mathbf{\Sigma}_f)\right|\!\right|\!\right|_2 < 1$, we have

$$\left|\!\left|\!\left|\widehat{\mathbf{\Theta}}_f - \mathbf{\Theta}_f\right|\!\right|\!\right|_2 \leq \frac{|||\mathbf{\Theta}_f|||_2\left|\!\left|\!\left|\mathbf{\Theta}_f(\widehat{\mathbf{\Sigma}}_f - \mathbf{\Sigma}_f)\right|\!\right|\!\right|_2}{1 - \left|\!\left|\!\left|\mathbf{\Theta}_f(\widehat{\mathbf{\Sigma}}_f - \mathbf{\Sigma}_f)\right|\!\right|\!\right|_2} = \mathcal{O}_P\left(\frac{K^{3/2}}{\sqrt{T}} + \frac{K^{5/2}}{\sqrt{p}}\right).$$

Let $\omega_{4T} = K^{3/2}/\sqrt{T} + K^{5/2}/\sqrt{p}$. Using the definition of ϱ_T from Theorem 2, it follows that $\varrho_T^{-1}\omega_{4T} = o_P(\omega_{4T}\omega_{3T}^{-1})$. Let $\widetilde{\gamma}_T \equiv \omega_{4T}\omega_{3T}^{-1}$. Consider $\varrho_T^{-1}\left|\!\left|\!\left|\widehat{\mathbf{\Theta}}_f - \mathbf{\Theta}_f\right|\!\right|\!\right|_2 = o_P(\gamma_T)$. The latter holds for any $\gamma_t \geq \widetilde{\gamma}_T$, with the tightest bound obtained when $\gamma_T = \widetilde{\gamma}_T$. For the ease of representation, we use $\gamma_T = 1/\sqrt{K}$ instead of $\widetilde{\gamma}_T$.

(d) We will bound each term in the definition of $\widehat{\mathbf{\Pi}} - \mathbf{\Pi}$. First, we have

$$\begin{aligned}
\left|\!\left|\!\left|\widehat{\mathbf{B}}'\widehat{\mathbf{\Theta}}_\varepsilon\widehat{\mathbf{B}} - (\mathbf{BH}')'\mathbf{\Theta}_\varepsilon(\mathbf{BH}')\right|\!\right|\!\right|_2 &\leq \left|\!\left|\!\left|\widehat{\mathbf{B}} - \mathbf{BH}'\right|\!\right|\!\right|_2 \left|\!\left|\!\left|\widehat{\mathbf{\Theta}}_\varepsilon\right|\!\right|\!\right|_2 \left|\!\left|\!\left|\widehat{\mathbf{B}}\right|\!\right|\!\right|_2 \\
&+ |||\mathbf{BH}'|||_2 \left|\!\left|\!\left|\widehat{\mathbf{\Theta}}_\varepsilon - \mathbf{\Theta}_\varepsilon\right|\!\right|\!\right|_2 \left|\!\left|\!\left|\widehat{\mathbf{B}}\right|\!\right|\!\right|_2 \\
&+ |||\mathbf{BH}'|||_2 |||\mathbf{\Theta}_\varepsilon|||_2 \left|\!\left|\!\left|\widehat{\mathbf{B}} - \mathbf{BH}'\right|\!\right|\!\right|_2 \\
&= \mathcal{O}_P\left(p \cdot s_T \cdot \varrho_T\right). \qquad (2.33)
\end{aligned}$$

Now we combine (2.33) with the results from Parts (b)-(c):

$$\varrho_T^{-1}\left|\!\left|\!\left|\mathbf{\Pi}\left(\widehat{\mathbf{\Pi}}^{-1} - \mathbf{\Pi}^{-1}\right)\right|\!\right|\!\right|_2 = \mathcal{O}_P\left(s_t\right).$$

Finally, since $\left|\!\left|\!\left|\mathbf{\Pi}\left(\widehat{\mathbf{\Pi}}^{-1} - \mathbf{\Pi}^{-1}\right)\right|\!\right|\!\right|_2 < 1$, we have

$$\varrho_T^{-1}\left|\!\left|\!\left|\widehat{\mathbf{\Pi}} - \mathbf{\Pi}\right|\!\right|\!\right|_2 \leq \varrho_T^{-1}\frac{|||\mathbf{\Pi}|||_2\left|\!\left|\!\left|\mathbf{\Pi}\left(\widehat{\mathbf{\Pi}}^{-1} - \mathbf{\Pi}^{-1}\right)\right|\!\right|\!\right|_2}{1 - \left|\!\left|\!\left|\mathbf{\Pi}\left(\widehat{\mathbf{\Pi}}^{-1} - \mathbf{\Pi}^{-1}\right)\right|\!\right|\!\right|_2} = \mathcal{O}_P\left(\frac{s_t}{p}\right).$$

■

2.A.6 Proof of Theorem 3

Using the Sherman-Morrison-Woodbury formula, we have

$$\begin{aligned}
\left\|\left|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\right|\right\|_l &\leq \left\|\left|\widehat{\mathbf{\Theta}}_\varepsilon - \mathbf{\Theta}_\varepsilon\right|\right\|_l + \left\|\left|(\widehat{\mathbf{\Theta}}_\varepsilon - \mathbf{\Theta}_\varepsilon)\widehat{\mathbf{B}}\widehat{\mathbf{\Pi}}\widehat{\mathbf{B}}'\widehat{\mathbf{\Theta}}_\varepsilon\right|\right\|_l + \left\|\left|\mathbf{\Theta}_\varepsilon(\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}')\widehat{\mathbf{\Pi}}\widehat{\mathbf{B}}'\widehat{\mathbf{\Theta}}_\varepsilon\right|\right\|_l \\
&+ \left\|\left|\mathbf{\Theta}_\varepsilon\mathbf{B}\mathbf{H}'(\widehat{\mathbf{\Pi}} - \mathbf{\Pi})\widehat{\mathbf{B}}'\widehat{\mathbf{\Theta}}_\varepsilon\right|\right\|_l + \left\|\left|\mathbf{\Theta}_\varepsilon\mathbf{B}\mathbf{H}'\mathbf{\Pi}(\widehat{\mathbf{B}} - \mathbf{B})'\widehat{\mathbf{\Theta}}_\varepsilon\right|\right\|_l \\
&+ \left\|\left|\mathbf{\Theta}_\varepsilon\mathbf{B}\mathbf{H}'\mathbf{\Pi}(\mathbf{B}\mathbf{H}')'(\widehat{\mathbf{\Theta}}_\varepsilon - \mathbf{\Theta}_\varepsilon)\right|\right\|_l \\
&= \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 + \Delta_5 + \Delta_6. \qquad (2.34)
\end{aligned}$$

We now bound the terms in (2.34) for $l = 2$ and $l = \infty$. We start with $l = 2$. First, note that $\varrho_T^{-1}\Delta_1 = \mathcal{O}_P(s_T)$ by Theorem 2. Second, using Lemmas 8-9 together with Theorem 2, we have $\varrho_T^{-1}(\Delta_2 + \Delta_6) = \mathcal{O}_P(s_T \cdot \sqrt{p} \cdot (1/p) \cdot \sqrt{p} \cdot 1) = \mathcal{O}_P(s_T)$. Third, $\varrho_T^{-1}(\Delta_3 + \Delta_5)$ is negligible according to Lemma 8(c). Finally, $\varrho_T^{-1}\Delta_4 = \mathcal{O}_P\Big(1 \cdot \sqrt{p} \cdot (s_T/p) \cdot \sqrt{p} \cdot 1\Big) = \mathcal{O}_P(s_T)$ by Lemmas 8-9 and Theorem 2.

Now consider $l = \infty$. First, similarly to the previous case, $\varrho_T^{-1}\Delta_1 = \mathcal{O}_P(s_T)$. Second, $\varrho_T^{-1}(\Delta_2 + \Delta_6) = \mathcal{O}_P\Big(s_T \cdot \sqrt{pK} \cdot (\sqrt{K}/p) \cdot \sqrt{pK} \cdot \sqrt{d_T}\Big) = \mathcal{O}_P(s_T K^{3/2}\sqrt{d_T})$, where we used the fact that for any $\mathbf{A} \in \mathcal{S}_p$ we have $|||\mathbf{A}|||_1 = |||\mathbf{A}|||_\infty \leq \sqrt{d(\mathbf{A})}|||\mathbf{A}|||_2$, where $d(\mathbf{A})$ measures the maximum vertex degree as described at the beginning of Section 4. Third, the term $\varrho_T^{-1}(\Delta_3 + \Delta_5)$ is negligible according to Lemma 8(c). Finally, $\varrho_T^{-1}\Delta_4 = \mathcal{O}_P(\sqrt{d_T} \cdot \sqrt{pK} \cdot \sqrt{K}(s_T)/p \cdot \sqrt{pK} \cdot \sqrt{d_T}) = \mathcal{O}_P(d_T K^{3/2} s_T)$.

2.A.7 Lemmas for Theorem 4

Lemma 10 *Under the assumptions of Theorem 4,*

(a) $\|\widehat{\mathbf{m}}-\mathbf{m}\|_{max}=\mathcal{O}_P(\sqrt{\log p/T})$*, where* $\mathbf{m}$ *is the unconditional mean of stock returns defined in Subsection 3.3, and* $\widehat{\mathbf{m}}$ *is the sample mean.*

(b) $|||\mathbf{\Theta}|||_1=\mathcal{O}(d_T K^{3/2})$*, where* d_T *was defined in Section 4.*

Proof.

(a) The proof of Part (a) is provided in Chang et al. (2018) (Lemma 1).

(b) To prove Part (b) we use the Sherman-Morrison-Woodbury formula:

$$\begin{aligned}|||\mathbf{\Theta}|||_1 &\leq |||\mathbf{\Theta}_\varepsilon|||_1+\Big|\Big|\Big|\mathbf{\Theta}_\varepsilon\mathbf{B}[\mathbf{\Theta}_f+\mathbf{B}'\mathbf{\Theta}_\varepsilon\mathbf{B}]^{-1}\mathbf{B}'\mathbf{\Theta}_\varepsilon\Big|\Big|\Big|_1\\ &=\mathcal{O}(\sqrt{d_T})+\mathcal{O}\Big(\sqrt{d_T}\cdot p\cdot\frac{\sqrt{K}}{p}\cdot K\cdot\sqrt{d_T}\Big)=\mathcal{O}(d_T K^{3/2}).\end{aligned}\tag{2.35}$$

The last equality in (2.35) is obtained under the assumptions of Theorem 4. This result is important in several aspects: it shows that the sparsity of the precision matrix of stock returns is controlled by the sparsity in the precision of the idiosyncratic returns. Hence, one does not need to impose an unrealistic sparsity assumption on the precision of returns a priori when the latter follow a factor structure - sparsity of the precision once the common movements have been taken into account would suffice.

■

Lemma 11 *Define* $a = \boldsymbol{\iota}_p'\boldsymbol{\Theta}\boldsymbol{\iota}_p/p$, $b = \boldsymbol{\iota}_p'\boldsymbol{\Theta}\mathbf{m}/p$, $d = \mathbf{m}'\boldsymbol{\Theta}\mathbf{m}/p$, $g = \sqrt{\mathbf{m}'\boldsymbol{\Theta}\mathbf{m}}/p$ *and* $\widehat{a} = \boldsymbol{\iota}_p'\widehat{\boldsymbol{\Theta}}\boldsymbol{\iota}_p/p$, $\widehat{b} = \boldsymbol{\iota}_p'\widehat{\boldsymbol{\Theta}}\widehat{\mathbf{m}}/p$, $\widehat{d} = \widehat{\mathbf{m}}'\widehat{\boldsymbol{\Theta}}\widehat{\mathbf{m}}/p$, $\widehat{g} = \sqrt{\widehat{\mathbf{m}}'\widehat{\boldsymbol{\Theta}}\widehat{\mathbf{m}}}/p$ *. Under the assumptions of Theorem 4 and assuming* $(ad - b^2) > 0$,

(a) $a \geq C_0 > 0$, $b = \mathcal{O}(1)$, $d = \mathcal{O}(1)$, *where* C_0 *is a positive constant representing the minimal eigenvalue of* $\boldsymbol{\Theta}$.

(b) $|\widehat{a} - a| = \mathcal{O}_P(\varrho_T d_T K^{3/2} s_T) = o_P(1)$.

(c) $\left|\widehat{b} - b\right| = \mathcal{O}_P(\varrho_T d_T K^{3/2} s_T) = o_P(1)$

(d) $\left|\widehat{d} - d\right| = \mathcal{O}_P(\varrho_T d_T K^{3/2} s_T) = o_P(1)$.

(e) $|\widehat{g} - g| = \mathcal{O}_P\Big([\varrho_T d_T K^{3/2} s_T]^{1/2}\Big) = o_P(1)$.

(f) $\left|(\widehat{a}\widehat{d} - \widehat{b}^2) - (ad - b^2)\right| = \mathcal{O}_P\Big(\varrho_T d_T K^{3/2} s_T\Big) = o_P(1)$.

(g) $\left|ad - b^2\right| = \mathcal{O}(1)$.

Proof.

(a) Part (a) is trivial and follows directly from $|||\boldsymbol{\Theta}|||_2 = \mathcal{O}(1)$ and $\|\mathbf{m}\|_\infty = \mathcal{O}(1)$ from Assumption **(B.1)**. We show the proof for d: recall, $d = \mathbf{m}'\boldsymbol{\Theta}\mathbf{m}/p \leq |||\boldsymbol{\Theta}|||_2^2 \|\mathbf{m}\|_2^2/p = \mathcal{O}(1)$.

(b) Using the Hölders inequality, we have

$$|\widehat{a} - a| = \left|\frac{\boldsymbol{\iota}_p'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\boldsymbol{\iota}_p}{p}\right| \leq \frac{\left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\boldsymbol{\iota}_p\right\|_1 \|\boldsymbol{\iota}_p\|_{\max}}{p} \leq \left|\left|\left|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right|\right|\right|_1$$
$$= \mathcal{O}_P\Big(\varrho_T d_T K^{3/2}(s_T + (1/p))\Big) = o_P(1),$$

where the last rate is obtained using the assumptions of Theorem 3.

(c) First, rewrite the expression of interest:

$$\widehat{b} - b = [\boldsymbol{\iota}_p'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})]/p + [\boldsymbol{\iota}_p'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\mathbf{m}]/p + [\boldsymbol{\iota}_p'\boldsymbol{\Theta}(\widehat{\mathbf{m}} - \mathbf{m})]/p. \tag{2.36}$$

We now bound each of the terms in (2.36) using the expressions derived in Callot et al. (2019) (see their Proof of Lemma A.2) and the fact that $\log p/T = o(1)$.

$$\left|\boldsymbol{\iota}_p'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})\right|/p \leq \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta} \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_1 \|\widehat{\mathbf{m}} - \mathbf{m}\|_{\max} = \mathcal{O}_P\Big(\varrho_T d_T K^{3/2} s_T \cdot \sqrt{\frac{\log p}{T}}\Big). \tag{2.37}$$

$$\left|\boldsymbol{\iota}_p'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\mathbf{m}\right|/p \leq \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta} \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_1 = \mathcal{O}_P\Big(\varrho_T d_T K^{3/2} s_T\Big). \tag{2.38}$$

$$\left|\boldsymbol{\iota}_p'\boldsymbol{\Theta}(\widehat{\mathbf{m}} - \mathbf{m})\right|/p \leq |||\boldsymbol{\Theta}|||_1 \|\widehat{\mathbf{m}} - \mathbf{m}\|_{\max} = \mathcal{O}_P\Big(d_T K^{3/2} \cdot \sqrt{\frac{\log p}{T}}\Big). \tag{2.39}$$

(d) First, rewrite the expression of interest:

$$\begin{aligned}\widehat{d} - d = {} & [(\widehat{\mathbf{m}} - \mathbf{m})'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})]/p + [(\widehat{\mathbf{m}} - \mathbf{m})'\boldsymbol{\Theta}(\widehat{\mathbf{m}} - \mathbf{m})]/p \\ & + [2(\widehat{\mathbf{m}} - \mathbf{m})'\boldsymbol{\Theta}\mathbf{m}]/p + [2\mathbf{m}'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})]/p \\ & + [\mathbf{m}'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\mathbf{m}]/p.\end{aligned} \tag{2.40}$$

We now bound each of the terms in (2.40) using the expressions derived in Callot et al. (2019) (see their Proof of Lemma A.3) and the fact that $\log p/T = o(1)$.

$$\begin{aligned}\Big|(\widehat{\mathbf{m}}-\mathbf{m})'(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta})(\widehat{\mathbf{m}}-\mathbf{m})\Big|/p &\leq \|\widehat{\mathbf{m}}-\mathbf{m}\|_{\max}^2 \Big|\Big|\Big|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\Big|\Big|\Big|_1 \\ &= \mathcal{O}_P\Big(\frac{\log p}{T}\cdot \varrho_T d_T K^{3/2} s_T\Big) \end{aligned} \tag{2.41}$$

$$\big|(\widehat{\mathbf{m}}-\mathbf{m})'\boldsymbol{\Theta}(\widehat{\mathbf{m}}-\mathbf{m})\big|/p \leq \|\widehat{\mathbf{m}}-\mathbf{m}\|_{\max}^2 |||\boldsymbol{\Theta}|||_1 = \mathcal{O}_P\Big(\frac{\log p}{T}\cdot d_T K^{3/2}\Big). \tag{2.42}$$

$$\big|(\widehat{\mathbf{m}}-\mathbf{m})'\boldsymbol{\Theta}\mathbf{m}\big|/p \leq \|\widehat{\mathbf{m}}-\mathbf{m}\|_{\max} |||\boldsymbol{\Theta}|||_1 = \mathcal{O}_P\Big(\sqrt{\frac{\log p}{T}}\cdot d_T K^{3/2}\Big). \tag{2.43}$$

$$\begin{aligned}\Big|\mathbf{m}'(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta})(\widehat{\mathbf{m}}-\mathbf{m})\Big|/p &\leq \|\widehat{\mathbf{m}}-\mathbf{m}\|_{\max} \Big|\Big|\Big|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\Big|\Big|\Big|_1 \\ &= \mathcal{O}_P\Big(\sqrt{\frac{\log p}{T}}\cdot \varrho_T d_T K^{3/2} s_T\Big). \end{aligned} \tag{2.44}$$

$$\Big|\mathbf{m}'(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta})\mathbf{m}\Big|/p \leq \Big|\Big|\Big|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\Big|\Big|\Big|_1 = \mathcal{O}_P\Big(\varrho_T d_T K^{3/2} s_T\Big). \tag{2.45}$$

(e) This is a direct consequence of Part (d) and the fact that $\sqrt{\widehat{d}-d} \geq \sqrt{\widehat{d}}-\sqrt{d}$.

(f) First, rewrite the expression of interest:

$$(\widehat{a}\widehat{d}-\widehat{b}^2)-(ad-b^2)=[(\widehat{a}-a)+a][(\widehat{d}-d)+d]-[(\widehat{b}-b)+b]^2,$$

therefore, using Lemma 11, we have

$$\begin{aligned}\left|(\widehat{a}\widehat{d}-\widehat{b}^2)-(ad-b^2)\right| &\leq \left[|\widehat{a}-a|\left|\widehat{d}-d\right|+|\widehat{a}-a|d+a\left|\widehat{d}-d\right|+(\widehat{b}-b)^2+2|b|\left|\widehat{b}-b\right|\right]\\ &=\mathcal{O}_P\left(\varrho_T d_T K^{3/2}s_T\right)=o_P(1).\end{aligned}$$

(g) This is a direct consequence of Part (a): $ad-b^2\leq ad=\mathcal{O}(1)$.

■

2.A.8 Proof of Theorem 4

Let us derive convergence rates for each portfolio weight formulas one by one. We start with GMV formulation.

$$\|\widehat{\mathbf{w}}_{\text{GMV}}-\mathbf{w}_{\text{GMV}}\|_1\leq\frac{a\frac{\left\|(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta})\boldsymbol{\iota}_p\right\|_1}{p}+|a-\widehat{a}|\frac{\|\boldsymbol{\Theta}\boldsymbol{\iota}_p\|_1}{p}}{|\widehat{a}|a}=\mathcal{O}_P\left(\varrho_T d_T^2 K^3 s_T\right)=o_P(1),$$

where the first inequality was shown in Callot et al. (2019) (see their expression A.50), and the rate follows from Lemmas 11 and 10.

We now proceed with the MWC weight formulation. First, let us simplify the weight expression as follows: $\mathbf{w}_{\text{MWC}} = \kappa_1(\boldsymbol{\Theta}\boldsymbol{\iota}_p/p) + \kappa_2(\boldsymbol{\Theta}\mathbf{m}/p)$, where

$$\kappa_1 = \frac{d - \mu b}{ad - b^2}$$
$$\kappa_2 = \frac{\mu a - b}{ad - b^2}.$$

Let $\widehat{\mathbf{w}}_{\text{MWC}} = \widehat{\kappa}_1(\widehat{\boldsymbol{\Theta}}\boldsymbol{\iota}_p/p) + \widehat{\kappa}_2(\widehat{\boldsymbol{\Theta}}\widehat{\mathbf{m}}/p)$, where $\widehat{\kappa}_1$ and $\widehat{\kappa}_2$ are the estimators of κ_1 and κ_2 respectively. As shown in Callot et al. (2019) (see their equation A.57), we can bound the quantity of interest as follows:

$$\begin{aligned}\|\widehat{\mathbf{w}}_{\text{MWC}} - \mathbf{w}_{\text{MWC}}\|_1 \leq\ & |(\widehat{\kappa}_1 - \kappa_1)| \left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\boldsymbol{\iota}_p\right\|_1/p + |(\widehat{\kappa}_1 - \kappa_1)| \|\boldsymbol{\Theta}\boldsymbol{\iota}_p\|_1/p \\ & + |\kappa_1| \left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\boldsymbol{\iota}_p\right\|_1/p \\ & + |(\widehat{\kappa}_2 - \kappa_2)| \left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})\right\|_1/p + |(\widehat{\kappa}_2 - \kappa_2)| \|\boldsymbol{\Theta}(\widehat{\mathbf{m}} - \mathbf{m})\|_1/p \\ & + |(\widehat{\kappa}_2 - \kappa_2)| \left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\mathbf{m}\right\|_1/p + |(\widehat{\kappa}_2 - \kappa_2)| \|\boldsymbol{\Theta}\mathbf{m}\|_1/p \\ & + |\kappa_2| \left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})\right\|_1/p + |\kappa_2| \left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\mathbf{m}\right\|_1/p. \end{aligned} \tag{2.46}$$

For the ease of representation, denote $y = ad - b^2$. Then, using similar technique as in Callot et al. (2019) we get

$$|(\widehat{\kappa}_1 - \kappa_1)| \leq \frac{y\left|\widehat{d} - d\right| + y\mu\left|\widehat{b} - b\right| + |\widehat{y} - y||d - \mu b|}{\widehat{y}y} = \mathcal{O}_P\left(\varrho_T d_T K^{3/2} s_T\right) = o_P(1),$$

where the rate trivially follows from Lemma 11.

Similarly, we get

$$|(\widehat{\kappa}_2 - \kappa_2)| = \mathcal{O}_P\left(\varrho_T d_T K^{3/2} s_T\right) = o_P(1).$$

Callot et al. (2019) showed that $|\kappa_1| = \mathcal{O}(1)$ and $|\kappa_2| = \mathcal{O}(1)$. Therefore, we can get the rate of (2.46):

$$\|\widehat{\mathbf{w}}_{\text{MWC}} - \mathbf{w}_{\text{MWC}}\|_1 = \mathcal{O}_P\left(\varrho_T d_T^2 K^3 s_T\right) = o_P(1).$$

We now proceed with the MRC weight formulation:

$$\begin{aligned}
&\|\widehat{\mathbf{w}}_{\text{MRC}} - \mathbf{w}_{\text{MRC}}\|_1 \le \frac{\frac{g}{p}\left[\left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})\right\|_1 + \left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\mathbf{m}\right\|_1 + \|\boldsymbol{\Theta}(\widehat{\mathbf{m}} - \mathbf{m})\|_1\right]}{|\widehat{g}|g} \\
&+ \frac{|\widehat{g} - g|\|\boldsymbol{\Theta}\mathbf{m}\|_1}{|\widehat{g}|g} \\
&\le \frac{\frac{g}{p}\left[p\left\vert\!\left\vert\!\left\vert\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right\vert\!\right\vert\!\right\vert_1 \|(\widehat{\mathbf{m}} - \mathbf{m})\|_{\max} + p\left\vert\!\left\vert\!\left\vert\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right\vert\!\right\vert\!\right\vert_1 \|\mathbf{m}\|_{\max} + p\vert\!\vert\!\vert\boldsymbol{\Theta}\vert\!\vert\!\vert_1 \|(\widehat{\mathbf{m}} - \mathbf{m})\|_{\max}\right.}{|\widehat{g}|g} \\
&\frac{\left. + p|\widehat{g} - g|\vert\!\vert\!\vert\boldsymbol{\Theta}\vert\!\vert\!\vert_1\|\mathbf{m}\|_{\max}\right]}{|\widehat{g}|g} = \mathcal{O}_P\left(\varrho_T d_T K^{3/2} s_T \cdot \sqrt{\frac{\log p}{T}}\right) + \mathcal{O}_P\left(\varrho_T d_T K^{3/2} s_T\right) \\
&+ \mathcal{O}_P\left(d_T K^{3/2} \cdot \sqrt{\frac{\log p}{T}}\right) + \mathcal{O}_P\left([\varrho_T d_T K^{3/2} s_T]^{1/2} \cdot d_T K^{3/2}\right) = o_P(1),
\end{aligned}$$

where we used Lemmas 10-11.

2.A.9 Proof of Theorem 5

We start with the GMV formulation. Using Lemma 11 (a)-(b), we get

$$\left|\frac{\hat{a}^{-1}}{a^{-1}}-1\right|=\frac{|a-\hat{a}|}{|\hat{a}|}=\mathcal{O}_P(\varrho_T d_T K^{3/2}s_T)=o_P(1).$$

Proceeding to the MWC risk exposure, we follow Callot et al. (2019) and introduce the following notation: $x=a\mu^2-2b\mu+d$ and $\hat{x}=\hat{a}\mu-2\hat{b}\mu+\hat{d}$ to rewrite $\widehat{\Phi}_{\text{MWC}}=p^{-1}(\hat{x}/\hat{y})$. As shown in Callot et al. (2019), $y/x=\mathcal{O}(1)$ (see their equation A.42). Furthermore, by Lemma 11 (b)-(d)

$$|\hat{x}-x|\leq|\hat{a}-a|\mu^2+2\left|\hat{b}-b\right|\mu+\left|\hat{d}-d\right|=\mathcal{O}_P(\varrho_T d_T K^{3/2}s_T)=o_P(1),$$

and by Lemma 11 (f):

$$|\hat{y}-y|=\left|\hat{a}\hat{d}-\hat{b}^2-(ad-b^2)\right|=\mathcal{O}_P(\varrho_T d_T K^{3/2}s_T)=o_P(1).$$

Using the above and the facts that $y=\mathcal{O}(1)$ and $x=\mathcal{O}(1)$ (which were derived by Callot et al. (2019) in A.45 and A.46), we have

$$\left|\frac{\widehat{\Phi}_{\text{MWC}}-\Phi_{\text{MWC}}}{\Phi_{\text{MWC}}}\right|=\left|\frac{(\hat{x}-x)y+x(y-\hat{y})}{\hat{y}y}\right|\mathcal{O}(1)\mathcal{O}_P(\varrho_T d_T K^{3/2}s_T)=o_P(1).$$

Finally, to bound MRC risk exposure, we use Lemma 11 (e) and rewrite

$$\frac{|g-\hat{g}|}{|\hat{g}|}=\mathcal{O}_P\Big([\varrho_T d_T K^{3/2}s_T]^{1/2})=o_P(1).$$

2.B Additional Simulations

2.B.1 Verifying Theoretical Rates

To compare the empirical rate with the theoretical expressions derived in Theorems 3-5, we use the facts from Theorem 2 that $\omega_{3T} \equiv K^2\sqrt{\log p/T} + K^3/\sqrt{p}$ and $\varrho_T^{-1}\omega_{3T} \xrightarrow{p} 0$ to introduce the following functions that correspond to the theoretical rates for the choice of parameters in the empirical setting:

$$\left.\begin{aligned} f_{\vert\kern-0.25ex\vert\kern-0.25ex\vert\cdot\vert\kern-0.25ex\vert\kern-0.25ex\vert_2} &= C_1 + C_2 \cdot \log_2(s_T\varrho_T) \\ g_{\vert\kern-0.25ex\vert\kern-0.25ex\vert\cdot\vert\kern-0.25ex\vert\kern-0.25ex\vert_1} &= C_3 + C_2 \cdot \log_2(d_T K^{3/2} s_T \varrho_T) \end{aligned}\right\} \text{for } \widehat{\boldsymbol{\Theta}} \tag{2.47}$$

$$h_1 = C_4 + C_2 \cdot \log_2(\varrho_T d_T^2 K^3 s_T) \qquad \text{for } \widehat{\mathbf{w}}_{\text{GMV}}, \widehat{\mathbf{w}}_{\text{MWC}} \tag{2.48}$$

$$h_2 = C_5 + C_6 \cdot \log_2([\varrho_T s_T]^{1/2} d_T^{3/2} K^3) \quad \text{for } \widehat{\mathbf{w}}_{\text{MRC}} \tag{2.49}$$

$$h_3 = C_7 + C_2 \cdot \log_2(d_T K^{3/2} s_T \varrho_T) \qquad \text{for } \widehat{\Phi}_{\text{GMV}}, \widehat{\Phi}_{\text{MWC}} \tag{2.50}$$

$$h_4 = C_8 + C_9 \cdot \log_2(d_T K^{3/2} s_T \varrho_T) \qquad \text{for } \widehat{\Phi}_{\text{MRC}} \tag{2.51}$$

where $C_1, \ldots, C_9$ are constants with $C_6 > C_2$ (by Theorem 4), $C_9 > C_2$ (by Theorem 5).

Figure 2.B.1 shows the averaged (over Monte Carlo simulations) errors of the estimators of $\boldsymbol{\Theta}$, $\mathbf{w}$ and Φ versus the sample size T in the logarithmic scale (base 2). In order to confirm the theoretical findings from Theorems 3-5, we also plot the theoretical rates of convergence given by the functions in (2.47)-(2.51). We verify that the empirical and theoretical rates are matched. Since the convergence rates for GMV and MWC portfolio weights $\mathbf{w}$ and risk exposures Φ are very similar, we only report the former. Note that as predicted by Theorem 3, the rate of convergence of the precision matrix in $\vert\kern-0.25ex\vert\kern-0.25ex\vert\cdot\vert\kern-0.25ex\vert\kern-0.25ex\vert_2$-norm is faster than the rate in $\|\cdot\|_1$-norm. Furthermore, the convergence rate of the GMV, MWC and MRC portfolio weights and risk exposures are close to the rate of the precision matrix $\mathbf{\Theta}$ in $\|\cdot\|_1$-norm, which is confirmed by Theorem 4. As evidenced by Figure 2.B.1, the convergence rate of the MRC risk exposure is slower than the rate of GMV and MWC exposures. This finding is in accordance with Theorem 5 and it is also consistent with the empirical findings that indicate higher overall risk associated with MRC portfolios.

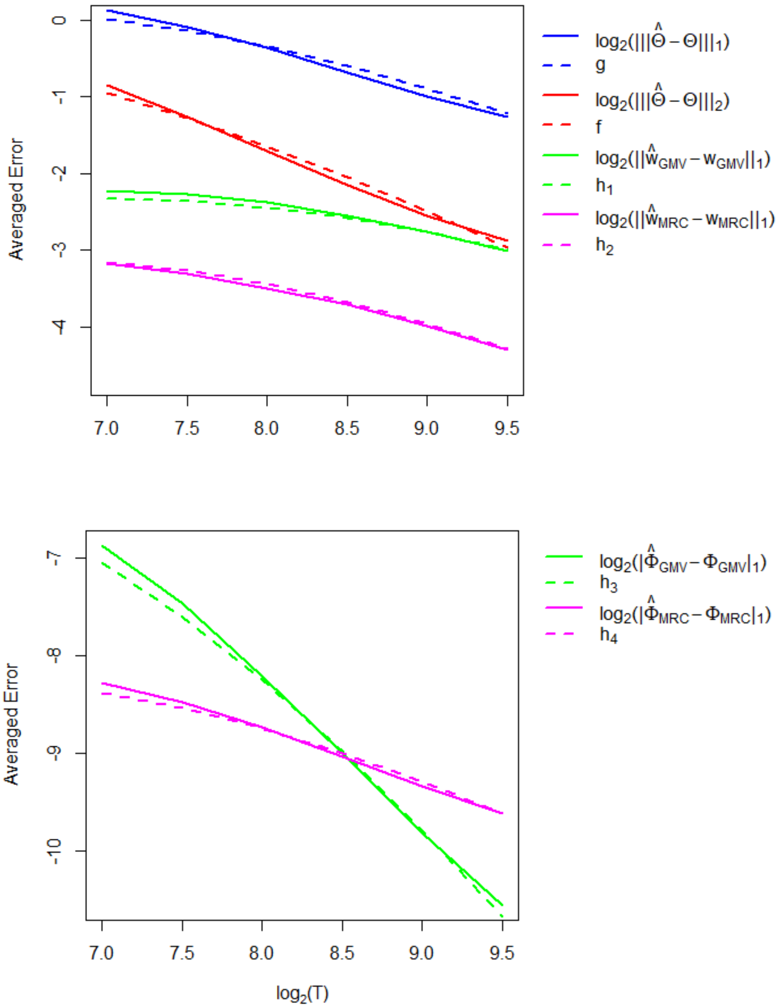

Figure 2.B.1: **Averaged empirical errors (solid lines) and theoretical rates of convergence (dashed lines) on logarithmic scale: $p = T^{0.85}$, $K = 2(\log T)^{0.5}$, $s_T = \mathcal{O}(T^{0.05})$.**

2.B.2 Results for Case 1

We compare the performance of FGL with the alternative models listed at the beginning of Section 5 for Case 1. The only instance when FGL is strictly but slightly dominated occurs in Figure 2.B.2: POET outperforms FGL in terms of convergence of precision matrix in the spectral norm. This is different from Case 2 in Figure 2.1 where FGL outperforms all the competing models.

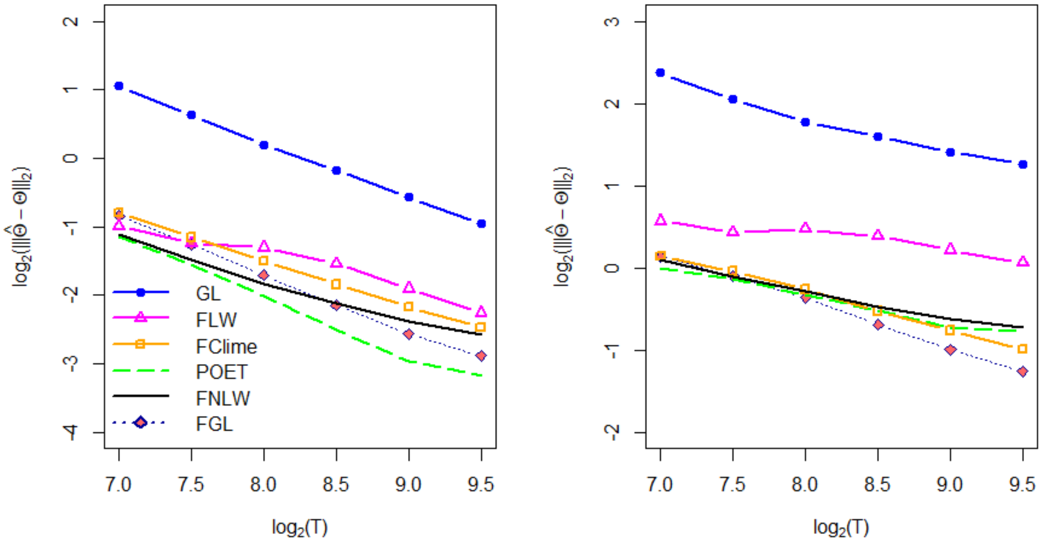

Figure 2.B.2: **Averaged errors of the estimators of Θ for Case 1 on logarithmic scale:** $p = T^{0.85}$**,** $K = 2(\log T)^{0.5}$**,** $s_T = \mathcal{O}(T^{0.05})$**.**

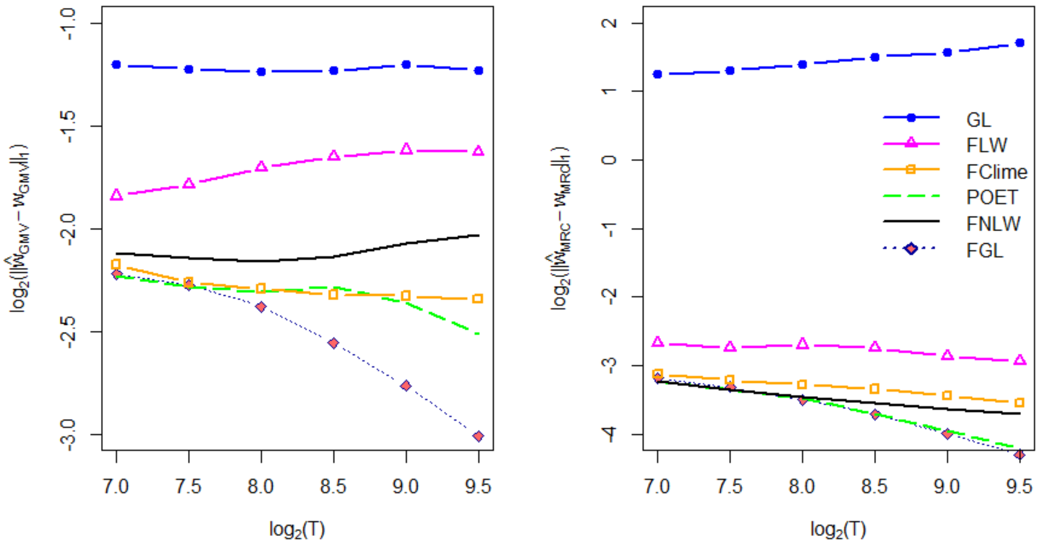

Figure 2.B.3: **Averaged errors of the estimators of $\mathbf{w}_{\mathbf{GMV}}$ (left) and $\mathbf{w}_{\mathbf{MRC}}$ (right) for Case 1 on logarithmic scale: $p = T^{0.85}$, $K = 2(\log T)^{0.5}$, $s_T = \mathcal{O}(T^{0.05})$.**

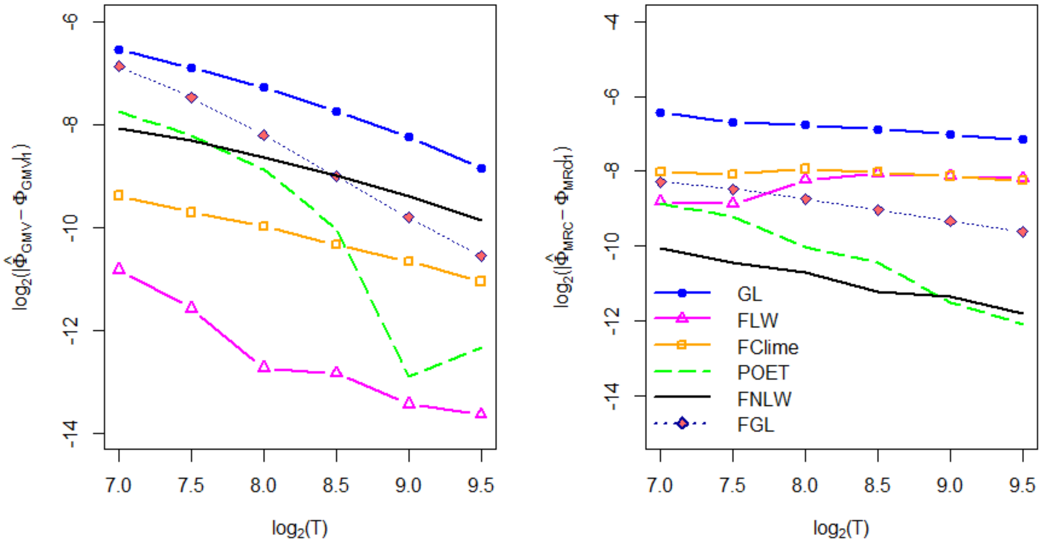

Figure 2.B.4: **Averaged errors of the estimators of $\Phi_{\mathbf{GMV}}$ (left) and $\Phi_{\mathbf{MRC}}$ (right) for Case 1 on logarithmic scale: $p = T^{0.85}$, $K = 2(\log T)^{0.5}$, $s_T = \mathcal{O}(T^{0.05})$.**

2.B.3 Robust FGL

The DGP for elliptical distributions is similar to [56]: let $(\mathbf{f}_t, \boldsymbol{\varepsilon}_t)$ from (2.11) jointly follow the multivariate t-distribution with the degrees of freedom ν. When $\nu = \infty$, this corresponds to the multivariate normal distribution, smaller values of ν are associated with thicker tails. We draw T independent samples of $(\mathbf{f}_t, \boldsymbol{\varepsilon}_t)$ from the multivariate t-distribution with zero mean and covariance matrix $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\Sigma}_f, \boldsymbol{\Sigma}_\varepsilon)$, where $\boldsymbol{\Sigma}_f = \mathbf{I}_K$. To construct $\boldsymbol{\Sigma}_\varepsilon$ we use a Toeplitz structure parameterized by $\rho = 0.5$, which leads to the sparse $\boldsymbol{\Theta}_\varepsilon = \boldsymbol{\Sigma}_\varepsilon^{-1}$. The rows of $\mathbf{B}$ are drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I}_K)$. We let $p = T^{0.85}$, $K = 2(\log T)^{0.5}$ and $T = [2^h]$, for $h \in \{7, 7.5, 8, \ldots, 9.5\}$. Figure 2.B.5-2.B.6 report the averaged (over Monte Carlo simulations) estimation errors (in the logarithmic scale, base 2) for $\boldsymbol{\Theta}$ and two portfolio weights (GMV and MRC) using FGL and Robust FGL for $\nu = 4.2$. Noticeably, the performance of FGL for estimating the precision matrix is comparable with that of Robust FGL: this suggests that our FGL algorithm is insensitive to heavy-tailed distributions even without additional modifications. Furthermore, FGL outperforms its Robust counterpart in terms of estimating portfolio weights, as evidenced by Figure 2.B.6. We further compare the performance of FGL and Robust FGL for different degrees of freedom: Figure 2.B.7 reports the log-ratios (base 2) of the averaged (over Monte Carlo simulations) estimation errors for $\nu = 4.2$, $\nu = 7$ and $\nu = \infty$. The results for the estimation of $\boldsymbol{\Theta}$ presented in Figure 2.B.7 are consistent with the findings in [56]: Robust FGL outperforms the non-robust counterpart for thicker tails.

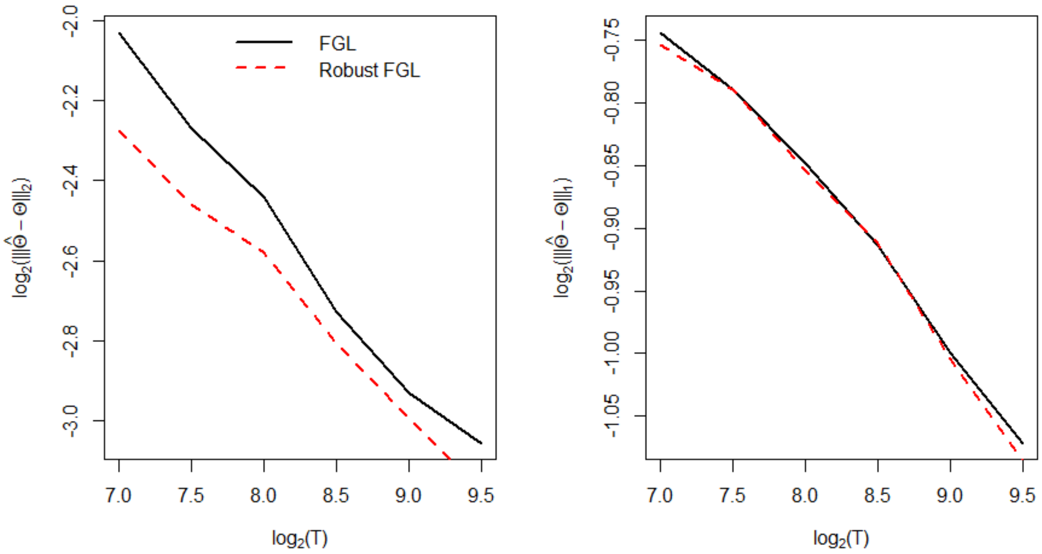

Figure 2.B.5: **Averaged errors of the estimators of $\boldsymbol{\Theta}$ on logarithmic scale: $p = T^{0.85}$, $K = 2(\log T)^{0.5}$, $\nu = 4.2$.**

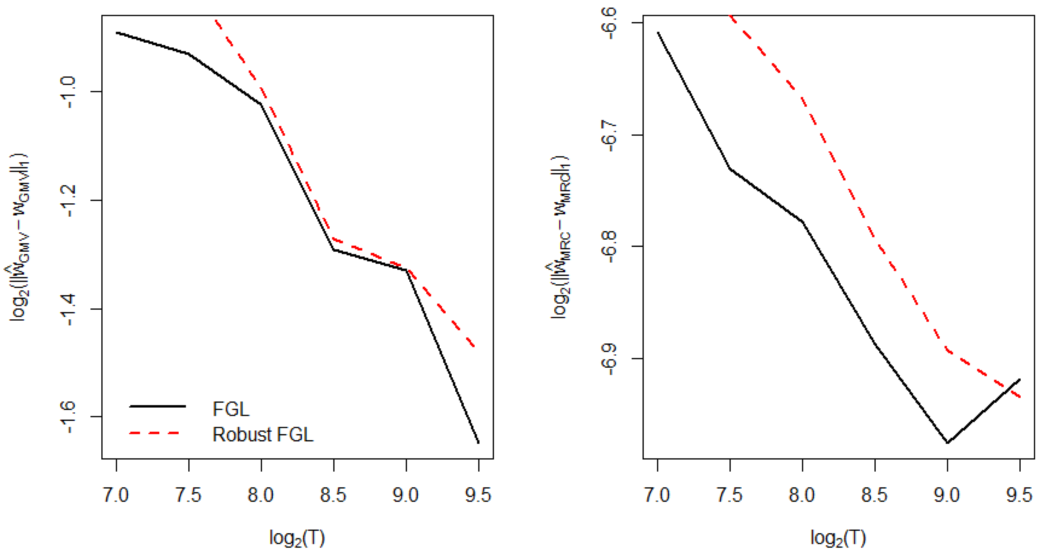

Figure 2.B.6: **Averaged errors of the estimators of $\mathbf{w}_{\mathrm{GMV}}$ (left) and $\mathbf{w}_{\mathrm{MRC}}$ (right) on logarithmic scale: $p = T^{0.85}$, $K = 2(\log T)^{0.5}$, $\nu = 4.2$.**

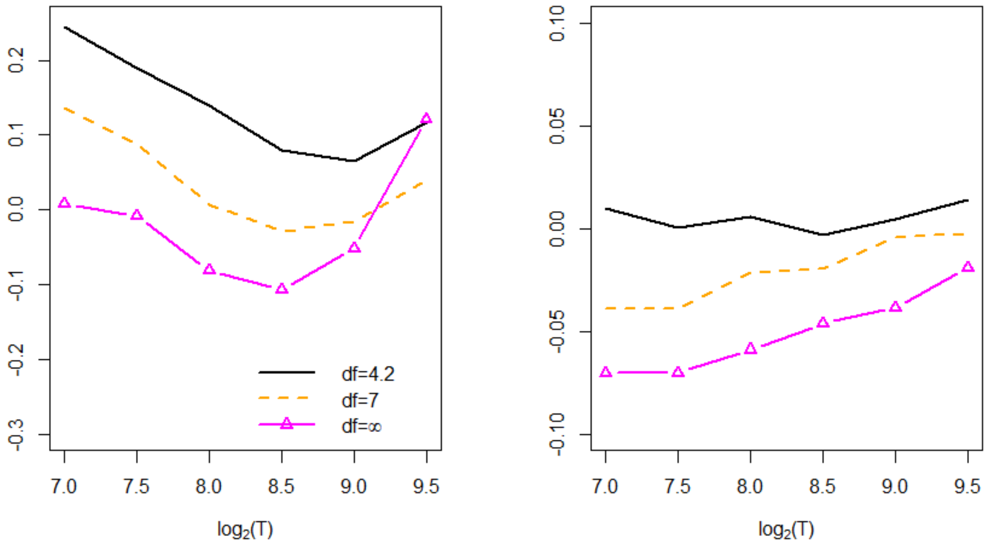

Figure 2.B.7: **Log ratios (base 2) of the averaged errors of the FGL and the Robust FGL estimators of $\boldsymbol{\Theta}$:** $\log_2\left(\frac{|||\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}|||_2}{|||\widehat{\boldsymbol{\Theta}}_{\mathbf{R}}-\boldsymbol{\Theta}|||_2}\right)$ **(left),** $\log_2\left(\frac{|||\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}|||_1}{|||\widehat{\boldsymbol{\Theta}}_{\mathbf{R}}-\boldsymbol{\Theta}|||_1}\right)$ **(right):** $p = T^{0.85}$**,** $K = 2(\log T)^{0.5}$**.**

2.B.4 Relaxing Pervasiveness Assumption

As pointed out by [113], the data on 100 industrial portfolios shows that there are no large gaps between eigenvalues i and $i+1$ of the sample covariance data except for $i = 1$. However, as is commonly believed, such data contains at least three factors. Therefore, the factor pervasiveness assumption suggests the existence of a large gap for $i \geq 3$. In order to examine sensitivity of portfolios to the pervasiveness assumption and quantify the degree of pervasiveness, we use the same DGP as in (2.23)-(2.24), but with $\sigma_{\varepsilon,ij} = \rho^{|i-j|}$ and $K = 3$. We consider $\rho \in \{0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ which corresponds to $\lambda_3/\lambda_4 \in \{3.1, 2.7, 2.6, 2.2, 1.5, 1.1\}$. In other words, as ρ increases, the systematic-idiosyncratic gap measured by $\hat{\lambda}_3/\hat{\lambda}_4$ decreases. Table 2.B.1-2.B.2 report the mean quality of the estimators for portfolio weights and risk over 100 replications for $T = 300$ and $p \in \{300, 400\}$. The sample size and the number of regressors are chosen to closely match the values from the empirical application. POET and Projected POET are the most sensitive to a reduction in the gap between the leading and bounded eigenvalues which is evident from a dramatic deterioration in the quality of these estimators. The remaining methods, including FGL, exhibit robust performance. Since the behavior of the estimators for portfolio weights is similar to that of the estimators of precision matrix, we only report the former for the ease of presentation. For $(T, p) = (300, 300)$, FClime shows the best performance followed by FGL and FLW, whereas for $(T, p) = (300, 400)$ FGL takes the lead. Despite inferior performance of POET and Projected POET in terms of estimating portfolio weights, risk exposure of the portfolios based on these estimators is competitive with the other approaches.

	$\rho = 0.4$	$\rho = 0.5$	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
	$(\lambda_3/\lambda_4 = 3.1)$	$(\lambda_3/\lambda_4 = 2.7)$	$(\lambda_3/\lambda_4 = 2.6)$	$(\lambda_3/\lambda_4 = 2.2)$	$(\lambda_3/\lambda_4 = 1.5)$	$(\lambda_3/\lambda_4 = 1.1)$
			$\lVert \widehat{\mathbf{w}}_{\text{GMV}} - \mathbf{w}_{\text{GMV}} \rVert_1$			
FGL	2.3198	2.3465	2.5177	2.4504	2.5010	2.7319
FClime	1.9554	1.9359	1.9795	1.9103	1.9813	1.9948
FLW	2.3445	2.3948	2.5328	2.4715	2.5918	3.0515
FNLW	2.2381	2.3009	2.3293	2.5497	2.9039	3.1980
POET	47.6746	82.1873	43.9722	54.1131	157.6963	235.8119
Projected POET	9.6335	7.8669	10.1546	10.6205	12.1795	15.2581
			$\left\lvert \widehat{\Phi}_{\text{GMV}} - \Phi_{\text{GMV}} \right\rvert$			
FGL	0.0033	0.0032	0.0034	0.0027	0.0021	0.0023
FClime	0.0012	0.0012	0.0012	0.0011	0.0010	0.0010
FLW	0.0049	0.0052	0.0061	0.0056	0.0049	0.0059
FNLW	0.0055	0.0060	0.0054	0.0052	0.0066	0.0057
POET	0.0070	0.0122	0.0058	0.0063	0.0103	0.0160
Projected POET	0.0021	0.0022	0.0019	0.0019	0.0018	0.0026
			$\lVert \widehat{\mathbf{w}}_{\text{MWC}} - \mathbf{w}_{\text{MWC}} \rVert_1$			
FGL	2.3766	2.4108	2.7411	2.6094	2.5669	3.4633
FClime	2.0502	2.0279	2.2901	2.1400	2.1028	3.0737
FLW	2.4694	2.5132	2.8902	2.7315	2.7210	4.0248
FNLW	2.7268	2.3060	2.8984	3.5902	2.9232	3.2076
POET	49.8603	34.2024	469.3605	108.1529	74.8016	99.4561
Projected POET	9.0261	7.4028	8.1899	9.4806	11.9642	13.3890
			$\left\lvert \widehat{\Phi}_{\text{MWC}} - \Phi_{\text{MWC}} \right\rvert$			
FGL	0.0033	0.0032	0.0034	0.0027	0.0021	0.0024
FClime	0.0012	0.0012	0.0013	0.0011	0.0010	0.0009
FLW	0.0050	0.0053	0.0062	0.0057	0.0050	0.0059
FNLW	0.0055	0.0060	0.0055	0.0053	0.0066	0.0057
POET	0.0068	0.0047	0.0363	0.0092	0.0060	0.0056
Projected POET	0.0022	0.0022	0.0020	0.0020	0.0018	0.0027
			$\lVert \widehat{\mathbf{w}}_{\text{MRC}} - \mathbf{w}_{\text{MRC}} \rVert_1$			
FGL	0.4872	0.1793	1.0044	0.6332	1.4568	2.3353
FClime	0.5160	0.2148	1.0188	0.6694	1.4855	2.3519
FLW	0.5333	0.2279	1.0345	0.6734	1.4904	2.3691
FNLW	0.8365	1.1285	1.1181	1.4419	1.7694	2.4612
POET	NaN	NaN	NaN	NaN	NaN	NaN
Projected POET	0.7414	0.6383	1.6686	1.8013	2.3297	3.2791
			$\left\lvert \widehat{\Phi}_{\text{MRC}} - \Phi_{\text{MRC}} \right\rvert$			
FGL	0.0004	0.0003	0.0025	0.0007	0.0021	0.0071
FClime	0.0005	0.0003	0.0024	0.0004	0.0016	0.0062
FLW	0.0002	0.0002	0.0021	0.0003	0.0018	0.0066
FNLW	0.0062	0.0062	0.0069	0.0119	0.0059	0.0143
POET	NaN	NaN	NaN	NaN	NaN	NaN
Projected POET	0.0003	0.0003	0.0027	0.0031	0.0069	0.0062

Table 2.B.1: Sensitivity of portfolio weights and risk exposure when the gap between the diverging and bounded eigenvalues decreases: $(T, p) = (300, 300)$.

	$\rho = 0.4$	$\rho = 0.5$	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
	$(\lambda_3/\lambda_4 = 3.1)$	$(\lambda_3/\lambda_4 = 2.7)$	$(\lambda_3/\lambda_4 = 2.6)$	$(\lambda_3/\lambda_4 = 2.2)$	$(\lambda_3/\lambda_4 = 1.5)$	$(\lambda_3/\lambda_4 = 1.1)$
$\|\widehat{\mathbf{w}}_{\text{GMV}} - \mathbf{w}_{\text{GMV}}\|_1$						
FGL	1.6900	1.8134	1.8577	1.8839	1.9843	2.0692
FClime	1.9073	1.9524	1.9997	1.9490	1.9898	2.0330
FLW	2.0239	2.0945	2.1195	2.1235	2.2473	2.4745
FNLW	2.0316	2.0790	2.1927	2.2503	2.4143	2.4710
POET	18.7934	28.0493	155.8479	32.4197	41.8098	71.5811
Projected POET	7.8696	8.4915	8.8641	10.7522	11.2092	19.0424
$\left\lvert\widehat{\Phi}_{\text{GMV}} - \Phi_{\text{GMV}}\right\rvert$						
FGL	8.62E-04	9.22E-04	7.23E-04	7.31E-04	6.83E-04	5.73E-04
FClime	8.40E-04	8.27E-04	8.02E-04	7.87E-04	7.36E-04	6.71E-04
FLW	1.59E-03	1.73E-03	1.57E-03	1.68E-03	1.69E-03	1.54E-03
FNLW	2.24E-03	2.10E-03	1.83E-03	1.88E-03	2.07E-03	1.29E-03
POET	1.11E-03	1.46E-03	3.59E-03	1.27E-03	1.88E-03	2.51E-03
Projected POET	8.97E-04	8.80E-04	6.83E-04	6.79E-04	7.98E-04	6.55E-04
$\|\widehat{\mathbf{w}}_{\text{MWC}} - \mathbf{w}_{\text{MWC}}\|_1$						
FGL	1.9034	2.2843	1.9118	3.2569	2.7055	2.8812
FClime	2.1193	2.4024	2.0540	3.3487	2.7277	2.8593
FLW	2.2573	2.5809	2.1790	3.5728	3.0072	3.3164
FNLW	2.3207	3.3335	3.5518	3.4282	2.6446	4.8827
POET	15.8824	100.1419	56.9827	33.6483	38.8961	103.0434
Projected POET	6.5386	7.2169	7.8583	9.7342	12.1420	17.7368
$\left\lvert\widehat{\Phi}_{\text{MWC}} - \Phi_{\text{MWC}}\right\rvert$						
FGL	8.72E-04	9.41E-04	7.26E-04	7.99E-04	7.12E-04	6.08E-04
FClime	8.52E-04	8.49E-04	8.06E-04	8.32E-04	7.50E-04	6.86E-04
FLW	1.59E-03	1.74E-03	1.57E-03	1.71E-03	1.70E-03	1.56E-03
FNLW	2.25E-03	2.22E-03	1.89E-03	1.91E-03	2.08E-03	1.56E-03
POET	1.14E-03	4.91E-03	1.78E-03	1.45E-03	1.57E-03	2.93E-03
Projected POET	9.19E-04	9.20E-04	7.11E-04	7.04E-04	8.26E-04	6.78E-04
$\|\widehat{\mathbf{w}}_{\text{MRC}} - \mathbf{w}_{\text{MRC}}\|_1$						
FGL	0.6683	0.7390	1.3103	1.5195	1.7124	3.0935
FClime	0.6903	0.7635	1.3238	1.5403	1.7415	3.1180
FLW	0.7132	0.7828	1.3430	1.5549	1.7517	3.1364
FNLW	0.4909	1.2121	1.4974	1.1996	1.8020	3.2989
POET	NaN	NaN	NaN	NaN	NaN	NaN
Projected POET	1.6851	1.4434	1.9628	2.6182	2.7716	4.1753
$\left\lvert\widehat{\Phi}_{\text{MRC}} - \Phi_{\text{MRC}}\right\rvert$						
FGL	1.02E-03	9.73E-04	4.63E-03	4.49E-03	3.23E-03	8.73E-03
FClime	1.14E-03	1.01E-03	4.55E-03	4.22E-03	2.70E-03	7.72E-03
FLW	6.62E-04	5.54E-04	4.19E-03	4.01E-03	2.71E-03	8.11E-03
FNLW	2.73E-04	6.93E-03	5.11E-03	1.93E-03	6.42E-03	2.98E-02
POET	NaN	NaN	NaN	NaN	NaN	NaN
Projected POET	3.59E-03	1.20E-03	1.49E-03	2.58E-03	7.86E-03	1.39E-02

Table 2.B.2: Sensitivity of portfolio weights and risk exposure when the gap between the diverging and bounded eigenvalues decreases: $(T, p) = (300, 400)$.

2.C Additional Empirical Results

Similarly to daily data, we use monthly returns of the components of the S&P500 from CRSP and Compustat. The full sample has 480 observations on 355 stocks from January 1, 1980 - December 1, 2019. We use January 1, 1980 - December 1, 1994 (180 obs) as a training (estimation) period and January 1, 1995 - December 1, 2019 (300 obs) as the out-of-sample test period. At the end of each month, prior to portfolio construction, we remove stocks with less than 15 years of historical stock return data. We set the return target $\mu = 0.7974\%$ which is equivalent to 10% yearly return when compounded. The target level of risk for MWC and MRC portfolios is set at $\sigma = 0.05$ which is the standard deviation of the monthly excess returns of the S&P500 index in the first training set. Transaction costs are taken to be the same as for the daily returns in Section 6.

Table 2.C.1 reports the results for monthly data. Some comments are in order: **(1)** interestingly, MRC produces portfolio return and Sharpe Ratio that are mostly higher than those for the weight-constrained allocations MWC and GMV. This means that relaxing the constraint that portfolio weights sum up to one leads to a large increase in the out-of-sample Sharpe Ratio and portfolio return which has not been previously well-studied in the empirical finance literature. **(2)** Similarly to the results from Table 2.1, FGL outperforms the competitors including EW and Index in terms of the out-of-sample Sharpe Ratio and turnover. **(3)** Similarly to the results in Table 2.1, the observable Fama-French factors produce the FGL portfolios with higher return and higher out-of-sample Sharpe Ratio compared to the FGL portfolios based on statistical factors. Again, this increase in return is not followed by higher risk.

	Markowitz Risk-Constrained				Markowitz Weight-Constrained				Global Minimum-Variance			
	Return	**Risk**	**SR**	**Turnover**	**Return**	**Risk**	**SR**	**Turnover**	**Return**	**Risk**	**SR**	**Turnover**
Without TC												
EW	0.0081	0.0519	0.1553	-	0.0081	0.0519	0.1553	-	0.0081	0.0519	0.1553	-
Index	0.0063	0.0453	0.1389	-	0.0063	0.0453	0.1389	-	0.0063	0.0453	0.1389	-
FGL	0.0256	0.0828	0.3099	-	0.0059	0.0329	0.1804	-	0.0065	0.0321	0.2023	-
FClime	0.0372	0.2337	0.1593	-	0.0067	0.0471	0.1434	-	0.0076	0.0466	0.1643	-
FLW	0.0296	0.1049	0.2817	-	0.0059	0.0353	0.1662	-	0.0063	0.0353	0.1774	-
FNLW	0.0264	0.0925	0.2853	-	0.0060	0.0333	0.1793	-	0.0064	0.0332	0.1930	-
POET	NaN	NaN	NaN	-	-0.1041	2.0105	-0.0518	-	0.5984	11.0064	0.0544	-
Projected POET	0.0583	0.3300	0.1766	-	0.0058	0.0546	0.1056	-	0.0069	0.0612	0.1128	-
FGL (FF1)	0.0275	0.0800	0.3433	-	0.0061	0.0316	0.1941	-	0.0073	0.0302	0.2427	-
FGL (FF3)	0.0274	0.0797	0.3437	-	0.0061	0.0314	0.1955	-	0.0073	0.0300	0.2440	-
FGL (FF5)	0.0273	0.0793	0.3443	-	0.0061	0.0314	0.1943	-	0.0073	0.0300	0.2426	-
With TC												
EW	0.0080	0.0520	0.1538	0.0630	0.0080	0.0520	0.1538	0.0630	0.0080	0.0520	0.1538	0.0630
FGL	0.0222	0.0828	0.2682	3.1202	0.0050	0.0329	0.1525	0.8786	0.0056	0.0321	0.1740	0.8570
FClime	0.0334	0.2334	0.1429	4.9174	0.0062	0.0471	0.1307	0.5945	0.0071	0.0466	0.1522	0.5528
FLW	0.0237	0.1052	0.2257	5.5889	0.0043	0.0353	0.1231	1.5166	0.0048	0.0354	0.1343	1.5123
FNLW	0.0224	0.0927	0.2415	3.7499	0.0049	0.0334	0.1463	1.0812	0.0053	0.0333	0.1596	1.0793
POET	NaN	NaN	NaN	NaN	-0.1876	1.7274	-0.1086	152.3298	1.0287	14.2676	0.0721	354.6043
Projected POET	0.0166	0.2859	0.0579	69.7600	-0.0002	0.0540	-0.0044	5.9131	-0.0002	0.0613	-0.0027	7.0030
FGL (FF1)	0.0243	0.0800	0.3036	2.8514	0.0054	0.0317	0.1692	0.7513	0.0066	0.0302	0.2176	0.7095
FGL (FF3)	0.0242	0.0797	0.3037	2.8708	0.0054	0.0314	0.1703	0.7545	0.0066	0.0300	0.2186	0.7127
FGL (FF5)	0.0241	0.0793	0.3037	2.8857	0.0053	0.0315	0.1686	0.7630	0.0065	0.0300	0.2167	0.7224

Table 2.C.1: Monthly portfolio returns, risk, Sharpe Ratio (SR) and turnover.

Chapter 3

Sparse Portfolios

Abstract

The existing approaches to sparse wealth allocations (1) are suboptimal due to the bias induced by ℓ_1-penalty; (2) require the number of assets to be less than the sample size; (3) do not model factor structure of stock returns in high dimensions. We address these shortcomings and develop a novel strategy which produces unbiased and consistent sparse allocations. We demonstrate that: (1) failing to correct for the bias leads to low out-of-sample portfolio return; (2) only sparse portfolios achieved positive cumulative return during several economic downturns, including the dot-com bubble of 2000, the financial crisis of 2007-09, and COVID-19 outbreak.

3.1 Introduction

The search for the optimal portfolio weights reduces to the questions (i) which stocks to buy and (ii) how much to invest in these stocks. Depending on the strategy used to address the first question, the existing allocation approaches can be further broken down into the ones that invest in all available stocks, and the ones that select a subset out of the stock universe. The latter is referred to as a *sparse portfolio*, since some assets will be excluded and get a zero weight leading to sparse wealth allocations. Any portfolio optimization problem requires the inverse covariance matrix, or *precision* matrix, of excess stock returns as an input. In the era of big data, a search for the optimal portfolio becomes a high-dimensional problem: the number of assets, p, is comparable to or greater than the sample size, T. Constructing non-sparse portfolios in high dimensions has been the main focus of the existing research on asset management for a long time. In particular, many papers focus on developing an improved covariance or precision estimator to achieve desirable statistical properties of portfolio weights. In contrast, the literature on constructing sparse portfolio is scarce: it is limited to a low-dimensional framework and lacks theoretical analysis of the resulting sparse allocations. In this paper we fill this gap and propose a novel approach to construct sparse portfolios in high dimensions. We obtain the oracle bounds of sparse weight estimators and provide guidance regarding their distribution. From the empirical perspective, we examine the merit of sparse portfolios during the periods of economic growth, moderate market decline and severe economic downturns. We find that in contrast to non-sparse counterparts, our strategy is robust to recessions and can be used as a hedging vehicle during such times.

As pointed out above, estimating high-dimensional covariance or precision matrix to improve portfolio performance of non-sparse strategies has received a lot of attention in the existing literature. [91, 95] developed linear and non-linear shrinkage estimators of covariance matrix, [53, 56] introduced a covariance matrix estimator when stock returns are driven by common factors under the assumption of a spiked covariance model. Once the covariance estimator is obtained, it is then inverted to get a precision matrix, the main input to any portfolio optimization problem. A parallel stream of literature has focused on estimating precision matrix directly, that is, avoiding the inversion step that leads to additional estimation errors, especially in high dimensions. [65] developed an iterative algorithm that estimates the entries of precision matrix column-wise using penalized Gaussian log-likelihood (Graphical Lasso); [108] used the relationship between regression coefficients and the entries of precision matrix to estimate the elements of the latter column by column (nodewise regression). [23] use constrained ℓ_1-minimization for inverse matrix estimation (CLIME). [24] examined the performance of high-dimensional portfolios constructed using covariance and precision estimators and found that precision-based models outperform covariance-based counterparts in terms of the out-of-sample (OOS) Sharpe Ratio and portfolio return.

From a practical perspective, apart from enjoying favorable statistical properties a successful wealth allocation strategy should be easy to maintain and monitor and it should be robust to economic downturns such that investors could use it as a hedging vehicle. Having this motivation in mind, we chose several popular covariance and precision-based estimators to construct non-sparse portfolios and explore their performance during the recent COVID-19 outbreak. Using daily returns of 495 constituents of the S&P500 from May 25, 2018 – September 24, 2020 (588 obs.), Table 3.1.1 reports the performance of the selected strategies: we included equal-weighted (EW) and Index portfolios, as well as precision-based nodewise regression estimator by [108] (motivated by the recent application of this statistical technique to portfolio studied in [24]), linear shrinkage covariance estimator by [91] and CLIME by [23]. We use May 25, 2018 – October 23, 2018 (105 obs.) as a training period and October 24, 2018 – September 24, 2020 (483 obs.) as the out-of-sample test period. We roll the estimation window over the test sample to rebalance the portfolios monthly. The left panel of Table 3.1.1 shows return, risk and Sharpe Ratio of portfolios over the training period, and the right panel reports cumulative excess return (CER) and risk over two sub-periods of interest: before the pandemic (January 2, 2019 – December 31, 2019) and during the first wave of COVID-19 outbreak in the US (January 2, 2020 – June 30, 2020). As evidenced by Table 3.1.1, none of the portfolios was robust to the downturn brought by pandemic and yielded negative CER. We noticed that similar pattern pertained in several other historic episodes of mild and severe downturns, such as the Global Financial Crisis (GFC) of 2007-09.[1]

Studies that examine the relationship between portfolio performance and the number of stock holdings are scarce. [131] used active US equity funds' quarterly data from January 2000 to December 2017 from Morningstar, Inc. to study the impact of concentration (measured by the number of holdings) on fund excess returns: they found that the effect was significant and fluctuated considerably over time. Notably, the relationship became negative in the period preceding and including the GFC. This indicates that holding sparse

[1] Please see the Empirical Application section for more details.

portfolios might be the key to hedging during downturns. To support this hypothesis, we further compare the performance of sparse vs non-sparse strategies in terms of utility gain to investors. Suppose we observe $i = 1, \dots, p$ excess returns over $t = 1, \dots, T$ period of time: $\mathbf{r}_t = (r_{1t}, \dots, r_{pt})' \sim \mathcal{D}(\mathbf{m}, \mathbf{\Sigma})$. Consider the following mean-variance utility problem: $\min_{\mathbf{w}} -U \equiv \frac{\gamma}{2}\mathbf{w}\mathbf{\Sigma}\mathbf{w} - \mathbf{w}'\mathbf{m}$, s.t. $\mathbf{w}'\boldsymbol{\iota} = 1$, $|\text{supp}(\mathbf{w})| \leq \bar{p}$, $\bar{p} \leq p$, where $\mathbf{w}$ is a $p \times 1$ vector of portfolio weights, $\text{supp}(\mathbf{w}) = \{i : w_i > 0\}$ is the cardinality constraint that controls sparsity, and γ determines the risk of an investor under the assumption of a normal distribution. When $\bar{p} = p$ the portfolio is non-sparse and the respective utility is denoted as $U^{\text{Non-Sparse}}$, while when $\bar{p} < p$ the utility of such sparse portfolio is denoted as U^{Sparse}. Figure 3.1.1 reports the ratio of utilities using monthly data from 2003:04 to 2009:12 on the constituents of the S&P100 as a function of $\bar{p}$: we set $\gamma = 3$ and vary $\bar{p} = \{5, 10, 15, 20, 30, \dots, 90\}$[2]. Our test sample includes two periods of particular interest: before the GFC (2004:01-2006:12) and during the GFC (2007:01-2009:12) As evidenced from Figure 3.1.1: (1) for both time periods there exists a lower-dimensional subset of stocks which brings greater utility compared to non-sparse portfolios; (2) the number of stocks minimizing the ratio of utilities is smaller during the GFC compared to the period preceding it. Both findings are consistent with the empirical result of [131] that including more stocks does not guarantee better performance and suggesting that holding a "basket half full" instead can help achieve superior performance even in stressed market scenarios.

[2] Since the optimization problem with a cardinality constraint is not convex, we find a solution using Lagrangian relaxation procedure of [124]

	Total OOS Performance 10/24/19–09/24/20			Before the Pandemic 01/02/19–12/31/19		During the Pandemic 01/02/20–06/30/20	
	Return (×100)	**Risk** (×100)	**Sharpe Ratio**	**CER** (×100)	**Risk** (×100)	**CER** (×100)	**Risk** (×100)
EW	0.0108	1.8781	0.0058	28.5420	0.8010	-19.7207	3.3169
Index	0.0351	1.7064	0.0206	27.8629	0.7868	-9.0802	2.9272
Nodewise Regr'n	0.0322	1.6384	0.0196	29.6292	0.6856	-11.7431	2.8939
CLIME	0.0793	3.1279	0.0373	31.5294	1.0215	-25.3004	3.8972
LW	0.0317	1.7190	0.0184	29.5513	0.7924	-14.9328	3.0115
Our Post-Lasso-based	0.1247	1.7254	0.0723	45.2686	1.0386	12.4196	2.8554
Our De-biased Estimator	0.0275	0.5231	0.0526	23.7629	0.4972	6.5813	0.5572

Table 3.1.1: Performance of non-sparse and sparse portfolios: return (×100), risk (×100) and Sharpe Ratio over the training period (left), CER (×100) and risk (×100) over two sub-periods (right). Weights are estimated using the standard Global Minimum Variance formula.

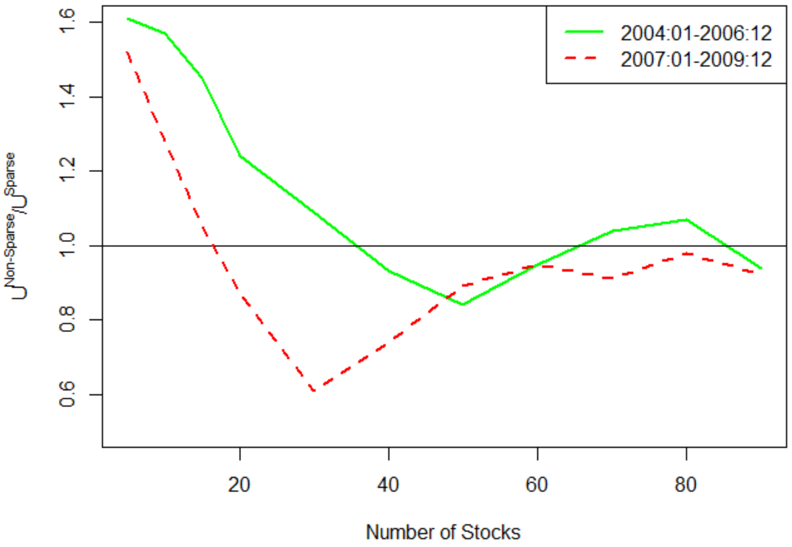

Figure 3.1.1: **The ratio of non-sparse and sparse portfolio utilities averaged over the test window.**

In order to create a sparse portfolio, that is, a portfolio with many zero entries in the weight vector, we can use an ℓ_1-penalty (Lasso) on the portfolio weights which shrinks some of them to zero (see [58], [2], [100], [17] among others). [19] proved the mathematical equivalence of adding an ℓ_1-penalty and controlling transaction costs associated with the bid-ask spread impact of single and sequential trades executed in a very short time. This indicates another advantage of sparse portfolios: market liquidity dries up during economic downturns which increases bid-ask spreads, a measure of liquidity costs. Henceforth, regularizing portfolio positions accounts for the increased liquidity risk associated with acquiring and liquidating positions. The existing literature on sparse wealth allocations is scarce and has several drawbacks: (1) it is limited to low-dimensional setup when $p < T$, whereas sparsity becomes especially important in high-dimensional scenarios; (2) it lacks theoretical analysis of sparse wealth allocations and their impact on portfolio exposure; (3) the use of an ℓ_1-penalty produces biased estimates (see [14, 77–79, 135, 141] among others), however, this issue has been overlooked in the context of portfolio allocation. This paper addresses the aforementioned drawbacks and develops an approach to construct sparse portfolios in high dimensions. Our contribution is twofold: from the theoretical perspective, we establish the oracle bounds of sparse weight estimators and provide guidance regarding their distribution. From the empirical perspective, we examine the merit of sparse portfolios during different market scenarios. We find that in contrast to non-sparse counterparts, our strategy is robust to recessions and can be used as a hedging vehicle during such times. To illustrate, the last two rows of Table 3.1.1 show the performance of two sparse strategies proposed in this paper: both approaches outperform non-sparse counterparts in terms of total OOS Sharpe Ratio, and they produce positive CER during the pandemic, as well as in the period preceding it. Figure 3.1 shows the stocks selected by post-Lasso in August, 2019 and in May, 2020: the colors serve as a visual guide to identify groups of closely-related stocks (stocks of the same color do not necessarily correspond to the same sector). Our framework makes use of the tool from the network theory called nodewise regression which not only satisfies desirable statistical properties, but also allows us to study whether certain industries could serve as safe havens during recessions. We find that such non-cyclical industries as consumer staples, healthcare, retail and food were driving the returns of the sparse portfolios during both GFC and COVID-19 outbreak, whereas insurance sector was the least attractive investment in both periods.

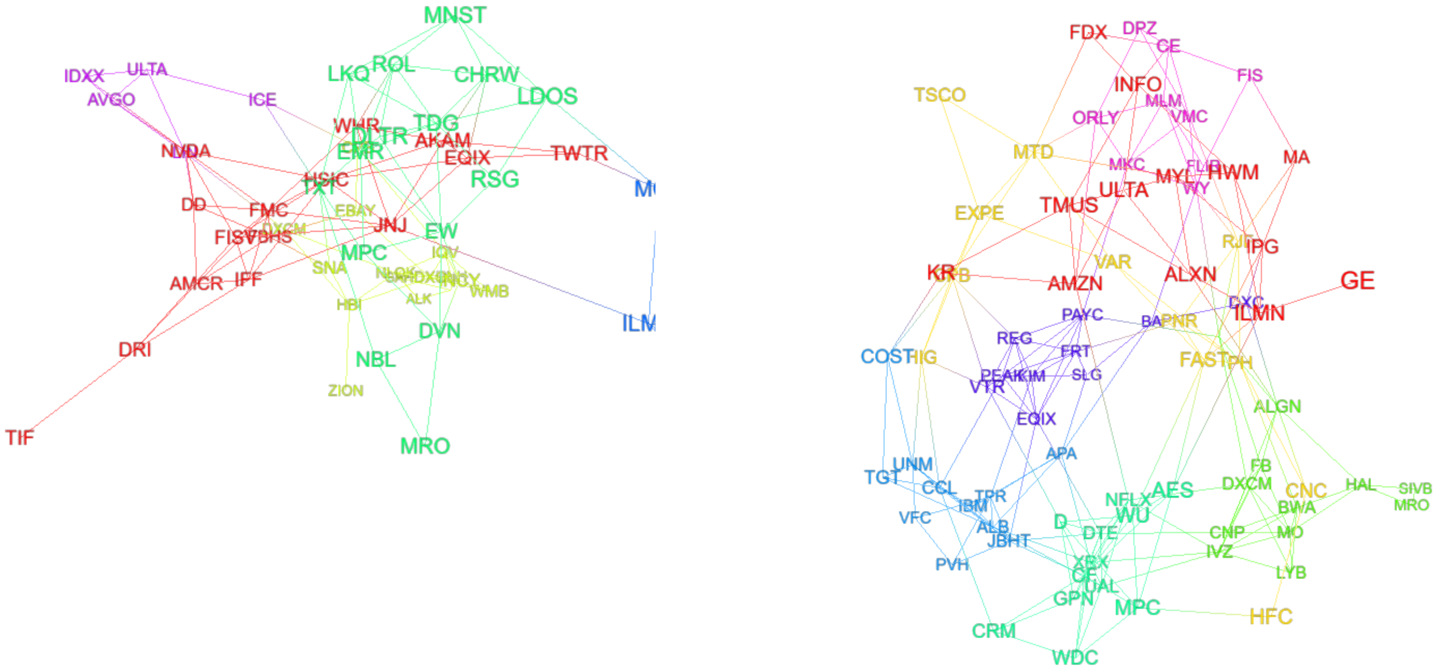

Figure 3.1.2: **Stocks selected by Post-Lasso strategy from Table 3.1.1:** August, 2019 (left) and May, 2020 (right)

This paper is organized as follows: Section 2 introduces sparse de-biased portfolio and sparse portfolio using post-Lasso. Section 3 develops a new high-dimensional precision estimator called Factor Nodewise regression. Section 4 develops a framework for factor investing. Section 5 contains theoretical results and Section 6 validates these results using simulations. Section 7 provides empirical application. Section 8 concludes.

Notation

For the convenience of the reader, we summarize the notation to be used throughout the paper. Let $\mathcal{S}_p$ denote the set of all $p \times p$ symmetric matrices. For any matrix $\mathbf{C}$, its (i,j)-th element is denoted as c_{ij}. Given a vector $\mathbf{u} \in \mathbb{R}^d$ and parameter $a \in [1, \infty)$, let $\|\mathbf{u}\|_a$ denote ℓ_a-norm. Given a matrix $\mathbf{U} \in \mathcal{S}_p$, let $\Lambda_{\max}(\mathbf{U}) \equiv \Lambda_1(\mathbf{U}) \geq \Lambda_2(\mathbf{U}) \geq \ldots \Lambda_{\min}(\mathbf{U}) \equiv \Lambda_p(\mathbf{U})$ be the eigenvalues of $\mathbf{U}$, and $\text{eig}_K(\mathbf{U}) \in \mathbb{R}^{K \times p}$ denote the first $K \leq p$ normalized eigenvectors corresponding to $\Lambda_1(\mathbf{U}), \ldots \Lambda_K(\mathbf{U})$. Given parameters $a, b \in [1, \infty)$, let $|||\mathbf{U}|||_{a,b} = \max_{\|\mathbf{y}\|_a=1} \|\mathbf{U}\mathbf{y}\|_b$ denote the induced matrix-operator norm. The special cases are $|||\mathbf{U}|||_1 \equiv \max_{1 \leq j \leq p} \sum_{i=1}^{p} |u_{i,j}|$ for the ℓ_1/ℓ_1-operator norm; the operator norm (ℓ_2-matrix norm) $|||\mathbf{U}|||_2^2 \equiv \Lambda_{\max}(\mathbf{U}\mathbf{U}')$ is equal to the maximal singular value of $\mathbf{U}$; $|||\mathbf{U}|||_\infty \equiv \max_{1 \leq j \leq p} \sum_{i=1}^{p} |u_{j,i}|$ for the ℓ_∞/ℓ_∞-operator norm. Finally, $\|\mathbf{U}\|_{\max} = \max_{i,j} |u_{i,j}|$ denotes the element-wise maximum, and $|||\mathbf{U}|||_F^2 = \sum_{i,j} u_{i,j}^2$ denotes the Frobenius matrix norm. We also use the following notations: $a \vee b = \max\{a, b\}$, and $a \wedge b = \min\{a, b\}$. For an event A, we say that A wp $\rightarrow 1$ when A occurs with probability approaching 1 as T increases.

3.2 Sparse Portfolios

There exist several widely used portfolio weight formulations depending on the type of optimization problem solved by an investor. Suppose we observe p assets (indexed by i) over T period of time (indexed by t). Let $\mathbf{r}_t = (r_{1t}, r_{2t}, \ldots, r_{pt})' \sim \mathcal{D}(\mathbf{m}, \mathbf{\Sigma})$ be a $p \times 1$ vector of *excess* returns drawn from a distribution $\mathcal{D}$, where $\mathbf{m}$ and $\mathbf{\Sigma}$ are unconditional mean and covariance of excess returns, and $\mathcal{D}$ belongs to either sub-Gaussian or elliptical families. When $\mathcal{D} = \mathcal{N}$, the precision matrix $\mathbf{\Sigma}^{-1} \equiv \mathbf{\Theta}$ contains information about conditional dependence between the variables. For instance, if θ_{ij}, which is the ij-th element of the precision matrix, is zero, then the variables i and j are conditionally independent, given the other variables. The goal of the Markowitz theory is to choose assets weights in a portfolio *optimally.* We will study two criteria of optimality: the first is a well-known Markowitz weight-constrained optimization problem, and the second formulation relaxes constraints on portfolio weights.

The first optimization problem, which will be referred to as *Markowitz weight-constrained problem (MWC)*, searches for assets weights such that the portfolio achieves a desired expected rate of return with minimum risk, under the restriction that all weights sum up to one. The aforementioned goal can be formulated as the following quadratic optimization problem:

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}'\mathbf{\Sigma}\mathbf{w}, \text{ s.t. } \mathbf{w}'\boldsymbol{\iota} = 1 \text{ and } \mathbf{m}'\mathbf{w} \geq \mu, \tag{3.1}$$

where $\mathbf{w}$ is a $p \times 1$ vector of assets weights in the portfolio, $\boldsymbol{\iota}$ is a $p \times 1$ vector of ones, and

μ is a desired expected rate of portfolio return. The constraint in (3.1) requires portfolio weights to sum up to one - this assumption can be easily relaxed and we will demonstrate the implications of this constraint on portfolio weights.

If $\mathbf{m}'\mathbf{w} > \mu$, then the solution to (3.1) yields the *global minimum-variance (GMV)* portfolio weights $\mathbf{w}_{GMV}$:

$$\mathbf{w}_{GMV} = (\boldsymbol{\iota}'\boldsymbol{\Theta}\boldsymbol{\iota})^{-1}\boldsymbol{\Theta}\boldsymbol{\iota}. \tag{3.2}$$

If $\mathbf{m}'\mathbf{w} = \mu$, the solution to (3.1) is

$$\mathbf{w}_{MWC} = (1 - a_1)\mathbf{w}_{GMV} + a_1\mathbf{w}_M, \tag{3.3}$$

$$\mathbf{w}_M = (\boldsymbol{\iota}'\boldsymbol{\Theta}\mathbf{m})^{-1}\boldsymbol{\Theta}\mathbf{m}, \tag{3.4}$$

$$a_1 = \frac{\mu(\mathbf{m}'\boldsymbol{\Theta}\boldsymbol{\iota})(\boldsymbol{\iota}'\boldsymbol{\Theta}\boldsymbol{\iota}) - (\mathbf{m}'\boldsymbol{\Theta}\boldsymbol{\iota})^2}{(\mathbf{m}'\boldsymbol{\Theta}\mathbf{m})(\boldsymbol{\iota}'\boldsymbol{\Theta}\boldsymbol{\iota}) - (\mathbf{m}'\boldsymbol{\Theta}\boldsymbol{\iota})^2}, \tag{3.5}$$

where $\mathbf{w}_{MWC}$ denotes the portfolio allocation with the constraint that the weights need to sum up to one and $\mathbf{w}_M$ captures all mean-related market information.

The second optimization problem, which will be referred to as *Markowitz risk-constrained (MRC)* problem, has the same objective as in (3.1), but portfolio weights are not required to sum up to one:

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} \quad \text{s.t. } \mathbf{m}'\mathbf{w} \geq \mu. \tag{3.6}$$

It can be easily shown that the solution to (2.6) is:

$$\mathbf{w}_1^* = \frac{\mu\boldsymbol{\Theta}\mathbf{m}}{\mathbf{m}'\boldsymbol{\Theta}\mathbf{m}}. \tag{3.7}$$

Alternatively, instead of searching for a portfolio with a specified desired expected rate of return and minimum risk, one can maximize expected portfolio return given a maximum risk-tolerance level:

$$\max_{\mathbf{w}} \mathbf{w}'\mathbf{m} \quad \text{s.t. } \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} \leq \sigma^2. \tag{3.8}$$

In this case, the solution to (2.8) yields:

$$\mathbf{w}_2^* = \frac{\sigma^2}{\mathbf{w}'\mathbf{m}}\boldsymbol{\Theta}\mathbf{m} = \frac{\sigma^2}{\mu}\boldsymbol{\Theta}\mathbf{m}. \tag{3.9}$$

To get the second equality in (2.9) we used the definition of μ from (3.1) and (3.6). It follows that if $\mu = \sigma\sqrt{\theta}$, where $\theta \equiv \mathbf{m}'\boldsymbol{\Theta}\mathbf{m}$ is the squared Sharpe Ratio, then the solution to (2.6) and (2.8) admits the following expression:

$$\mathbf{w}_{MRC} = \frac{\sigma}{\sqrt{\mathbf{m}'\boldsymbol{\Theta}\mathbf{m}}}\Theta\mathbf{m} = \frac{\sigma}{\sqrt{\theta}}\boldsymbol{\alpha}, \tag{3.10}$$

where $\boldsymbol{\alpha} \equiv \boldsymbol{\Theta}\mathbf{m}$. Equation (2.10) tells us that once an investor specifies the desired return, μ, and maximum risk-tolerance level, σ, this pins down the Sharpe Ratio of the portfolio which makes the optimization problems of minimizing risk and maximizing expected return of the portfolio in (2.6) and (2.8) identical.

This brings us to three alternative portfolio allocations commonly used in the existing literature: Global Minimum-Variance Portfolio in (3.2), weight-constrained Markowitz Mean-Variance in (3.3) and maximum-risk-constrained Markowitz Mean-Variance in (2.10).

Below we summarize the aforementioned portfolio weight expressions:

$$
\begin{aligned}
&\text{GMV:} && \mathbf{w}_{GMV} = (\boldsymbol{\iota}'\boldsymbol{\Theta}\boldsymbol{\iota})^{-1}\boldsymbol{\Theta}\boldsymbol{\iota}, && (3.11)\\
&\text{MWC} && \mathbf{w}_{MWC} = (1-a_1)\mathbf{w}_{GMV} + a_1\mathbf{w}_M, && (3.12)\\
&\text{where} && \mathbf{w}_M = (\boldsymbol{\iota}'\boldsymbol{\Theta}\mathbf{m})^{-1}\boldsymbol{\Theta}\mathbf{m}, &&\\
& && a_1 = \frac{\mu(\mathbf{m}'\boldsymbol{\Theta}\boldsymbol{\iota})(\boldsymbol{\iota}'\boldsymbol{\Theta}\boldsymbol{\iota}) - (\mathbf{m}'\boldsymbol{\Theta}\boldsymbol{\iota})^2}{(\mathbf{m}'\boldsymbol{\Theta}\mathbf{m})(\boldsymbol{\iota}'\boldsymbol{\Theta}\boldsymbol{\iota}) - (\mathbf{m}'\boldsymbol{\Theta}\boldsymbol{\iota})^2}, &&\\
&\text{MRC:} && \mathbf{w}_{MRC} = \frac{\sigma}{\sqrt{\theta}}\boldsymbol{\alpha}, && (3.13)\\
&\text{where} && \boldsymbol{\alpha} = \boldsymbol{\Theta}\mathbf{m}, \quad \theta = \mathbf{m}'\boldsymbol{\Theta}\mathbf{m} &&
\end{aligned}
$$

So far we have considered allocation strategies that put non-zero weights to all assets in the financial portfolio. As an implication, an investor needs to buy a certain amount of each security even if there are a lot of small weights. However, oftentimes investors are interested in managing a few assets which significantly reduces monitoring and transaction costs and was shown to outperform equal weighted and index portfolios in terms of the Sharpe Ratio and cumulative return (see [58], [2], [100], [17] among others). This strategy is based on holding a *sparse portfolio*, that is, a portfolio with many zero entries in the weight vector.

3.2.1 Sparse De-Biased Portfolio

Let us first introduce some notations. The sample mean and sample covariance matrix have standard formulas: $\widehat{\mathbf{m}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_t$ and $\widehat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{r}_t - \widehat{\mathbf{m}})(\mathbf{r}_t - \widehat{\mathbf{m}})'$. Our empirical application shows that risk-constrained Markowitz allocation in (3.13) outperforms GMV and MWC portfolios in (3.11)-(3.12). Therefore, we first study sparse MRC portfolios. Our goal is to construct a sparse vector of portfolio weights given by (3.13). To achieve this we use the following equivalent and unconstrained regression representation of the mean-variance optimization in (2.6) and (2.8):

$$\mathbf{w}_{MRC} = \underset{\mathbf{w}}{\text{argmin}}\, \mathbb{E}\big[y - \mathbf{w}'\mathbf{r}_t\big]\,, \quad \text{where} \quad y \equiv \frac{1+\theta}{\theta}\mu \equiv \sigma\frac{1+\theta}{\sqrt{\theta}}. \tag{3.14}$$

The sample counterpart of (3.14) is written as:

$$\mathbf{w}_{MRC} = \underset{\mathbf{w}}{\text{argmin}}\, \frac{1}{T}\sum_{t=1}^{T}(y - \mathbf{w}'\mathbf{r}_t)^2. \tag{3.15}$$

[2] prove that the weight allocation from (3.14) is equivalent to (3.13). The sparsity is introduced through Lasso which yields the following constrained optimization problem:

$$\mathbf{w}_{\text{MRC, SPARSE}} = \underset{\mathbf{w}}{\text{argmin}}\, \frac{1}{T}\sum_{t=1}^{T}(y - \mathbf{w}'\mathbf{r}_t)^2 + 2\lambda\|\mathbf{w}\|_1. \tag{3.16}$$

Now we propose two extensions to the setup (3.16). First, the estimator $\mathbf{w}_{\text{MRC, SPARSE}}$ is infeasible since θ used for constructing y is unknown. [2] construct an estimator of θ under normally distributed excess returns, assuming $p/T \to \rho \in (0,1)$ and the sample size T is required to be larger than the number of assets p. Their paper uses an unbiased estimator proposed in [84]: $\hat{\theta} = ((T-p-2)\widehat{\mathbf{m}}'\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{m}} - p)/T$, where $\widehat{\mathbf{m}}$ and $\widehat{\boldsymbol{\Sigma}}^{-1}$ are sample mean and inverse of the sample covariance matrix respectively. One of the limitations of the model studied by [2] is that it cannot handle high dimensions. In both simulations and empirical application the maximum number of stocks used by the authors is limited to 100. Another limitation of [2] approach is that they do not correct the bias introduced by imposing ℓ_1-constraint in (3.16). However, it is well-known that the estimator in (3.16) is biased and the existing literature proposes several de-biasing techniques (see [14, 77–79, 135, 141] among others).

To address the first aforementioned limitation, we propose to use an estimator of a high-dimensional precision matrix discussed in the next section. The suggested estimator is appropriate for high-dimensional settings, it can handle cases when the sample size is less than the number of assets, and it is always non-negative by construction[3]. Consequently, the estimator of y is

$$\widehat{y} \equiv \frac{1+\hat{\theta}}{\hat{\theta}}\mu \equiv \sigma\frac{1+\hat{\theta}}{\sqrt{\hat{\theta}}}. \tag{3.17}$$

To approach the second limitation, motivated by [135], we propose the de-biasing technique that uses the nodewise regression estimator of the precision matrix. First, let $\mathbf{R}$ be a $T \times p$ matrix of excess returns stacked over time and $\widehat{\mathbf{y}}$ be a $T \times 1$ constant vector. Consider a high-dimensional linear model

$$\widehat{\mathbf{y}} = \mathbf{R}\mathbf{w} + \mathbf{e}, \quad \text{where} \quad \mathbf{e} \sim \mathcal{D}(\mathbf{0}, \sigma_e^2\mathbf{I}). \tag{3.18}$$

We study high-dimensional framework $p \geq T$ and in the asymptotic results we require $\log p/T = o(1)$. Let us rewrite (3.16):

$$\mathbf{w}_{\text{MRC, SPARSE}} = \arg\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{T}\|\widehat{\mathbf{y}} - \mathbf{R}\mathbf{w}\|_2^2 + 2\lambda\|\mathbf{w}\|_1. \tag{3.19}$$

[3] Our empirical results suggest that the unbiased estimator $\hat{\theta} = ((T-p-2)\widehat{\mathbf{m}}'\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{m}} - p)/T$ is oftentimes negative even after using the adjusted estimator defined in [84] (p. 2906).

The estimator in (3.16) satisfies the following KKT conditions:

$$-\mathbf{R}'(\widehat{\mathbf{y}}-\mathbf{R}\widehat{\mathbf{w}})/T+\lambda\widehat{\mathbf{g}}=0, \tag{3.20}$$

$$\|\widehat{\mathbf{g}}\|_\infty \leq 1 \quad \text{and} \quad \hat{g}_i=\text{sign}(\hat{w}_i) \quad \text{if} \quad \hat{w}_i \neq 0. \tag{3.21}$$

where $\widehat{\mathbf{g}}$ is a $p \times 1$ vector arising from the subgradient of $\|\mathbf{w}\|_1$. Let $\widehat{\mathbf{\Sigma}}=\mathbf{R}'\mathbf{R}/T$, then we can rewrite the KKT conditions:

$$\widehat{\mathbf{\Sigma}}(\widehat{\mathbf{w}}-\mathbf{w})+\lambda\widehat{\mathbf{g}}=\mathbf{R}'\mathbf{e}/T. \tag{3.22}$$

Multiply both sides of (3.22) by $\widehat{\mathbf{\Theta}}$ obtained from Algorithm 5, add and subtract $(\widehat{\mathbf{w}}-\mathbf{w})$, and rearrange the terms:

$$\widehat{\mathbf{w}}-\mathbf{w}+\widehat{\mathbf{\Theta}}\lambda\widehat{\mathbf{g}}=\widehat{\mathbf{\Theta}}\mathbf{R}'\mathbf{e}/T-\underbrace{\sqrt{T}(\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}}-\mathbf{I}_p)(\widehat{\mathbf{w}}-\mathbf{w})}_{\Delta}/\sqrt{T}. \tag{3.23}$$

In the section with the theoretical results we show that Δ is asymptotically negligible under certain sparsity assumptions[4]. Combining (3.20) and (3.23) brings us to the de-biased estimator of portfolio weights:

$$\widehat{\mathbf{w}}_{\text{MRC, DEBIASED}}=\widehat{\mathbf{w}}+\widehat{\mathbf{\Theta}}\lambda\widehat{\mathbf{g}}=\widehat{\mathbf{w}}+\widehat{\mathbf{\Theta}}\mathbf{R}'(\widehat{\mathbf{y}}-\mathbf{R}\widehat{\mathbf{w}})/T. \tag{3.24}$$

The properties of the proposed de-biased estimator are examined in Section 5.

[4] Note that we cannot directly apply Theorem 2.2 of [135] since $\mathbf{r}_c$ needs to be estimated and we first need to show consistency of the respective estimator.

3.2.2 Sparse Portfolio Using Post-Lasso

One of the drawbacks of the de-biased portfolio weights in (3.24) is that the weight formula is tailored to a specific portfolio choice that maximizes an unconstrained Sharpe Ratio (i.e. MRC in (3.13)). However, it is desirable to accommodate preferences of different types of investors who might be interested in weight allocations corresponding to GMV (3.11) or MWC (3.12) portfolios. At the same time, we are willing to stay within the framework of sparse allocations. One of the difficulties that precludes us from pursuing a similar technique as in (3.16) is the fact that once the weight constraint is added, the optimization problem in (3.16) has two solutions depending on whether $\boldsymbol{\iota}'\mathbf{\Theta}\mathbf{m}$ is positive or negative. As shown in [104], when $\boldsymbol{\iota}'\mathbf{\Theta}\mathbf{m} < 0$, the minimum value cannot be achieved exactly for a specified portfolio allocation that satisfies the full investment constraint. Hence, one can design an approximate solution to approach the supremum as closely as desired.

To overcome this difficulty, we propose to use Lasso regression in (3.19) for selecting a subset of stocks, and then constructing a financial portfolio using any of the weight formulations in (3.11)-(3.13). The procedure to estimate sparse portfolio using post-Lasso is described in Algorithm 3.

Algorithm 3: Sparse Portfolio Using Post-Lasso

```
1: Use Lasso regression in (3.19) to select the model $\widehat{\Xi} \equiv \text{support}(\widehat{\mathbf{w}})$
     • Apply additional thresholding to remove stocks with small estimated weights:
           $\widehat{\mathbf{w}}(t) = (\widehat{w}_j \mathbb{1}\left[|\widehat{w}_j| > t\right],\ j = 1, \ldots, p),$
       where $t \geq 0$ is the thresholding level.
     • The corresponding selected model is denoted as $\widehat{\Xi}(t) \equiv \text{support}(\widehat{\mathbf{w}}(t))$. When
       $t = 0$, $\widehat{\Xi}(t) = \widehat{\Xi}$.
2: Choose a desired portfolio formulation in (3.11)-(3.13) and apply it to the selected
   subset of stocks $\widehat{\Xi}(t)$.
     • When $\text{card}(\widehat{\Xi}(t)) < \widetilde{t}$, use the inverse of the sample covariance matrix as an
       estimator of $\boldsymbol{\Theta}$. Otherwise, apply the estimator of precision matrix described in
       Section 3.
```

3.3 Factor Nodewise Regression

In this section we first review a nodewise regression ([108]), a popular approach to estimate a precision matrix. After that we propose a novel estimator which accounts for the common factors in the excess returns.

In the high-dimensional settings it is necessary to regularize the precision matrix, which means that some of the entries θ_{ij} will be zero. In other words, to achieve consistent estimation of the inverse covariance, the estimated precision matrix should be sparse.

3.3.1 Nodewise Regression

One of the approaches to induce sparsity in the estimation of precision matrix is to solve for $\widehat{\mathbf{\Theta}}$ one column at a time via linear regressions, replacing population moments by their sample counterparts. When we repeat this procedure for each variable $j = 1, \ldots, p$, we will estimate the elements of $\widehat{\mathbf{\Theta}}$ column by column using $\{\mathbf{r}_t\}_{t=1}^{T}$ via p linear regressions. [108] use this approach to incorporate sparsity into the estimation of the precision matrix. They fit p separate Lasso regressions using each variable (node) as the response and the others as predictors to estimate $\widehat{\mathbf{\Theta}}$. This method is known as the "nodewise" regression and it is reviewed below based on [135] and [24].

Let $\mathbf{r}_j$ be a $T \times 1$ vector of observations for the j-th regressor, the remaining covariates are collected in a $T \times (p-1)$ matrix $\mathbf{R}_{-j}$. For each $j = 1, \ldots, p$ we run the following Lasso regressions:

$$\widehat{\boldsymbol{\gamma}}_j = \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^{p-1}} \Big(\|\mathbf{r}_j - \mathbf{R}_{-j}\boldsymbol{\gamma}\|_2^2 / T + 2\lambda_j \|\boldsymbol{\gamma}\|_1 \Big), \tag{3.25}$$

where $\widehat{\boldsymbol{\gamma}}_j = \{\widehat{\gamma}_{j,k}; j = 1, \ldots, p, k \neq j\}$ is a $(p-1) \times 1$ vector of the estimated regression coefficients that will be used to construct the estimate of the precision matrix, $\widehat{\boldsymbol{\Theta}}$. Define

$$\widehat{\mathbf{C}} = \begin{pmatrix} 1 & -\widehat{\gamma}_{1,2} & \cdots & -\widehat{\gamma}_{1,p} \\ -\widehat{\gamma}_{2,1} & 1 & \cdots & -\widehat{\gamma}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\widehat{\gamma}_{p,1} & -\widehat{\gamma}_{p,2} & \cdots & 1 \end{pmatrix}. \tag{3.26}$$

For $j = 1, \ldots, p$, define

$$\hat{\tau}_j^2 = \|\mathbf{r}_j - \mathbf{R}_{-j}\widehat{\boldsymbol{\gamma}}_j\|_2^2/T + \lambda_j \|\widehat{\boldsymbol{\gamma}}_j\|_1 \tag{3.27}$$

and write

$$\widehat{\mathbf{T}}^2 = \mathrm{diag}(\hat{\tau}_1^2, \ldots, \hat{\tau}_p^2). \tag{3.28}$$

The approximate inverse is defined as

$$\widehat{\boldsymbol{\Theta}} = \widehat{\mathbf{T}}^{-2}\widehat{\mathbf{C}}. \tag{3.29}$$

The procedure to estimate the precision matrix using nodewise regression is summarized in Algorithm 4.

Algorithm 4: Nodewise Regression by [108] (MB)

```
1: Repeat for $j = 1, \ldots, p$ :
    • Estimate $\widehat{\gamma}_j$ using (3.25) for a given $\lambda_j$.
    • Select $\lambda_j$ using a suitable information criterion.
2: Calculate $\widehat{\mathbf{C}}$ and $\widehat{\mathbf{T}}^2$ .
3: Return $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{T}}^{-2}\widehat{\mathbf{C}}$.
```

One of the caveats to keep in mind when using the nodewise regression method is that the estimator in (3.29) is not self-adjoint. [24] show (see their Lemma A.1) that $\widehat{\mathbf{\Theta}}$ in (3.29) is positive definite with high probability, however, it could still occur that $\widehat{\mathbf{\Theta}}$ is not positive definite in finite samples. To resolve this issue we use the matrix symmetrization procedure as in [56] and then use eigenvalue cleaning as in [25] and [72]. First, the symmetric matrix is constructed as

$$\widehat{\theta}_{ij}^s = \widehat{\theta}_{ij}\mathbb{1}\left[\left|\widehat{\theta}_{ij}\right| \leq \left|\widehat{\theta}_{ji}\right|\right] + \widehat{\theta}_{ji}\mathbb{1}\left[\left|\widehat{\theta}_{ij}\right| > \left|\widehat{\theta}_{ji}\right|\right], \tag{3.30}$$

where $\widehat{\theta}_{ij}$ is the (i, j)-th element of the estimated precision matrix from (3.29). Second, we use eigenvalue cleaning to make $\widehat{\mathbf{\Theta}}^s$ positive definite: write the spectral decomposition $\widehat{\mathbf{\Theta}}^s = \widehat{\mathbf{V}}'\widehat{\mathbf{\Lambda}}\widehat{\mathbf{V}}$, where $\widehat{\mathbf{V}}$ is a matrix of eigenvectors and $\widehat{\mathbf{\Lambda}}$ is a diagonal matrix with p eigenvalues $\widehat{\mathbf{\Lambda}}_i$ on its diagonal. Let $\mathbf{\Lambda}_m \equiv \min\{\widehat{\mathbf{\Lambda}}_i | \widehat{\mathbf{\Lambda}}_i > 0\}$. We replace all $\widehat{\mathbf{\Lambda}}_i < \mathbf{\Lambda}_m$ with $\mathbf{\Lambda}_m$ and define the diagonal matrix with cleaned eigenvalues as $\widetilde{\mathbf{\Lambda}}$. We use $\widetilde{\mathbf{\Theta}} = \widehat{\mathbf{V}}'\widetilde{\mathbf{\Lambda}}\widehat{\mathbf{V}}$ which is symmetric and positive definite.

3.3.2 Factor Nodewise Regression

The arbitrage pricing theory (APT), developed by [119], postulates that expected returns on securities should be related to their covariance with the common components or factors only. The goal of the APT is to model the tendency of asset returns to move together via factor decomposition. Assume that the return generating process ($\mathbf{r}_t$) follows a K-factor model:

$$\underbrace{\mathbf{r}_t}_{p\times 1} = \mathbf{B}\underbrace{\mathbf{f}_t}_{K\times 1} + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T \tag{3.31}$$

where $\mathbf{f}_t = (f_{1t}, \ldots, f_{Kt})'$ are the factors, $\mathbf{B}$ is a $p \times K$ matrix of factor loadings, and $\boldsymbol{\varepsilon}_t$ is the idiosyncratic component that cannot be explained by the common factors. Factors in (3.31) can be either observable, such as in [46, 47], or can be estimated using statistical factor models.

In this subsection we examine how to approach the portfolio allocation problems in (3.11)-(3.13) using a factor structure in the returns. Our approach, called *Factor Nodewise Regression*, uses the estimated common factors to obtain sparse precision matrix of the idiosyncratic component. The resulting estimator is used to obtain the precision of the asset returns necessary to form portfolio weights.

As in [53], we consider a spiked covariance model when the first K principal eigenvalues of $\boldsymbol{\Sigma}$ are growing with p, while the remaining $p - K$ eigenvalues are bounded and grow slower than p.

Rewrite equation (3.31) in matrix form:

$$\underbrace{\mathbf{R}}_{p\times T} = \underbrace{\mathbf{B}}_{p\times K}\mathbf{F} + \mathbf{E}. \tag{3.32}$$

Let $\mathbf{\Sigma} = T^{-1}\mathbf{R}\mathbf{R}'$, $\mathbf{\Sigma}_{\varepsilon} = T^{-1}\mathbf{E}\mathbf{E}'$ and $\mathbf{\Sigma}_f = T^{-1}\mathbf{F}\mathbf{F}'$ be covariance matrices of stock returns, idiosyncratic components and factors, and let $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, $\mathbf{\Theta}_{\varepsilon} = \mathbf{\Sigma}_{\varepsilon}^{-1}$ and $\mathbf{\Theta}_f = \mathbf{\Sigma}_f^{-1}$ be their inverses. The factors and loadings in (3.32) are estimated by solving $(\widehat{\mathbf{B}}, \widehat{\mathbf{F}}) = \text{argmin}_{\mathbf{B},\mathbf{F}}\|\mathbf{R} - \mathbf{B}\mathbf{F}\|_F^2$ s.t. $\frac{1}{T}\mathbf{F}\mathbf{F}' = \mathbf{I}_K,\ \mathbf{B}'\mathbf{B}$ is diagonal. The constraints are needed to identify the factors ([56]). It was shown ([126]) that $\widehat{\mathbf{F}} = \sqrt{T}\text{eig}_K(\mathbf{R}'\mathbf{R})$ and $\widehat{\mathbf{B}} = T^{-1}\mathbf{R}\widehat{\mathbf{F}}'$. Given $\widehat{\mathbf{F}}, \widehat{\mathbf{B}}$, define $\widehat{\mathbf{E}} = \mathbf{R} - \widehat{\mathbf{B}}\widehat{\mathbf{F}}$.

Since our interest is in constructing portfolio weights, our goal is to estimate a precision matrix of the excess returns. However, as pointed out by [87], when common factors are present across the excess returns, the precision matrix cannot be sparse because all pairs of the returns are partially correlated given other excess returns through the common factors. Therefore, we impose a sparsity assumption on the precision matrix of the idiosyncratic errors, $\mathbf{\Theta}_{\varepsilon}$, which is obtained using the estimated residuals after removing the co-movements induced by the factors (see [11, 18, 87]).

We use the nodewise regression as a shrinkage technique to estimate the precision matrix of residuals. Once the precision $\mathbf{\Theta}_f$ of the low-rank component is also obtained, similarly to [52], we use the Sherman-Morrison-Woodbury formula to estimate the precision of excess returns:

$$\mathbf{\Theta} = \mathbf{\Theta}_{\varepsilon} - \mathbf{\Theta}_{\varepsilon}\mathbf{B}[\mathbf{\Theta}_f + \mathbf{B}'\mathbf{\Theta}_{\varepsilon}\mathbf{B}]^{-1}\mathbf{B}'\mathbf{\Theta}_{\varepsilon}. \tag{3.33}$$

To obtain $\widehat{\mathbf{\Theta}}_f = \widehat{\mathbf{\Sigma}}_f^{-1}$, we use the inverse of the sample covariance of the estimated factors $\widehat{\mathbf{\Sigma}}_f = T^{-1}\widehat{\mathbf{F}}\widehat{\mathbf{F}}'$. To get $\widehat{\mathbf{\Theta}}_\varepsilon$, we apply Algorithm 4 to the estimated idiosyncratic errors, $\widehat{\boldsymbol{\varepsilon}}_t$. Once we have estimated $\widehat{\mathbf{\Theta}}_f$ and $\widehat{\mathbf{\Theta}}_\varepsilon$, we can get $\widehat{\mathbf{\Theta}}$ using a sample analogue of (3.33). The proposed procedure is called *Factor Nodewise Regression* and is summarized in Algorithm 5.

Algorithm 5: Factor Nodewise Regression by [108] (FMB)

1: Estimate factors, $\widehat{\mathbf{F}}$, and factor loadings, $\widehat{\mathbf{B}}$, using PCA. Obtain $\widehat{\mathbf{\Sigma}}_f = T^{-1}\widehat{\mathbf{F}}\widehat{\mathbf{F}}'$, $\widehat{\mathbf{\Theta}}_f = \widehat{\mathbf{\Sigma}}_f^{-1}$ and $\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{r}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t$.

2: Estimate a sparse $\mathbf{\Theta}_\varepsilon$ using nodewise regression: run Lasso regressions in (3.25) for $\widehat{\boldsymbol{\varepsilon}}_t$

$$\widehat{\gamma}_j = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{p-1}} \left(\left\| \widehat{\varepsilon}_j - \widehat{\mathbf{E}}_{-j}\boldsymbol{\gamma} \right\|_2^2 / T + 2\lambda_j \|\boldsymbol{\gamma}\|_1 \right), \tag{3.34}$$

to get $\widehat{\mathbf{\Theta}}_\varepsilon$.

3: Use $\widehat{\mathbf{\Theta}}_f$ from Step 1 and $\widehat{\mathbf{\Theta}}_\varepsilon$ from Step 2 to estimate $\mathbf{\Theta}$ using the sample counterpart of the Sherman-Morrison-Woodbury formula in (2.16):

$$\widehat{\mathbf{\Theta}} = \widehat{\mathbf{\Theta}}_\varepsilon - \widehat{\mathbf{\Theta}}_\varepsilon\widehat{\mathbf{B}}[\widehat{\mathbf{\Theta}}_f + \widehat{\mathbf{B}}'\widehat{\mathbf{\Theta}}_\varepsilon\widehat{\mathbf{B}}]^{-1}\widehat{\mathbf{B}}'\widehat{\mathbf{\Theta}}_\varepsilon. \tag{3.35}$$

Algorithm 5 involves a tuning parameter λ_j in (3.34): we choose shrinkage intensity by minimizing the generalized information criterion (GIC). Let $\left|\widehat{S}_j(\lambda_j)\right|$ denote the estimated number of nonzero parameters in the vector $\widehat{\boldsymbol{\gamma}}_j$:

$$\text{GIC}(\lambda_j) = \log\left(\left\|\widehat{\boldsymbol{\varepsilon}}_j - \widehat{\mathbf{E}}_{-j}\widehat{\boldsymbol{\gamma}}_j\right\|_2^2/T\right) + \left|\widehat{S}_j(\lambda_j)\right|\frac{\log(p)}{T}\log(\log(T)).$$

We can use $\widehat{\boldsymbol{\Theta}}$ obtained in (3.35) to estimate y in equation (3.17) and obtain sparse portfolio weights in (3.24) and Algorithm 3.

3.4 Factor Investing is Allowed

In this section we allow an investor to hold a portfolio of assets and factors, in other words, factors are assumed to be tradable. Note that in contrast with [2], the distinction between tradable and non-tradable factors is not pinned down by the fact that the excess returns are driven by the common factors. That is, factor structure of returns is allowed independently of whether factors are tradable or not. We assume that only observable factors can be tradable. Denote a $K_1 \times 1$ vector of observable factors as $\widetilde{\mathbf{f}}_t$, and $K_2 \times 1$ vector of unobservable factors as $\mathbf{f}_t^{PCA}$, where $K1 + K2 = K$. The goal of factor investing is to decide how much weight is allocated to factors $\widetilde{\mathbf{f}}_t$ and stocks $\mathbf{r}_t$. Let $r_{t,all}$ be the return of portfolio at time t:

$$r_{t,all} = \underbrace{\mathbf{w}'_{all,t}}_{1\times(p+K_1)}\mathbf{x}_t. \tag{3.36}$$

where $\mathbf{x}_t = (\widetilde{\mathbf{f}}'_t, \mathbf{r}'_t)'$ is a $(p + K_1) \times 1$ vector of excess returns of observable factors and stocks and $\mathbf{w}_{all,t} = (\mathbf{w}'_{ft}, \mathbf{w}'_t)'$ is a vector of weights with $\mathbf{w}_{ft}$ invested in $\widetilde{\mathbf{f}}_t$ and $\mathbf{w}_t$ invested in stocks. We treat $\widetilde{\mathbf{f}}_t$ as additional K_1 investments vehicles which will contribute to the return of the total portfolio. Now consider K_2-factor model for $\mathbf{x}_t$:

$$\mathbf{x}_t = \mathbf{B}\underbrace{\mathbf{f}_t^{PCA}}_{K_2\times 1} + \mathbf{e}_t, \quad t = 1, \dots, T \tag{3.37}$$

Rewrite equation (3.37) in matrix form:

$$\underbrace{\mathbf{X}}_{(p+K_1)\times T} = \mathbf{B}\underbrace{\mathbf{F}^{PCA}}_{K_2\times T} + \mathbf{E}, \tag{3.38}$$

which can be estimated using the standard PCA techniques as in (3.32): $\widehat{\mathbf{F}}^{PCA} = \sqrt{T}\text{eig}_{K_2}(\mathbf{X}'\mathbf{X})$ and $\widehat{\mathbf{B}} = T^{-1}\mathbf{X}\widehat{\mathbf{F}}'^{PCA}$. Given $\widehat{\mathbf{F}}^{PCA}, \widehat{\mathbf{B}}$, define $\widehat{\mathbf{E}} = \mathbf{X} - \widehat{\mathbf{B}}\widehat{\mathbf{F}}^{PCA}$.

Similarly to Algorithm 5, we use (2.16) to estimate the precision of the augmented excess returns, $\mathbf{\Theta}_x$. To get $\widehat{\mathbf{\Theta}}_{f^{PCA}} = \widehat{\mathbf{\Sigma}}_{f^{PCA}}^{-1}$, we use the inverse of the sample covariance of the estimated factors $\widehat{\mathbf{\Sigma}}_{f^{PCA}} = T^{-1}\widehat{\mathbf{F}}^{PCA}\widehat{\mathbf{F}}'^{PCA}$. To get $\widehat{\mathbf{\Theta}}_e$, we first apply Algorithm 4 to the estimated idiosyncratic errors, $\widehat{\mathbf{e}}_t$ in (3.37). Once we have estimated $\widehat{\mathbf{\Theta}}_{f^{PCA}}$ and $\widehat{\mathbf{\Theta}}_e$, we can get $\widehat{\mathbf{\Theta}}_x$ using a sample analogue of (2.16). This procedure is summarized in Algorithm 6.

Algorithm 6: Factor Investing Using FMB

```
Estimate the residuals from equation (3.37): $\widehat{\mathbf{e}}_t = \mathbf{x}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t^{PCA}$ using PCA.
Estimate a sparse $\mathbf{\Theta}_e$ using nodewise regression: apply Algorithm 4 to $\widehat{e}_t$.
Estimate $\mathbf{\Theta}_x$ using the Sherman-Morrison-Woodbury formula in (2.16).
```

We can use $\widehat{\boldsymbol{\Theta}}_x$ obtained from Algorithm 6 to estimate portfolio weights $\mathbf{w}_{all,t}$ using either a de-biased technique from section 2.1 ((3.24)), or post-Lasso (Algorithm 3). Once we obtain $\widehat{\mathbf{w}}_{all,t} = (\widehat{\mathbf{w}}'_{ft}, \widehat{\mathbf{w}}'_t)'$, we can test whether factor investing significantly contributes to the portfolio return by testing whether $\mathbf{w}_{ft} = 0$.

3.5 Asymptotic Properties

In this section we study asymptotic properties of the de-biased estimator of weights for sparse portfolio in (3.24) and post-Lasso estimator from Algorithm 3.

Denote $S_0 \equiv \{j; \mathbf{w}_j \neq 0\}$ to be the active set of variables, where $\mathbf{w}$ is a vector of true portfolio weights in equation (3.18). Also, let $s_0 \equiv |S_0|$. Further, let $S_j \equiv \{k; \gamma_{j,k} \neq 0\}$ be the active set for row $\boldsymbol{\gamma}_j$ for the nodewise regression in (3.25), and let $s_j \equiv |S_j|$. Define $\bar{s} \equiv \max_{1\leq j\leq p} s_j$.

Consider a factor model from equation (3.31):

$$\underbrace{\mathbf{r}_t}_{p\times 1} = \mathbf{B} \underbrace{\mathbf{f}_t}_{K\times 1} + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T \tag{3.39}$$

We study the case when the factors are not known, i.e. the only observable variable in equation (3.39) is the excess returns $\mathbf{r}_t$. In this paper our main interest lies in establishing asymptotic properties of sparse portfolio weights and the out-of-sample Sharpe Ratio for the high-dimensional case. We assume that the number of common factors, K, is fixed.

3.5.1 Assumptions

We now list the assumptions on the model (2.11):

(A.1) (Spiked covariance model) As $p \to \infty$, $\Lambda_1(\mathbf{\Sigma}) > \Lambda_2(\mathbf{\Sigma}) > \ldots > \Lambda_K(\mathbf{\Sigma}) \gg \Lambda_{K+1}(\mathbf{\Sigma}) \geq \ldots \geq \Lambda_p(\mathbf{\Sigma}) \geq 0$, where $\Lambda_j(\mathbf{\Sigma}) = \mathcal{O}(p)$ for $j \leq K$, while the non-spiked eigenvalues are bounded, $\Lambda_j(\mathbf{\Sigma}) = o(p)$ for $j > K$.

(A.2) (Pervasive factors) There exists a positive definite $K \times K$ matrix $\breve{\mathbf{B}}$ such that $\left|\left|\left| p^{-1}\mathbf{B}'\mathbf{B} - \breve{\mathbf{B}} \right|\right|\right|_2 \to 0$ and $\Lambda_{\min}(\breve{\mathbf{B}})^{-1} = \mathcal{O}(1)$ as $p \to \infty$.

Similarly to [30] and [24], we also impose beta mixing condition.

(A.3) (Beta mixing) Let $\mathcal{F}_{-\infty}^{t}$ and $\mathcal{F}_{t+k}^{\infty}$ denote the σ-algebras that are generated by $\{\boldsymbol{\varepsilon}_u : u \leq t\}$ and $\{\boldsymbol{\varepsilon}_u : u \geq t+k\}$ respectively. Then $\{\boldsymbol{\varepsilon}\}_u$ is β-mixing in the sense that $\beta_k \to 0$ as $k \to \infty$, where the mixing coefficient is defined as

$$\beta_k = \sup_t \mathbb{E}\left[\sup_{B \in \mathcal{F}_{t+k}^{\infty}} \left|\Pr\left(B|\mathcal{F}_{-\infty}^{t}\right) - \Pr\left(B\right)\right|\right]. \tag{3.40}$$

Some comments regarding the aforementioned assumptions are in order. Assumptions **(A.1)**-**(A.2)** are the same as in [56], and assumption **(A.3)** is required to consistently estimate precision matrix for de-biasing portfolio weights. Assumption **(A.1)** divides the eigenvalues into the diverging and bounded ones. Without loss of generality, we assume that K largest eigenvalues have multiplicity of 1. The assumption of a spiked covariance model is common in the literature on approximate factor models, however, we note that the model studied in this paper can be characterized as a "very spiked model". In other words, the gap between the first K eigenvalues and the rest is increasing with p. As pointed out by [56], **(A.1)** is typically satisfied by the factor model with pervasive factors, which brings us to the assumption **(A.2)**: the factors impact a non-vanishing proportion of individual time-series. Assumption **(A.3)** allows for weak dependence in the residuals of the factor model in 2.11: causal ARMA processes, certain stationary Markov chains and stationary GARCH models with finite second moments satisfy this assumption. We note that our Assumption **(A.3)** is much weaker than in [24], the latter requires weak dependence of the returns series, whereas we only restrict dependence of the idiosyncratic components.

Let $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}'$, where $\mathbf{\Sigma}$ is the covariance matrix of returns that follow factor structure described in equation (2.11). Define $\widehat{\mathbf{\Sigma}}, \widehat{\mathbf{\Lambda}}_K, \widehat{\mathbf{\Gamma}}_K$ to be the estimators of $\mathbf{\Sigma}, \mathbf{\Lambda}, \mathbf{\Gamma}$. We further let $\widehat{\mathbf{\Lambda}}_K = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_K)$ and $\widehat{\mathbf{\Gamma}}_K = (\hat{v}_1, \ldots, \hat{v}_K)$ to be constructed by the first K leading empirical eigenvalues and the corresponding eigenvectors of $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{B}}\widehat{\mathbf{B}}' = \widehat{\mathbf{\Gamma}}_K\widehat{\mathbf{\Lambda}}_K\widehat{\mathbf{\Gamma}}'_K$. Similarly to [56], we require the following bounds on the componentwise maximums of the estimators:

(B.1) $\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_{\max} = \mathcal{O}_P\Big(\sqrt{\log p/T}\Big)$,

(B.2) $\left\|(\widehat{\mathbf{\Lambda}}_K - \mathbf{\Lambda})\mathbf{\Lambda}^{-1}\right\|_{\max} = \mathcal{O}_P\Big(\sqrt{\log p/T}\Big)$,

(B.3) $\left\|\widehat{\mathbf{\Gamma}}_K - \mathbf{\Gamma}\right\|_{\max} = \mathcal{O}_P\Big(\sqrt{\log p/(Tp)}).$

Let $\widehat{\mathbf{\Sigma}}^{SG}$ be the sample covariance matrix, with $\widehat{\mathbf{\Lambda}}_K^{SG}$ and $\widehat{\mathbf{\Gamma}}_K^{SG}$ constructed with the first K leading empirical eigenvalues and eigenvectors of $\widehat{\mathbf{\Sigma}}^{SG}$ respectively. Also, let $\widehat{\mathbf{\Sigma}}^{EL1} = \widehat{\mathbf{D}}\widehat{\mathbf{R}}_1\widehat{\mathbf{D}}$, where $\widehat{\mathbf{R}}_1$ is obtained using the Kendall's tau correlation coefficients and $\widehat{\mathbf{D}}$ is a robust estimator of variances constructed using the Huber loss. Furthermore, let $\widehat{\mathbf{\Sigma}}^{EL2} = \widehat{\mathbf{D}}\widehat{\mathbf{R}}_2\widehat{\mathbf{D}}$, where $\widehat{\mathbf{R}}_2$ is obtained using the spatial Kendall's tau estimator. Define $\widehat{\mathbf{\Lambda}}_K^{EL}$ to be the matrix of the first K leading empirical eigenvalues of $\widehat{\mathbf{\Sigma}}^{EL1}$, and $\widehat{\mathbf{\Gamma}}_K^{EL}$ is the matrix of the first K leading empirical eigenvectors of $\widehat{\mathbf{\Sigma}}^{EL2}$. For more details regarding constructing $\widehat{\mathbf{\Sigma}}^{SG}$, $\widehat{\mathbf{\Sigma}}^{EL1}$ and $\widehat{\mathbf{\Sigma}}^{EL2}$ see [56], Sections 3 and 4.

Theorem 6 *([56])*

*For sub-Gaussian distributions, $\widehat{\mathbf{\Sigma}}^{SG}$, $\widehat{\mathbf{\Lambda}}_K^{SG}$ and $\widehat{\mathbf{\Gamma}}_K^{SG}$ satisfy **(B.1)-(B.3)**.*

*For elliptical distributions, $\widehat{\mathbf{\Sigma}}^{EL1}$, $\widehat{\mathbf{\Lambda}}_K^{EL}$ and $\widehat{\mathbf{\Gamma}}_K^{EL}$ satisfy **(B.1)-(B.3)**.*

Theorem 6 is essentially a rephrasing of the results obtained in [56], Sections 3 and 4. Since there is no separate statement of these results in their paper (it is rather a summary of several theorems), we separated it as a Theorem for the convenience of the reader. As evidenced from the above Theorem, $\widehat{\mathbf{\Sigma}}^{EL2}$ is only used for estimating the eigenvectors. This is necessary due to the fact that, in contrast with $\widehat{\mathbf{\Sigma}}^{EL2}$, the theoretical properties of the eigenvectors of $\widehat{\mathbf{\Sigma}}^{EL}$ are mathematically involved because of the sin function.

In addition, the following structural assumption on the model is imposed:

(C.1) $\|\mathbf{\Sigma}\|_{\max} = \mathcal{O}(1)$ and $\|\mathbf{B}\|_{\max} = \mathcal{O}(1)$,

which is a natural assumption on the population quantities.

In contrast to [56], instead of estimating and inverting covariance matrix, we focus on obtaining precision matrix directly since it is the ultimate input to any portfolio optimization problem.

3.5.2 Asymptotic Properties of Non-Sparse Portfolio Weights

Recall that we used equation (2.16) to estimate $\mathbf{\Theta}$. Therefore, in order to establish consistency of the estimator in (2.16), we first show consistency of $\widehat{\mathbf{\Theta}}_{\varepsilon}$. Proofs of all the theorems are in Appendix.

Theorem 7 *Suppose that Assumptions* ***(A.1)-(A.3)***, ***(B.1)-(B.3)*** *and* ***(C.1)*** *hold. Let* $\omega_T \equiv \sqrt{\log p/T}+1/\sqrt{p}$. *Then* $\max_{i\leq p}(1/T)\sum_{t=1}^{T}|\hat{\varepsilon}_{it}-\varepsilon_{it}| = \mathcal{O}_P(\omega_T^2)$ *and* $\max_{i,t}|\hat{\varepsilon}_{it}-\varepsilon_{it}| = \mathcal{O}_P(\omega_T) = o_P(1)$. *Under the sparsity assumption* $\bar{s}^2\omega_T = o(1)$, *with* $\lambda_j \asymp \omega_T$, *we have*

$$\max_{1\leq j\leq p}\left\|\widehat{\mathbf{\Theta}}_{\varepsilon,j}-\mathbf{\Theta}_{\varepsilon,j}\right\|_1 = \mathcal{O}_P(\bar{s}\omega_T),$$
$$\max_{1\leq j\leq p}\left\|\widehat{\mathbf{\Theta}}_{\varepsilon,j}-\mathbf{\Theta}_{\varepsilon,j}\right\|_2^2 = \mathcal{O}_P(\bar{s}\omega_T^2)$$

Some comments are in order. First, the sparsity assumption $\bar{s}^2\omega_T = o(1)$ is stronger than that required for convergence of $\widehat{\mathbf{\Theta}}_\varepsilon$: this is necessary to ensure consistency for $\widehat{\mathbf{\Theta}}$ established in Theorem 8, so we impose a stronger assumption at the beginning. We also note that at the first glance, our sparsity assumption in Theorem 8 is stronger than that required by [135] and [24], however, recall that we impose sparsity on $\mathbf{\Theta}_\varepsilon$, not $\mathbf{\Theta}$ as opposed to the two aforementioned papers. Hence, this assumption can be easily satisfied once the common factors have been accounted for and the precision of the idiosyncratic components is expected to be sparse. The bounds derived in Theorem 7 help us establish the convergence properties of the precision matrix of stock returns in equation (2.16).

Theorem 8 *Under the assumptions of Theorem 7 and, in addition, assuming* $\left\|\mathbf{\Theta}_{\varepsilon,j}\right\|_2 = \mathcal{O}(1)$, *we have*

$$\max_{1\leq j\leq p}\left\|\widehat{\mathbf{\Theta}}_{j}-\mathbf{\Theta}_{j}\right\|_1 = \mathcal{O}_P(\bar{s}^2\omega_T),$$
$$\max_{1\leq j\leq p}\left\|\widehat{\mathbf{\Theta}}_{j}-\mathbf{\Theta}_{j}\right\|_2^2 = \mathcal{O}_P(\bar{s}\omega_T^2).$$

Using Theorem 8 we can then establish the consistency of the non-sparse counterpart of the estimated MRC portfolio weight in (3.19).

Theorem 9 *Under the assumptions of Theorem 8, Algorithm 5 consistently estimates non-sparse MRC portfolio weights such that* $\|\widehat{\mathbf{w}}_{MRC} - \mathbf{w}_{MRC}\|_1 = \mathcal{O}_P(\bar{s}^2\omega_T)$.

Note that the rate in Theorem 9 depends on the sparsity of $\mathbf{\Theta}_\varepsilon$. If, instead, sparsity on $\mathbf{\Theta}$ is imposed, the rate becomes similar to the one derived by [24]: $\bar{s}(\mathbf{\Theta})^{3/2}\omega_T = o_P(1)$, where $\bar{s}(\mathbf{\Theta})$ is the maximum vertex degree of $\mathbf{\Theta}$. In their case, if the precision matrix of stock returns is not sparse, consistent estimation of portfolio weights is possible if $(p-1)^{3/2}(\sqrt{\log p/T} + 1/\sqrt{p}) = o(1)$. However, this excludes high-dimensional cases since p is required to be less than $T^{1/3}$.

3.5.3 Asymptotic Properties of De-Biased Portfolio Weights

We now proceed to examining the properties of sparse MRC portfolio weights for de-biased portfolio, as summarized by the following Theorem:

Theorem 10 *Let* $\widehat{\mathbf{\Sigma}}$ *be an estimator of covariance matrix satisfying **(B.1)**, and* $\widehat{\mathbf{\Theta}}$ *be the estimator of precision obtained using FMB in Algorithm 5. Under the assumptions of Theorem 8, consider the linear model* (3.18) *with* $\mathbf{e} \sim \mathcal{D}(\mathbf{0}, \sigma_e\mathbf{I})$, *where* $\sigma_e^2 = \mathcal{O}(1)$. *Consider a suitable choice of the regularization parameters* $\lambda \asymp \omega_T$ *for the Lasso regression in* (3.19) *and* $\lambda_j \asymp \omega_T$ *uniformly in* j *for the Lasso for nodewise regression in* (3.25). *Assume*

$(s_0 \vee \bar{s}^2)\Big(\log p/\sqrt{T} + \sqrt{T}/p\Big) = o(1)$. Then

$$\sqrt{T}(\widehat{\mathbf{w}}_{DEBIASED} - \mathbf{w}) = W + \Delta,$$

$$W = \widehat{\mathbf{\Theta}}\mathbf{R}'\mathbf{e}/\sqrt{T},$$

$$\left\|\Delta\right\|_\infty = \mathcal{O}_P\Big((s_0 \vee \bar{s}^2)\Big(\log p/\sqrt{T} + \sqrt{T}/p\Big)\Big) = o_P(1).$$

Furthermore, if $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2\mathbf{I})$, let $\widehat{\mathbf{\Omega}} \equiv \widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Theta}}'$. Then $W|\mathbf{R} \sim \mathcal{N}_p(\mathbf{0}, \sigma_e^2\widehat{\mathbf{\Omega}})$ and $\left\|\widehat{\mathbf{\Omega}} - \mathbf{\Theta}\right\|_\infty = o_P(1)$.

Some comments are in order. Our Theorem 8 is an extension of Theorem 2.4 of [135] for non-iid case, where the latter is achieved with a help of [30]. Furthermore, there are several fundamental differences between Theorem 8 and Theorem 2.4 of [135]: first, we apply nodewise regression to estimate sparse precision matrix of factor-adjusted returns, which explains the difference in convergence rates. Concretely, [135] have $\omega_T = \sqrt{\log p/T}$, whereas we have $\omega_T = \sqrt{\log p/T} + 1/\sqrt{p}$, where $1/\sqrt{p}$ arises due to the fact that factors need to be estimated. However, we note that since we deal with high-dimensional regime $p \geq T$, this additional term is asymptotically negligible, we only keep it for identification purposes. Second, in contrast with [135], the dependent variable in the Lasso regression in (3.19) is unknown and needs to be estimated. Lemma 13 shows that $\widehat{y}$ constructed using the precision matrix estimator from Theorem 8 is consistent and shares the same rate as the ℓ_1-bound in Theorem 8. Third, interestingly, the sparsity assumption on the Lasso regression in (3.19) is the same as in [135]: as shown in the Appendix, this condition is still sufficient to ensure that the bias term is asymptotically negligible even when the stock returns follow factor structure with unknown factors. Once we impose Gaussianity of $\mathbf{e}$ in (3.18), we can infer the distribution of portfolio weights. Note that in this case normally distributed errors do not imply that the stock returns are also Gaussian: we did not assume $\varepsilon_t \sim \mathcal{N}_p(\cdot)$ in (2.11). The unknown σ_e^2 can be replaced by a consistent estimator. Finally, even when Gaussianity of $\mathbf{e}$ is relaxed, we can use the central limit theorem argument to obtain approximate Gaussianity of components of $W|\mathbf{R}$ of fixed dimension, or moderately growing dimensions (see [135] for more details), however, in order not to divert the focus of this paper, we leave it for future research.

3.5.4 Asymptotic Properties of Post-Lasso Portfolio Weights

To establish the properties of the post-Lasso estimator in Algorithm 3, we focus on MRC weight formulation, since it satisfies the standard post-Lasso assumptions. For GMV and MWC formulations, the procedure described in Algorithm 3 is not "post-Lasso" in the usual sense. Concretely, the latter assumes that both steps in Algorithm 3 have the same objective function, which is violated for GMV and MWC. Consequently, we leave rigorous theoretical derivations of these two portfolio formulation for future research. For MRC, we use the post-model selection results established in [13]. Specifically, we have the following theorem:

Theorem 11 *Suppose the restricted eigenvalue condition and the restricted sparse eigenvalue condition on the empirical Gram matrix hold (see Condition RE($\bar{c}$) and Condition RSE(m) of [13], p. 529). Let $\widehat{\mathbf{w}}$ be the post-Lasso weight estimator from Algorithm 3, we have*

$$\|\widehat{\mathbf{w}} - \mathbf{w}\|_1 = \mathcal{O}_P \begin{cases} \sigma_e\Big((s_0\omega_T) \vee (\bar{s}^2\omega_T)\Big), & \textit{in general,} \\ \sigma_e s_0\Big(\sqrt{\frac{1}{T}} + \frac{1}{\sqrt{p}}\Big), & \textit{if } s_0 \geq \bar{s}^2 \textit{ and } \Xi = \hat{\Xi} \text{ wp} \to 1. \end{cases}$$

The proof of Theorem 11 easily follows from the proof of Corollary 2 of [13] and is omitted here. Let us comment on the upper bounds for post-Lasso estimator: first, the term $(s_0\omega_T)\vee(\bar{s}^2\omega_T)$ appears since one needs to estimate the dependent variable in equation (3.18), which creates the difference between the bound in [13] and our Theorem 11. Second, similarly to [13], the upper bound undergoes a transition from the oracle rate enjoyed by the standard Lasso to the faster rate that improves on the latter when (1) the precision matrix of the idiosyncratic components is sparse enough and (2) the oracle model has well-separated coefficients. Noticeably, the upper bounds in Theorem 11 hold despite the fact that the first-stage Lasso regression in Algorithm 3 may fail to correctly select the oracle model Ξ as a subset, that is, $\Xi \notin \hat{\Xi}$.

Finally, let us compare the rates of non-sparse MRC portfolio weights in Theorem 9, de-biased weights in Theorem 8, and post-lasso weights in Theorem 11: de-biased estimator exhibits fastest convergence, followed by post-lasso and non-sparse weights. This result is further supported by our simulations presented in the next section.

3.6 Monte Carlo

We study the consistency for estimating portfolio weights in (2.10) of (i) sparse portfolios that use the standard Lasso without de-biasing in (3.19), (ii) Lasso with de-biasing in (3.24), (iii) post-Lasso in Algorithm 3, and (iv) non-sparse portfolios that use FMB from Algorithm 5. Our simulation results are divided into two parts: the first part examines the performance of models (i)-(iv) under the Gaussian setting, and the second part examines the robustness of performance under the elliptical distributions (to be described later). Each part is further subdivided into two cases: with $p < T$ (**Case 1**) and with $p > T$ (**Case 2**), in both cases we allow the number of stocks to increase with the sample size, i.e. $p = p_T \to \infty$ as $T \to \infty$. In Case 1 we let $p = T^{\delta}$, $\delta = 0.85$ and $T = [2^h]$, for $h = 7, 7.5, 8, \ldots, 9.5$, in Case 2 we let $p = 3 \cdot T^{\delta}$, $\delta = 0.85$, all else equal.

First, consider the following data generating process for stock returns:

$$\underbrace{\mathbf{r}_t}_{p\times 1} = \mathbf{m} + \mathbf{B} \underbrace{\mathbf{f}_t}_{K\times 1} + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T \tag{3.41}$$

where $\mathbf{m}_i \sim \mathcal{N}(1,1)$ independently for each $i = 1, \ldots, p$, $\boldsymbol{\varepsilon}_t$ is a $p \times 1$ random vector of idiosyncratic errors following $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\varepsilon})$, with a Toeplitz matrix $\boldsymbol{\Sigma}_{\varepsilon}$ parameterized by ρ: that is, $\boldsymbol{\Sigma}_{\varepsilon} = (\boldsymbol{\Sigma}_{\varepsilon})_{ij}$, where $(\boldsymbol{\Sigma}_{\varepsilon})_{ij} = \rho^{|i-j|}$, $i, j \in 1, \ldots, p$ which leads to sparse $\boldsymbol{\Theta}_{\varepsilon}$, $\mathbf{f}_t$ is a $K \times 1$ vector of factors drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_f = \mathbf{I}_K/10)$, $\mathbf{B}$ is a $p \times K$ matrix of factor loadings drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I}_K/100)$. We set $\rho = 0.5$ and fix the number of factors $K = 3$.

Let $\boldsymbol{\Sigma} = \mathbf{B}\boldsymbol{\Sigma}_f\mathbf{B}' + \boldsymbol{\Sigma}_{\varepsilon}$. To create sparse MRC portfolio weights we use the following procedure: first, we threshold the vector $\boldsymbol{\Sigma}^{-1}\mathbf{m}$ to keep the top $p/2$ entries with largest absolute values. This yields sparse vector $\boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}\mathbf{m}$ defined in (3.13). We use $\boldsymbol{\Sigma}\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ as the values for the mean and covariance matrix parameters to generate multivariate Gaussian returns in (3.41). Note that the low rank plus sparse structure of the covariance matrix is preserved under this transformation.

Figure 3.6.1 shows the averaged (over Monte Carlo simulations) errors of the estimators of the weight $\mathbf{w}_{\text{MRC}}$ versus the sample size T in the logarithmic scale (base 2). As evidenced by Figure 3.6.1, (1) sparse estimators outperform non-sparse counterparts; (2) using de-biasing or post-Lasso improves the performance compared to the standard Lasso estimator. As expected from Theorems 8-11, the Lasso, de-biased Lasso and post-Lasso exhibit similar rates, but the two latter estimators enjoy lower estimation error. The ranking remains similar for Case 2, however, as illustrated in Figure 3.6.1, the performance of all estimators slightly deteriorates.

Gaussian-tail assumption is too restrictive for modeling the behavior of financial returns. Hence, as a second exercise we check the robustness of our sparse portfolio allocation estimators under the elliptical distributions, which we briefly review based on [56]. Elliptical distribution family generalizes the multivariate normal distribution and multivariate t-distribution. Let $\mathbf{m} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$. A p-dimensional random vector $\mathbf{r}$ has an elliptical distribution, denoted by $\mathbf{r} \sim \text{ED}_p(\mathbf{m}, \boldsymbol{\Sigma}, \zeta)$, if it has a stochastic representation

$$\mathbf{r} \overset{d}{=} \mathbf{m} + \zeta\mathbf{A}\mathbf{U}, \tag{3.42}$$

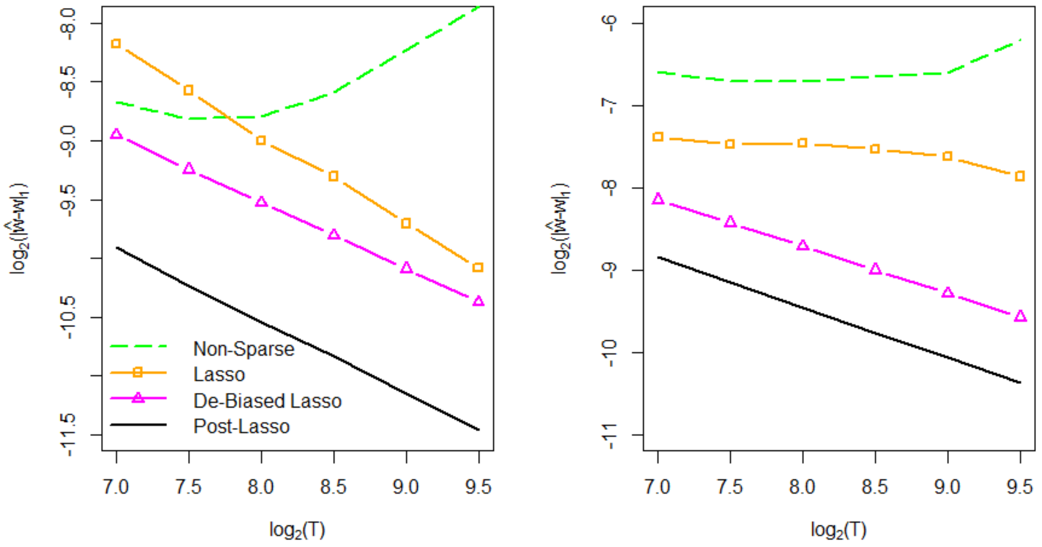

Figure 3.6.1: **Averaged errors of the estimators of $\mathbf{w}_{\mathrm{MRC}}$ for Case 1 on logarithmic scale (left):** $p = T^{0.85}$, $K = 3$ **and for Case 2 on logarithmic scale (right):** $p = 3 \cdot T^{0.85}$, $K = 3$.

where $\mathbf{U}$ is a random vector uniformly distributed on the unit sphere $\mathcal{S}^{q-1}$ in $\mathbb{R}^q$, $\zeta \geq 0$ is a scalar random variable independent of $\mathbf{U}$, $\mathbf{A} \in \mathbb{R}^{p \times q}$ is a deterministic matrix satisfying $\mathbf{A}\mathbf{A}' = \mathbf{\Sigma}$. As pointed out in [56], the representation in (3.42) is not identifiable, hence, we require $\mathbb{E}\left[\zeta^2\right] = q$, such that $\text{Cov}(\mathbf{r}) = \mathbf{\Sigma}$. We only consider continuous elliptical distributions with $\Pr[\zeta = 0] = 0$. The advantage of the elliptical distribution for the financial returns is its ability to model heavy-tailed data and the tail dependence between variables.

Having reviewed the elliptical distribution, we proceed to the second part of simulation results. The data generating process is similar to [56]: let $(\mathbf{f}_t, \boldsymbol{\varepsilon}_t)$ from (3.41) jointly follow the multivariate t-distribution with the degrees of freedom ν. When $\nu = \infty$, this corresponds to the multivariate normal distribution, smaller values of ν are associated with thicker tails. We draw T independent samples of $(\mathbf{f}_t, \boldsymbol{\varepsilon}_t)$ from the multivariate t-distribution with zero mean and covariance matrix $\mathbf{\Sigma} = \text{diag}(\mathbf{\Sigma}_f, \mathbf{\Sigma}_\varepsilon)$, where $\mathbf{\Sigma}_f = \mathbf{I}_K$. To construct $\mathbf{\Sigma}_\varepsilon$ we use a Toeplitz structure parameterized by $\rho = 0.5$, which leads to the sparse $\mathbf{\Theta}_\varepsilon = \mathbf{\Sigma}_\varepsilon^{-1}$. The rows of $\mathbf{B}$ are drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I}_K/100)$. Figure 3.6.2 reports the results for $\nu = 4.2$[5]: the performance of the standard Lasso estimator significantly deteriorates, which is further amplified in the high-dimensional case where it exhibits the worst performance. Noticeably, post-Lasso still achieves the lowest estimation error, followed by de-biased estimator.

[5] The results for larger degrees of freedom do not provide any additional insight, hence we do not report them here. However, they are available upon request.

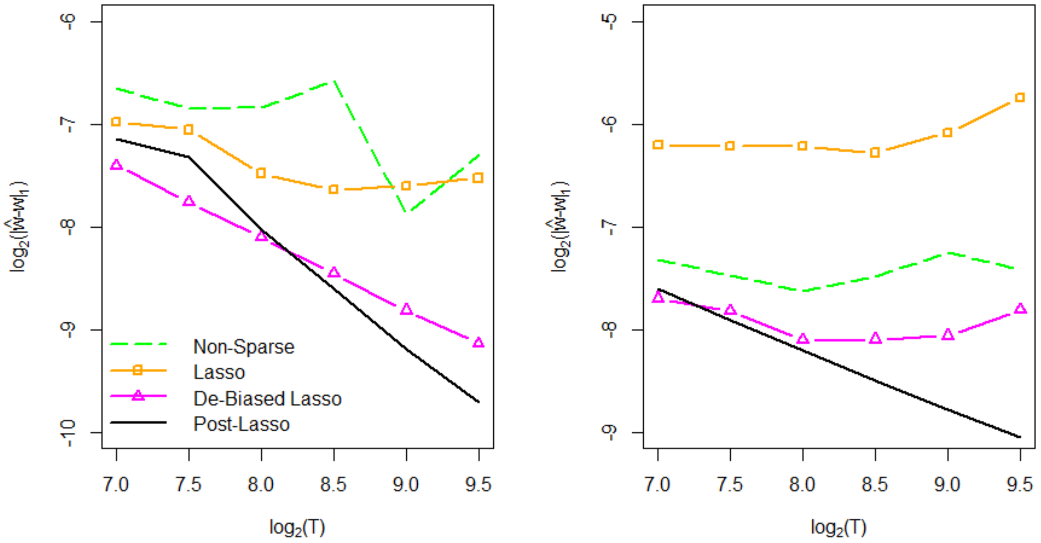

Figure 3.6.2: **Elliptical Distribution** ($\nu = 4.2$)**: Averaged errors of the estimators of $\mathbf{w}_{\mathbf{MRC}}$ for Case 1 on logarithmic scale (left):** $p = T^{0.85}$, $K = 3$ and for Case 2 on logarithmic scale (right): $p = 3 \cdot T^{0.85}$, $K = 3$.

3.7 Empirical Application

This section is divided into three main parts. First, we examine the performance of several non-sparse portfolios, including the equal-weighted and Index portfolios (reported as the composite S&P500 index listed as ^GSPC). Second, we study the performance of sparse portfolios that are based on de-biasing and post-Lasso. Third, we consider several interesting periods that include different states of the economy: we examine the merit of sparse vs non-sparse portfolios during the periods of economic growth, moderate market decline and severe economic downturns.

3.7.1 Data

We use monthly returns of the components of the S&P500 index[6]. The data on historical S&P500 constituents and stock returns is fetched from CRSP and Compustat using SAS interface. The full sample has 480 observations on 355 stocks from January 1, 1980 - December 1, 2019. We use January 1, 1980 - December 1, 1994 (180 obs) as a training period and January 1, 1995 - December 1, 2019 (300 obs) as the out-of-sample test period. We roll the estimation window over the test sample to rebalance the portfolios monthly. At the end of each month, prior to portfolio construction, we remove stocks with less than 15 years of historical stock return data. For sparse portfolio we employ the following strategy to choose the tuning parameter λ in (3.16): we use the first two thirds of the training data (which we call the training window) to estimate weights and tune the shrinkage intensity λ in the remaining one third of the training sample to yield the highest Sharpe Ratio which serves as a validation window. We estimate factors and factor loadings in the training window and validation window combined. The risk-free rate and Fama-French factors are taken from Kenneth R. French's data library.

3.7.2 Performance Measures

Similarly to [24], we consider four metrics commonly reported in finance literature: the Sharpe Ratio, the portfolio turnover, the average return and risk of a portfolio. We consider two scenarios: with and without transaction costs. Let T denote the total number of observations, the training sample consists of m observations, and the test sample is

[6] The conclusions from using daily data are the same as those for monthly returns, hence we do not report them in the main manuscript text. However, they are available upon request.

$n = T - m$. When transaction costs are not taken into account, the out-of-sample average portfolio return, risk and Sharpe Ratio are

$$\hat{\mu}_{\text{test}} = \frac{1}{n} \sum_{t=m}^{T-1} \widehat{\mathbf{w}}_t' \mathbf{r}_{t+1}, \tag{3.43}$$

$$\hat{\sigma}_{\text{test}} = \sqrt{\frac{1}{n-1} \sum_{t=m}^{T-1} (\widehat{\mathbf{w}}_t' \mathbf{r}_{t+1} - \hat{\mu}_{\text{test}})^2}, \tag{3.44}$$

$$\text{SR} = \hat{\mu}_{\text{test}} / \hat{\sigma}_{\text{test}}. \tag{3.45}$$

We follow [9,24,39,100] to account for transaction costs (tc). In line with the aforementioned papers, we set $c = 50$bps. Define the excess portfolio at time $t+1$ with transaction costs as

$$r_{t+1,\text{portfolio}} = \widehat{\mathbf{w}}_t' \mathbf{r}_{t+1} - c(1 + \widehat{\mathbf{w}}_t' \mathbf{r}_{t+1}) \sum_{j=1}^{p} \left| \hat{w}_{t+1,j} - \hat{w}_{t,j}^{+} \right|, \tag{3.46}$$

$$\text{where} \quad \hat{w}_{t,j}^{+} = \hat{w}_{t,j} \frac{1 + r_{t+1,j} + r_{t+1}^{f}}{1 + r_{t+1,\text{portfolio}} + r_{t+1}^{f}}, \tag{3.47}$$

where $r_{t+1,j} + r_{t+1}^{f}$ is sum of the excess return of the j-th asset and risk-free rate, and $r_{t+1,\text{portfolio}} + r_{t+1}^{f}$ is the sum of the excess return of the portfolio and risk-free rate. The out-of-sample average portfolio return, risk, Sharpe Ratio and turnover are defined accordingly:

$$\hat{\mu}_{\text{test,tc}} = \frac{1}{n}\sum_{t=m}^{T-1} r_{t,\text{portfolio}}, \tag{3.48}$$

$$\hat{\sigma}_{\text{test,tc}} = \sqrt{\frac{1}{n-1}\sum_{t=m}^{T-1}(r_{t,\text{portfolio}} - \hat{\mu}_{\text{test,tc}})^2}, \tag{3.49}$$

$$\text{SR}_{\text{tc}} = \hat{\mu}_{\text{test,tc}}/\hat{\sigma}_{\text{test,tc}}, \tag{3.50}$$

$$\text{Turnover} = \frac{1}{n}\sum_{t=m}^{T-1}\sum_{j=1}^{p}\left|\hat{w}_{t+1,j} - \hat{w}_{t,j}^{+}\right|. \tag{3.51}$$

3.7.3 Results and Discussion

The first set of results explores the performance of several non-sparse portfolios: equal-weighted portfolio (EW), Index portfolio (Index), MB from Algorithm 4, FMB from Algorithm 5, linear shrinkage estimator of covariance that incorporates factor structure through the Sherman-Morrison inversion formula ([91], further referred to as LW), CLIME ([23]). We consider two scenarios, when the factors are unknown and estimated using the standard PCA (statistical factors), and when the factors are known. For the statistical factors, we determine the number of factors, K, in a standard data-driven way using the information criteria discussed in [6] and [85] among others. For the scenario with known factors we include up to 5 Fama-French factors: FF1 includes the excess return on the market, FF3 includes FF1 plus size factor (Small Minus Big, SMB) and value factor (High Minus Low, HML), and FF5 includes FF3 plus profitability factor (Robust Minus Weak, RMW) and risk factor (Conservative Minus Agressive, CMA). In Table 3.7.1, we report monthly portfolio performance for three alternative portfolio allocations in (3.11), (3.12) and (3.11). We set a return target $\mu = 0.7974\%$ which is equivalent to 10% yearly return when compounded. The target level of risk for the weight-constrained and risk-constrained Markowitz portfolio (MWC and MRC) is set at $\sigma = 0.05$ which is the standard deviation of the monthly excess returns of the S&P500 index in the first training set.

	Markowitz (risk-constrained)				Markowitz (weight-constrained)				Global Minimum-Variance			
	Return	Risk	SR	Turnover	Return	Risk	SR	Turnover	Return	Risk	SR	Turnover
Without TC												
EW	0.0081	0.0520	0.1553	-	0.0081	0.0520	0.1553	-	0.0081	0.0520	0.1553	-
Index	0.0063	0.0458	0.1389	-	0.0063	0.0458	0.1389	-	0.0063	0.0458	0.1389	-
MB	0.0539	0.2522	0.2138	-	0.0070	0.0021	0.1539	-	0.0082	0.0020	0.1860	-
FMB (PC)	0.0287	0.1049	0.2743	-	0.0069	0.0346	0.1968	-	0.0076	0.0346	0.2211	-
CLIME	0.0372	0.2337	0.1593	-	0.0067	0.0471	0.1434	-	0.0076	0.0466	0.1643	-
LW	0.0296	0.1049	0.2817	-	0.0059	0.0353	0.1662	-	0.0063	0.0353	0.1774	-
FMB (FF1)	0.0497	0.2200	0.2258	-	0.0071	0.0447	0.1582	-	0.0083	0.0436	0.1921	-
FMB (FF3)	0.0384	0.1319	0.2908	-	0.0067	0.0387	0.1754	-	0.0080	0.0361	0.2223	-
FMB (FF5)	0.0373	0.1277	0.2921	-	0.0068	0.0374	0.1788	-	0.0081	0.0361	0.2250	-
With TC												
EW	0.0080	0.0520	0.1538	0.0630	0.0080	0.0027	0.1538	0.0630	0.0080	0.0027	0.1538	0.0630
MB	0.0512	0.0637	0.2027	2.9458	0.0067	0.0021	0.1461	0.3223	0.0080	0.0020	0.1804	0.2152
FMB (PC)	0.0248	0.1049	0.2368	3.7190	0.0059	0.0346	0.1687	0.9872	0.0067	0.0346	0.1929	0.9686
CLIME	0.0334	0.2334	0.1429	4.9174	0.0062	0.0471	0.1307	0.5945	0.0071	0.0466	0.1522	0.5528
LW	0.0237	0.1052	0.2257	5.5889	0.0043	0.0353	0.1231	1.5166	0.0048	0.0354	0.1343	1.5123
FMB (FF1)	0.0470	0.2202	0.2136	2.7245	0.0067	0.0447	0.1498	0.3489	0.0080	0.0436	0.1857	0.2486
FMB (FF3)	0.0356	0.1319	0.2694	2.4670	0.0062	0.0387	0.1622	0.4728	0.0076	0.0361	0.2106	0.3920
FMB (FF5)	0.0345	0.1277	0.2699	2.4853	0.0063	0.0387	0.1653	0.4847	0.0076	0.0361	0.2129	0.4057

Table 3.7.1: Monthly portfolio returns, risk, Sharpe Ratio and turnover.

We now comment on the results which are presented in Table 3.7.1: **(1)** accounting for the factor structure in stock returns improves portfolio performance in terms of the OOS Sharpe Ratio. Specifically, EW, Index, MB and CLIME which ignore factor structure perform worse than FMB and LW. **(2)** The models that use an improved estimator of covariance or precision matrix outperform EW and Index on the test sample. As a downside, such models have higher Turnover. This implies that superior performance is achieved at the cost of larger variability of portfolio positions over time and, as a consequence, increased risk associated with it.

The second set of results studies the performance of sparse portfolios: we include our proposed methods based on de-biasing and post-Lasso, as well as the approach studied in [2] (Lasso) without factor investing. For post-Lasso we first use Lasso-based weight estimator in (3.19) for selecting stocks with absolute value of weights above a small threshold ϵ (we use $\epsilon = 0.0001$), then we form portfolio with the selected stocks using three alternative portfolio allocations in (3.11)-(3.13).

	De-Biasing				Post-Lasso											
	Markowitz (RC)				Markowitz (RC)				Markowitz (WC)				GMV			
	Return	Risk	SR	Turnover	Return	Risk	SR	Turnover	Return	Risk	SR	Turnover	Return	Risk	SR	Turnover
Without TC																
Lasso (PC0)	0.0007	0.0048	0.1406	-												
De-biased Lasso (PC0)	0.0023	0.0100	0.2266	-	0.0287	0.1217	0.2362	-	-0.0174	0.4987	-0.0350	-	-0.0187	0.4941	-0.0379	-
Lasso (PC)	0.0006	0.0052	0.1122	-												
De-biased Lasso (PC)	0.0067	0.0265	0.2542	-	0.0290	0.1005	0.2882	-	0.0075	0.0624	0.1205	-	0.0087	0.0458	0.1901	-
Lasso (FF1)	0.0007	0.0039	0.1902	-												
De-biased Lasso (FF1)	0.0109	0.0346	0.3213	-	0.0207	0.1192	0.1738	-	0.0031	0.2468	0.0124	-	-0.0222	0.6047	-0.0367	-
Lasso (FF3)	0.0004	0.0040	0.1113	-												
De-biased Lasso (FF3)	0.0072	0.0265	0.2721	-	0.0157	0.1245	0.1263	-	0.0136	0.1153	0.1182	-	-0.0226	0.6054	-0.0373	-
Lasso (FF5)	0.0002	0.0042	0.0577	-												
De-biased Lasso (FF5)	0.0073	0.0300	0.2467	-	0.0212	0.1127	0.1879	-	0.0093	0.0693	0.1342	-	0.0094	0.0980	0.0959	-
With TC																
Lasso (PC0)	0.0006	0.0049	0.1189	0.0719												
De-biased Lasso (PC0)	0.0020	0.0100	0.1953	0.7952	0.0262	0.1212	0.2155	2.1249	-0.0191	0.4990	-0.0383	1.5373	-0.0199	0.4945	-0.0402	1.0737
Lasso (PC)	0.0004	0.0052	0.0845	0.1136												
De-biased Lasso (PC)	0.0055	0.0265	0.2061	1.2113	0.0268	0.1005	0.2668	2.1756	0.0059	0.0624	0.0940	1.5777	0.0076	0.0458	0.1652	1.1026
Lasso (FF1)	0.0006	0.0038	0.1654	0.0789												
De-biased Lasso (FF1)	0.0100	0.0346	0.2949	0.8298	0.0186	0.1192	0.1559	2.1589	0.0013	0.2458	0.0052	1.7374	-0.0234	0.6056	-0.0386	1.1113
Lasso (FF3)	0.0003	0.0040	0.0852	0.0785												
De-biased Lasso (FF3)	0.0062	0.0265	0.2352	0.9142	0.0134	0.1245	0.1077	2.2245	0.0120	0.1149	0.1046	1.6208	-0.0236	0.6058	-0.0390	1.0482
Lasso (FF5)	0.0001	0.0042	0.0310	0.0861												
De-biased Lasso (FF5)	0.0062	0.0300	0.2124	0.9507	0.0184	0.1122	0.1639	2.2542	0.0076	0.0693	0.1098	1.6033	0.0083	0.0980	0.0844	1.0944

Table 3.7.2: Sparse portfolio (FMB is used for de-biasing): monthly portfolio returns, risk, Sharpe Ratio and turnover.

Let us comment on the results presented in Table 3.7.2: **(1)** column one demonstrates that de-biasing leads to significant performance improvement in terms of the return and the OOS Sharpe Ratio. Even though the risk of de-biased portfolio is also higher, it still satisfies the risk-constraint. This result emphasizes the importance of correcting for the bias introduced by ℓ_1-regularization. **(2)** Comparing two bias-correction methods, de-biasing and post-Lasso, we find that the latter is characterized by higher return and higher risk. However, such increase in portfolio return is, overall, not sufficient to outperform de-biasing approach in terms of the OOS Sharpe Ratio. **(3)** Sparse portfolios have lower return, risk and turnover compared to non-sparse counterparts in Table 3.7.1, however, the OOS Sharpe Ratio is comparable. Therefore, incorporating sparsity allows investors to reduce portfolio risk at the cost of lower return while maintaining the Sharpe Ratio comparable to holding a non-sparse portfolio.

Tables 3.7.3-3.7.4 compare the performance of non-sparse and sparse (de-biased. “DL”, and post-Lasso, “PL”) portfolios for different time periods in terms of the cumulative excess return (CER) over the period of interest and risk. The first period of interest (1997-98, “Period I”) corresponds to economic growth since Index exhibited positive CER during this time. “Period II”, corresponds to moderate market decline since EW and Index had relatively small negative CER. Finally, “Period III”, corresponds to severe economic downturn and significant drop in the performance of EW and Index. The references to the specific crises in Tables 3.7.3-3.7.4 do not intend to limit these economic periods to these time spans. They merely provide the context for the time intervals of interest. Since the performance of MWC portfolios is similar to GMV, we only report MRC and GMV.

	Asian & Rus. Fin. Crisis (1997-1998)		Argen. Great Depr. & dot-com bubble (1999-2002)		Fin. Crisis (2007-2009)	
	CER	**Risk**	**CER**	**Risk**	**CER**	**Risk**
EW	0.2712	0.0547	-0.0322	0.0519	-0.4987	0.1203
Index	0.3222	0.0508	-0.1698	0.0539	-0.4924	0.0929
			Markowitz Risk-Constrained (MRC)			
MB	2.1662	0.3381	-0.1140	0.2916	-3.0688	0.5101
CLIME	1.3285	0.0892	0.4241	0.1297	-3.0470	0.4735
LW	0.9134	0.1021	0.3677	0.1412	-0.3196	0.2751
FMB (PC)	1.3153	0.0883	0.5016	0.1286	-0.1312	0.1219
FMB (FF1)	2.0379	0.3029	0.0861	0.2660	-2.7247	0.4301
			Global Minimum-Variance Portfolio (GMV)			
MB	0.2791	0.0496	-0.0470	0.0476	-0.4637	0.1015
CLIME	0.3960	0.0374	-0.1224	0.0510	-0.4588	0.0987
LW	0.3127	0.0415	-0.0952	0.0483	-0.4013	0.0693
FMB (PC)	0.4117	0.0364	-0.1227	0.0505	-0.3444	0.0393
FMB (FF1)	0.2784	0.0487	-0.0396	0.0468	-0.4570	0.0986

Table 3.7.3: Cumulative excess return (CER) and risk of non-sparse portfolios using monthly data.

	Asian & Rus. Fin. Crisis (1997-1998)		Argen. Great Depr. & dot-com bubble (1999-2002)		Fin. Crisis (2007-2009)	
	CER	**Risk**	**CER**	**Risk**	**CER**	**Risk**
EW	0.2712	0.0547	-0.0322	0.0519	-0.4987	0.1203
Index	0.3222	0.0508	-0.1698	0.0539	-0.4924	0.0929
			Debiased MRC			
DL(PC)	0.2962	0.0261	0.1567	0.0217	0.1129	0.0408
DL(FF1)	0.4149	0.0277	0.1681	0.0240	-0.0258	0.0230
DL(FF3)	0.2123	0.0142	0.1782	0.0186	-0.0406	0.0202
			Post-Lasso MRC			
PL(PC)	3.0881	0.2211	1.7153	0.1281	2.6131	0.1862
PL(FF1)	2.3433	0.1568	1.4470	0.1828	2.8639	0.2404
PL(FF3)	0.6691	0.1887	-0.1561	0.1799	-0.9998	0.1410
			Post-Lasso GMV			
PL(PC)	0.4403	0.0593	0.8150	0.0955	-0.3694	0.1243
PL(FF1)	0.3385	0.0616	0.8151	0.0877	-0.5545	0.1213
PL(FF3)	0.0711	0.0713	0.1458	0.1061	0.0295	0.0694

Table 3.7.4: Cumulative excess return (CER) and risk of sparse portfolios using monthly data.

Let us summarize the findings from Tables 3.7.3-3.7.4: **(1)** In Period I non-sparse portfolios that rely on the estimation of covariance or precision matrix outperformed EW and Index in terms of CER for both MRC and GMV. However, in Period II GMV portfolios exhibited slightly negative CER, whereas MRC portfolios had higher risk but positive CER (albeit being lower compared to Period I). Note that in Period III none of the non-sparse portfolios generated positive CER and portfolio risk increased rapidly. Examining the performance of sparse portfolios in Table 3.7.4, we see that **(2)** our proposed sparse portfolios produce positive CER during all three periods of interest. Furthermore, the return generated by PL is higher than that by non-sparse portfolios even during Periods I and II. Interestingly, DL produces positive CER without having high risk exposure. This suggests that our de-biased estimator of portfolio weights exhibits minimax properties. We leave the formal theoretical treatment of the latter for the future research.

3.8 Conclusion

This paper develops an approach to construct sparse portfolios in high dimensions that addresses the shortcomings of the existing sparse portfolio allocation techniques. We establish the oracle bounds of sparse weight estimators and provide guidance regarding their distribution. From the empirical perspective, we examine the merit of sparse portfolios during different market scenarios. We find that in contrast to non-sparse counterparts, our strategy is robust to recessions and can be used as a hedging vehicle during such times. Our framework makes use of the tool from the network theory called nodewise regression which not only satisfies desirable statistical properties, but also allows us to study whether certain industries could serve as safe havens during recessions. We find that such non-cyclical industries as consumer staples, healthcare, retail and food were driving the returns of the sparse portfolios during both the global financial crisis of 2007-09 and COVID-19 outbreak, whereas insurance sector was the least attractive investment in both periods. Finally, we develop a simple framework that provides clear guidelines how to implement factor investing using the methodology developed in this paper.

Appendices

In this Appendix we collected the proofs of Theorems 2-5.

3.A Proof of Theorem 7

The first part of Theorem 7 was proved in [56] (see their proof of Theorem 2.1) under the assumptions **(A.1)-(A.3)**, **(B.1)-(B.3)** and $\log p = o(T)$. To prove the convergence rates for the precision matrix of the factor-adjusted returns, we follow [30], [27] and [24]. Using the facts that $\max_{i \le p}(1/T)\sum_{t=1}^{T}|\hat{\varepsilon}_{it} - \varepsilon_{it}| = \mathcal{O}_P(\omega_T^2)$ and $\max_{i,t}|\hat{\varepsilon}_{it} - \varepsilon_{it}| = \mathcal{O}_P(\omega_T) = o_P(1)$, we get

$$\max_{1 \le j \le p} \left\| \widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j \right\|_1 = \mathcal{O}_P(\bar{s}\omega_T), \tag{3.52}$$

where $\widehat{\boldsymbol{\gamma}}_j$ was defined in (3.25). The proof of 3.52 is similar to the proof of the equation (23) of [30], with $\omega_T = \sqrt{\log p/T}$ for their case.

Similarly to [24], consider the following linear model:

$$\widehat{\varepsilon}_j = \widehat{\mathbf{E}}_{-j}\boldsymbol{\gamma}_j + \boldsymbol{\eta}_j, \text{ for } j = 1, \ldots, p, \tag{3.53}$$
$$\mathbb{E}\left[\boldsymbol{\eta}_j'\widehat{\mathbf{E}}_{-j}\right] = 0.$$

[135] and [30] showed that

$$\max_{1\leq j\leq p}\left\|\boldsymbol{\eta}_j'\widehat{\mathbf{E}}_{-j}\right\|_{\infty}/T = \mathcal{O}_P(\omega_T). \tag{3.54}$$

Let $\tau_j^2 \equiv \mathbb{E}\left[\boldsymbol{\eta}_j'\boldsymbol{\eta}_j\right]$, then we have

$$\max_{1\leq j\leq p}\left\|\boldsymbol{\eta}_j'\boldsymbol{\eta}_j/T - \tau_j^2\right\| = \mathcal{O}_P(\omega_T). \tag{3.55}$$

Note that the rate in (3.55) is the same as in Lemma 1 of [30] with $\omega_T = \sqrt{\log p/T}$ for their case. However, the rate in (3.55) is different from the one derived in [135] since we allow time-dependence between factor-adjusted returns.

Recall that $\hat{\tau}_j^2 = \left\|\widehat{\varepsilon}_j - \widehat{\mathbf{E}}_{-j}\widehat{\boldsymbol{\gamma}}_j\right\|_2^2/T + \lambda_j\|\widehat{\boldsymbol{\gamma}}_j\|_1$. Using triangle inequality, we have:

$$\max_{1\leq j\leq p}\left|\hat{\tau}_j^2 - \tau_j^2\right| \leq \underbrace{\max_{1\leq j\leq p}\left|\boldsymbol{\eta}_j'\boldsymbol{\eta}_j/T - \tau_j^2\right|}_{\text{I}} + \underbrace{\max_{1\leq j\leq p}\left|\boldsymbol{\eta}_j'\widehat{\mathbf{E}}_{-j}(\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j)/T\right|}_{\text{II}}$$
$$+ \underbrace{\max_{1\leq j\leq p}\left|\boldsymbol{\eta}_j'\widehat{\mathbf{E}}_{-j}\boldsymbol{\gamma}_j/T\right|}_{\text{III}} + \underbrace{\max_{1\leq j\leq p}\boldsymbol{\gamma}_j'\widehat{\mathbf{E}}_{-j}'\widehat{\mathbf{E}}_{-j}(\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j)/T}_{\text{IV}}.$$

The first term was bounded in 3.55, we now bound the remaining terms:

$$\mathrm{II} \leq \max_{1\leq j\leq p}\left\|\boldsymbol{\eta}_j'\widehat{\mathbf{E}}_{-j}/T\right\|_\infty \max_{1\leq j\leq p}\|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\|_1 = \mathcal{O}_P(\bar{s}\omega_T^2),$$

where we used 3.52 and 3.54. For III we have

$$\mathrm{III} \leq \max_{1\leq j\leq p}\left\|\boldsymbol{\eta}_j'\widehat{\mathbf{E}}_{-j}/T\right\|_\infty \max_{1\leq j\leq p}\|\boldsymbol{\gamma}_j\|_1 = \mathcal{O}_P(\sqrt{\bar{s}}\omega_T),$$

where we used 3.54 and the fact that $\|\boldsymbol{\gamma}_j\|_1 \leq \sqrt{s_j}\|\boldsymbol{\gamma}_j\|_2 = \mathcal{O}(\sqrt{s_j})$. To bound the last term, we use KKT conditions in node-wise regression:

$$\max_{1\leq j\leq p}\left\|\widehat{\mathbf{E}}_{-j}'\widehat{\mathbf{E}}_{-j}(\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j)/T\right\|_\infty \leq \max_{1\leq j\leq p}\left\|\widehat{\mathbf{E}}_{-j}'\boldsymbol{\eta}_j/T\right\|_\infty + \max_{1\leq j\leq p}\lambda_j = \mathcal{O}_P(\omega_T),$$

where we used 3.54 and $\lambda_j \asymp \omega_T$. It follows that

$$\mathrm{IV} = \mathcal{O}_P(\omega_T)\max_{1\leq j\leq p}\|\boldsymbol{\gamma}_j\|_1 = \mathcal{O}_P(\sqrt{\bar{s}}\omega_T).$$

Therefore, we now have shown that

$$\max_{1\leq j\leq p}\left|\hat{\tau}_j^2 - \tau_j^2\right| = \mathcal{O}_P(\sqrt{\bar{s}}\omega_T). \tag{3.56}$$

Using the fact that $1/\tau_j^2 = \mathcal{O}(1)$, we also have

$$1/\hat{\tau}_j^2 - 1/\tau_j^2 = \mathcal{O}_P(\sqrt{\bar{s}}\omega_T). \tag{3.57}$$

Finally, using the analysis in (B.51)-(B.53) of [27], we get

$$\max_{1\leq j\leq p}\left\|\widehat{\boldsymbol{\Theta}}_{\varepsilon,j}-\boldsymbol{\Theta}_{\varepsilon,j}\right\|_1=\mathcal{O}_P(s_T\omega_T). \tag{3.58}$$

To prove the second rate for the precision of the factor-adjusted returns, we note that

$$\max_{1\leq j\leq p}\|\widehat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\|_2=\mathcal{O}_P(\sqrt{\bar{s}}\omega_T), \tag{3.59}$$

which was obtained in [30] (see their Lemma 2). We can write

$$\max_{1\leq j\leq p}\left\|\widehat{\boldsymbol{\Theta}}_{\varepsilon,j}-\boldsymbol{\Theta}_{\varepsilon,j}\right\|_2\leq\max_{1\leq j\leq p}[\|\widehat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\|_2/\hat{\tau}_j^2+\|\boldsymbol{\gamma}_j\|_2 1/\hat{\tau}_j^2-1/\tau_j^2]=\mathcal{O}_P(\sqrt{\bar{s}}\omega_T). \tag{3.60}$$

3.B Proof of Theorem 8

Let $\widehat{\mathbf{J}}=\widehat{\boldsymbol{\Lambda}}^{1/2}\widehat{\boldsymbol{\Gamma}}'\widehat{\boldsymbol{\Theta}}_\varepsilon\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Lambda}}^{1/2}$ and $\widetilde{\mathbf{J}}=\widetilde{\boldsymbol{\Lambda}}^{1/2}\widetilde{\boldsymbol{\Gamma}}'\boldsymbol{\Theta}_\varepsilon\widetilde{\boldsymbol{\Gamma}}\widetilde{\boldsymbol{\Lambda}}^{1/2}$. Also, define

$$\Delta_{\text{inv}}=\widehat{\boldsymbol{\Theta}}_\varepsilon\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Lambda}}^{1/2}(\mathbf{I}_K+\widehat{\mathbf{J}})^{-1}\widehat{\boldsymbol{\Lambda}}^{1/2}\widehat{\boldsymbol{\Gamma}}'\widehat{\boldsymbol{\Theta}}_\varepsilon-\boldsymbol{\Theta}_\varepsilon\widetilde{\boldsymbol{\Gamma}}\widetilde{\boldsymbol{\Lambda}}^{1/2}(\mathbf{I}_K+\widetilde{\mathbf{J}})^{-1}\widetilde{\boldsymbol{\Lambda}}^{1/2}\widetilde{\boldsymbol{\Gamma}}'\boldsymbol{\Theta}_\varepsilon.$$

Using Sherman-Morrison-Woodbury formulas in 2.16, we have

$$\left\|\left\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\right\|\right\|_1\leq\left\|\left\|\widehat{\boldsymbol{\Theta}}_\varepsilon-\boldsymbol{\Theta}_\varepsilon\right\|\right\|_1+\||\Delta_{\text{inv}}|\|_1. \tag{3.61}$$

As pointed out by [56], $|||\Delta_{\text{inv}}|||_1$ can be bounded by the following three terms:

$$\begin{aligned}
&\left|\left|\left|(\widehat{\mathbf{\Theta}}_\varepsilon - \mathbf{\Theta}_\varepsilon)\widetilde{\mathbf{\Gamma}}\widetilde{\mathbf{\Lambda}}^{1/2}(\mathbf{I}_K + \widetilde{\mathbf{J}})^{-1}\widetilde{\mathbf{\Lambda}}^{1/2}\widetilde{\mathbf{\Gamma}}'\mathbf{\Theta}_\varepsilon\right|\right|\right|_1 = \mathcal{O}_P(\bar{s}\omega_T \cdot p \cdot \frac{1}{p} \cdot \sqrt{\bar{s}}),\\
&\left|\left|\left|\mathbf{\Theta}_\varepsilon(\widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Lambda}}^{1/2} - \widetilde{\mathbf{\Gamma}}\widetilde{\mathbf{\Lambda}}^{1/2})(\mathbf{I}_K + \widetilde{\mathbf{J}})^{-1}\widetilde{\mathbf{\Lambda}}^{1/2}\widetilde{\mathbf{\Gamma}}'\mathbf{\Theta}_\varepsilon\right|\right|\right|_1 = \mathcal{O}_P(\sqrt{\bar{s}} \cdot p\omega_T \cdot \frac{1}{p} \cdot \sqrt{\bar{s}}),\\
&\left|\left|\left|\mathbf{\Theta}_\varepsilon\widetilde{\mathbf{\Lambda}}^{1/2}\widetilde{\mathbf{\Gamma}}'((\mathbf{I}_K + \widehat{\mathbf{J}})^{-1} - (\mathbf{I}_K + \widetilde{\mathbf{J}})^{-1})\widetilde{\mathbf{\Gamma}}'\mathbf{\Theta}_\varepsilon\right|\right|\right|_1 = \mathcal{O}_P(\sqrt{\bar{s}} \cdot \frac{1}{p} \cdot p\bar{s}\omega_T\sqrt{\bar{s}}).
\end{aligned}$$

To derive the above rates we used **(B.1)-(B.3)**, Theorem 7 and the fact that $\left\|\widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Gamma}}' - \mathbf{B}\mathbf{B}'\right\|_{\text{F}} = \mathcal{O}_P(p\omega_T)$. The second rate in Theorem 8 can be easily obtained using the technique described above for the l_2-norm.

3.C Lemmas for Theorems 9-10

Lemma 12 *Under the assumptions of Theorem 8,*

(a) $\|\widehat{\mathbf{m}} - \mathbf{m}\|_{max} = \mathcal{O}_P(\sqrt{\log p/T})$*, where* $\mathbf{m}$ *is the unconditional mean of stock returns defined in Subsection 3.2, and* $\widehat{\mathbf{m}}$ *is the sample mean.*

(b) $|||\mathbf{\Theta}|||_1 = \mathcal{O}(\bar{s})$.

Proof.

(a) The proof of Part (a) is provided in [30] (Lemma 1).

(b) To prove Part (b) we use Sherman-Morrison-Woodbury formula in 2.16:

$$\begin{aligned}
|||\mathbf{\Theta}|||_1 &\leq |||\mathbf{\Theta}_\varepsilon|||_1 + \left|\left|\left|\mathbf{\Theta}_\varepsilon\mathbf{B}[\mathbf{I}_K + \mathbf{B}'\mathbf{\Theta}_\varepsilon\mathbf{B}]^{-1}\mathbf{B}'\mathbf{\Theta}_\varepsilon\right|\right|\right|_1\\
&= \mathcal{O}(\sqrt{\bar{s}}) + \mathcal{O}(\sqrt{\bar{s}} \cdot p \cdot \frac{1}{p} \cdot \sqrt{\bar{s}}) = \mathcal{O}(\bar{s}). \qquad (3.62)
\end{aligned}$$

The last equality in (3.62) is obtained under the assumptions of Theorem 10. This result is important in several aspects: it shows that the sparsity of the precision matrix of stock returns is controlled by the sparsity in the precision of the idiosyncratic returns. Hence, one does not need to impose an unrealistic sparsity assumption on the precision of returns a priori when the latter follow a factor structure - sparsity of the precision once the common movements have been taken into account would suffice.

■

Lemma 13 *Define* $\theta = \mathbf{m}'\mathbf{\Theta}\mathbf{m}/p$ *and* $g = \sqrt{\mathbf{m}'\mathbf{\Theta}\mathbf{m}}/p$. *Also, let* $\widehat{\theta} = \widehat{\mathbf{m}}'\widehat{\mathbf{\Theta}}\widehat{\mathbf{m}}/p$ *and* $\widehat{g} = \sqrt{\widehat{\mathbf{m}}'\widehat{\mathbf{\Theta}}\widehat{\mathbf{m}}}/p$. *Under the assumptions of Theorem 8:*

(a) $\theta = \mathcal{O}(1)$.

(b) $\left|\widehat{\theta} - \theta\right| = \mathcal{O}_P(\bar{s}^2\omega_T) = o_P(1)$.

(c) $|\widehat{y} - y| = \mathcal{O}_P(\bar{s}^2\omega_T) = o_P(1)$, *where* y *was defined in* (3.17).

(d) $|\widehat{g} - g| = \mathcal{O}_P\left([\bar{s}^2\omega_T]^{1/2}\right) = o_P(1)$.

Proof.

(a) Part (a) is trivial and follows directly from $|||\mathbf{\Theta}|||_2 = \mathcal{O}(1)$.

(b) First, rewrite the expression of interest:

$$\begin{aligned}\widehat{\theta} - \theta &= [(\widehat{\mathbf{m}} - \mathbf{m})'(\widehat{\mathbf{\Theta}} - \mathbf{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})]/p + [(\widehat{\mathbf{m}} - \mathbf{m})'\mathbf{\Theta}(\widehat{\mathbf{m}} - \mathbf{m})]/p \\ &+ [2(\widehat{\mathbf{m}} - \mathbf{m})'\mathbf{\Theta}\mathbf{m}]/p + [2\mathbf{m}'(\widehat{\mathbf{\Theta}} - \mathbf{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})]/p \\ &+ [\mathbf{m}'(\widehat{\mathbf{\Theta}} - \mathbf{\Theta})\mathbf{m}]/p. \end{aligned} \tag{3.63}$$

We now bound each of the terms in (3.63) using the expressions derived in [24] (see their Proof of Lemma A.3), Lemma 12 and the fact that $\log p/T = o(1)$.

$$\begin{aligned} \left|(\widehat{\mathbf{m}}-\mathbf{m})'(\widehat{\mathbf{\Theta}}-\mathbf{\Theta})(\widehat{\mathbf{m}}-\mathbf{m})\right|/p &\le \|\widehat{\mathbf{m}}-\mathbf{m}\|_{\max}^2 \left\|\left|\widehat{\mathbf{\Theta}}-\mathbf{\Theta}\right|\right\|_1 \\ &= \mathcal{O}_P\Big(\frac{\log p}{T} \cdot \bar{s}^2 \omega_T\Big) \end{aligned} \tag{3.64}$$

$$\left|(\widehat{\mathbf{m}}-\mathbf{m})'\mathbf{\Theta}(\widehat{\mathbf{m}}-\mathbf{m})\right|/p \le \|\widehat{\mathbf{m}}-\mathbf{m}\|_{\max}^2 |||\mathbf{\Theta}|||_1 = \mathcal{O}_P\Big(\frac{\log p}{T} \cdot \bar{s}\Big). \tag{3.65}$$

$$\left|(\widehat{\mathbf{m}}-\mathbf{m})'\mathbf{\Theta}\mathbf{m}\right|/p \le \|\widehat{\mathbf{m}}-\mathbf{m}\|_{\max} |||\mathbf{\Theta}|||_1 = \mathcal{O}_P\Big(\sqrt{\frac{\log p}{T}} \cdot \bar{s}\Big). \tag{3.66}$$

$$\begin{aligned} \left|\mathbf{m}'(\widehat{\mathbf{\Theta}}-\mathbf{\Theta})(\widehat{\mathbf{m}}-\mathbf{m})\right|/p &\le \|\widehat{\mathbf{m}}-\mathbf{m}\|_{\max} \left\|\left|\widehat{\mathbf{\Theta}}-\mathbf{\Theta}\right|\right\|_1 \\ &= \mathcal{O}_P\Big(\sqrt{\frac{\log p}{T}} \cdot \bar{s}^2 \omega_T\Big). \end{aligned} \tag{3.67}$$

$$\left|\mathbf{m}'(\widehat{\mathbf{\Theta}}-\mathbf{\Theta})\mathbf{m}\right|/p \le \left\|\left|\widehat{\mathbf{\Theta}}-\mathbf{\Theta}\right|\right\|_1 = \mathcal{O}_P\Big(\bar{s}^2 \omega_T\Big). \tag{3.68}$$

(c) Part (c) trivially follows from Part (b).

(d) This is a direct consequence of Part (b) and the fact that $\sqrt{\widehat{\theta}-\theta} \ge \sqrt{\widehat{\theta}} - \sqrt{\theta}$.

■

3.D Proof of Theorem 9

Using the definition of MRC weight in (3.13), we can rewrite

$$
\begin{aligned}
&\|\widehat{\mathbf{w}}_{\text{MRC}} - \mathbf{w}_{\text{MRC}}\|_1 \leq \frac{\frac{g}{p}\left[\left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})\right\|_1 + \left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\mathbf{m}\right\|_1 + \|\boldsymbol{\Theta}(\widehat{\mathbf{m}} - \mathbf{m})\|_1\right]}{|\widehat{g}|g} \\
&+ \frac{|\widehat{g} - g|\|\boldsymbol{\Theta}\mathbf{m}\|_1}{|\widehat{g}|g} \\
&\leq \frac{\frac{g}{p}\left[p\left\|\left|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right|\right\|_1\|(\widehat{\mathbf{m}} - \mathbf{m})\|_{\max} + p\left\|\left|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right|\right\|_1\|\mathbf{m}\|_{\max} + p|\!|\!|\boldsymbol{\Theta}|\!|\!|_1\|(\widehat{\mathbf{m}} - \mathbf{m})\|_{\max}\right]}{|\widehat{g}|g} \\
&+ \frac{p|\widehat{g} - g||\!|\!|\boldsymbol{\Theta}|\!|\!|_1\|\mathbf{m}\|_{\max}}{|\widehat{g}|g} \\
&= \mathcal{O}_P\left(\bar{s}^2\omega_T \cdot \sqrt{\frac{\log p}{T}}\right) + \mathcal{O}_P\left(\bar{s}^2\omega_T\right) + \mathcal{O}_P\left(\bar{s} \cdot \sqrt{\frac{\log p}{T}}\right) + \mathcal{O}_P\left([\bar{s}^2\omega_T]^{1/2} \cdot \bar{s}\right) = o_P(1),
\end{aligned}
$$

where we used Lemmas 1-2 to obtain the rates.

3.E Proof of Theorem 10

The KKT conditions for the nodewise Lasso in (3.25) imply that

$$
\hat{\tau}_j^2 = (\widehat{\boldsymbol{\varepsilon}}_j - \widehat{\mathbf{E}}_{-j}\widehat{\boldsymbol{\gamma}}_j)'\widehat{\boldsymbol{\varepsilon}}_j/T, \text{ hence, } \quad \widehat{\boldsymbol{\varepsilon}}_j'\widehat{\mathbf{E}}\widehat{\boldsymbol{\Theta}}_{\varepsilon,j}'/T = 1.
$$

As shown in [135], these KKT conditions also imply that

$$
\left\|\widehat{\mathbf{E}}_{-j}'\widehat{\mathbf{E}}\widehat{\boldsymbol{\Theta}}_{\varepsilon,j}\right\|_\infty/T \leq \lambda_j/\hat{\tau}_j^2. \tag{3.69}
$$

Therefore, the estimator of precision matrix needs to satisfy the following “extended KKT” condition:

$$\left\|\widehat{\mathbf{\Sigma}}_{\varepsilon}\widehat{\mathbf{\Theta}}'_{\varepsilon,j} - \mathbf{e}_j\right\|_{\infty} \leq \lambda_j/\hat{\tau}_j^2, \tag{3.70}$$

where $\mathbf{e}_j$ is the j-th unit column vector. Combining the rate in ℓ_1 norm in Theorem 8 and (3.70), we have:

$$\left\|\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Theta}}'_{j} - \mathbf{e}_j\right\|_{\infty} \leq \lambda_j/\hat{\tau}_j^2, \tag{3.71}$$

Using the definition of Δ in (3.23), it is straightforward to see that

$$\|\Delta\|_{\infty}/\sqrt{T} = \left\|(\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}} - \mathbf{I}_p)(\widehat{\mathbf{w}} - \mathbf{w})\right\|_{\infty} \leq \left\|\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}} - \mathbf{I}_p\right\|_{\infty}\|\widehat{\mathbf{w}} - \mathbf{w}\|_1. \tag{3.72}$$

Therefore, combining (3.71) and (3.72), we have

$$\|\Delta\|_{\infty} \leq \sqrt{T}\|\widehat{\mathbf{w}} - \mathbf{w}\|_1 \max_j \lambda_j/\hat{\tau}_j^2 = \mathcal{O}_P\Big(\sqrt{T}\cdot(s_0 \vee \bar{s}^2)\omega_T \cdot \omega_T\Big) \tag{3.73}$$

$$= \mathcal{O}_P\Big((s_0 \vee \bar{s}^2)\Big(\log p/\sqrt{T} + \sqrt{T}/p\Big)\Big) = o_P(1). \tag{3.74}$$

Finally, we show that $\left\|\widehat{\mathbf{\Omega}} - \mathbf{\Theta}\right\|_{\infty} = o_P(1)$. Using Theorem 8 and Lemma 12 we have $\left\|\widehat{\mathbf{\Theta}}_j\right\|_1 = \mathcal{O}_P(s_j)$. Also,

$$\widehat{\mathbf{\Omega}} = \widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Theta}}' = (\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}} - \mathbf{I}_p)\widehat{\mathbf{\Theta}}' + \widehat{\mathbf{\Theta}}'. \tag{3.75}$$

And using 3.71 and 3.72 together with $\max_j \lambda_j s_j^2 = o_P(1)$:

$$\left\|(\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}} - \mathbf{I}_p)\widehat{\mathbf{\Theta}}'\right\|_\infty \leq \max_j \lambda_j \left\|\widehat{\mathbf{\Theta}}_j\right\|_1 / \hat{\tau}_j^2 = o_P(1). \tag{3.76}$$

It follows that

$$\left\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\right\|_\infty \leq \max_j \left\|\widehat{\mathbf{\Theta}}_j - \mathbf{\Theta}_j\right\|_2 \leq \max_j \lambda_j \sqrt{s_j} = o_P(1). \tag{3.77}$$

Combining 3.75-3.77 completes the proof.

Chapter 4

Learning from Forecast Errors: A New Approach to Forecast Combinations

Abstract

[1] Forecasters often use common information and hence make common mistakes. We propose a new approach, Factor Graphical Model (FGM), to forecast combinations that separates idiosyncratic forecast errors from the common errors. FGM exploits the factor structure of forecast errors and the sparsity of the precision matrix of

[1] This paper is co-authored with Dr. Tae-Hwy Lee and is circulated under the name "Learning from Forecast Errors: A New Approach to Forecast Combinations".

the idiosyncratic errors. We prove the consistency of forecast combination weights and mean squared forecast error estimated using FGM, supporting the results with extensive simulations. Empirical applications to forecasting macroeconomic series shows that forecast combination using FGM outperforms combined forecasts using equal weights and graphical models without incorporating factor structure of forecast errors.

4.1 Introduction

A search for the best forecast combination has been an important on-going research question in economics. [33] pointed out that combining forecasts is "practical, economical and useful. Many empirical tests have demonstrated the value of composite forecasting. We no longer need to justify that methodology". However, as demonstrated by [42], there are still some unresolved issues. Despite the findings based on the theoretical grounds, equal-weighted forecasts have proved surprisingly difficult to beat. Many methodologies that seek for the best forecast combination use equal weights as a benchmark: for instance, [42] develop "partially egalitarian Lasso".

The success of equal weights is partly due to the fact that the forecasters use the same set of public information to make forecasts, hence, they tend to make common mistakes. For example, in the European Central Bank's Survey of Professional forecasters of Euro-area real GDP growth, the forecasters tend to *jointly* understate or overstate GDP growth. Therefore, we stipulate that the forecast errors include common and idiosyncratic components, which allows the forecast errors to move together due to the common error component. Our paper provides a simple framework to learn from analyzing forecast errors: we separate unique errors from the common errors to improve the accuracy of the combined forecast and support the merits of such approach using an empirical application to a large dataset of macroeconomic time series.

Dating back to [12], the well-known expression for the optimal forecast combination weights requires an estimator of inverse covariance (precision) matrix. Graphical models are a powerful tool to estimate precision matrix directly, avoiding the step of obtaining an estimator of covariance matrix to be inverted. Prominent examples of graphical models include Graphical Lasso ([65]) and nodewise regression ([108]). Despite using different strategies for estimating precision matrix, all graphical models assume that the latter is sparse: many entries of precision matrix are zero, which is a necessary condition to consistently estimate inverse covariance. Our paper demonstrates that such assumption contradicts the stylized fact that experts tend to make common mistakes and hence the forecast errors move together through common factors. We show that graphical models fail to recover entries of precision matrix under the factor structure.

This paper overcomes the aforementioned challenge and develops a new precision matrix estimator for the forecast errors under the approximate factor model with unobserved factors. We call our algorithm the *Factor Graphical Model*. We use a factor model to estimate an idiosyncratic component of the forecast errors, and then apply a Graphical model (Graphical Lasso or nodewise regression) for the estimation of the precision matrix of the idiosyncratic component.

There are a few papers that used graphical models in different contexts to estimate the covariance matrix of the idiosyncratic component when the factors are known and the loadings are assumed to be constant. [18] estimate a sparse covariance matrix for high-frequency data and construct the realized network for financial data. [11] develop a power-law partial correlation network based on the Gaussian graphical models. [87] uses the Weighted Graphical Lasso to estimate a sparse covariance matrix of the idiosyncratic component for a factor model with observable factors for high-frequency financial data.

Our paper makes several contributions. First, we allow the forecast errors to be highly correlated due to the common component which is motivated by the stylized fact that the forecasters tend to jointly understate or overstate the predicted series of interest. Second, we develop a high-dimensional precision matrix estimator which combines the benefits of the *factor* structure and *sparsity* of the precision matrix of the idiosyncratic component for the forecast combination under the approximate factor model. We prove consistency of forecast combination weights and the Mean Squared Forecast Error (MSFE) estimated using Factor Graphical models. Third, an empirical application to forecasting macroeconomic series in big data environment shows that incorporating the factor structure of the forecast errors into the graphical models improves the performance of a combined forecast over forecast combination using equal weights and graphical models without factors.

The paper is structured as follows: Section 2 reviews Graphical Lasso and nodewise regression. Section 3 studies the approximate factor models for the forecast combination. Section 4 introduces the Factor Graphical Models and discusses the choice of the tuning parameters. Section 5 contains theoretical results and Section 6 validates these results using simulations. Section 7 studies an empirical application for macroeconomic time-series. Section 8 concludes and Section 9 collects the proofs of the theorems.

Notation. For the convenience of the reader, we summarize the notation to be used throughout the paper. Let $\mathcal{S}_p$ denote the set of all $p \times p$ symmetric matrices. For any matrix $\mathbf{C}$, its (i, j)-th element is denoted as c_{ij}. Given a vector $\mathbf{u} \in \mathbb{R}^d$ and a parameter $a \in [1, \infty)$, let $\|\mathbf{u}\|_a$ denote ℓ_a-norm. Given a matrix $\mathbf{U} \in \mathcal{S}_p$, let $\Lambda_{\max}(\mathbf{U}) \equiv \Lambda_1(\mathbf{U}) \geq \Lambda_2(\mathbf{U}) \geq \ldots \geq \Lambda_{\min}(\mathbf{U}) \equiv \Lambda_p(\mathbf{U})$ be the eigenvalues of $\mathbf{U}$. Given a matrix $\mathbf{U} \in \mathbb{R}^{p \times p}$ and parameters $a, b \in [1, \infty)$, let $|||\mathbf{U}|||_{a,b} \equiv \max_{\|\mathbf{y}\|_a = 1} \|\mathbf{U}\mathbf{y}\|_b$ denote the induced matrix-operator norm. The special cases are $|||\mathbf{U}|||_1 \equiv \max_{1 \leq j \leq p} \sum_{i=1}^{p} |u_{i,j}|$ for the ℓ_1/ℓ_1-operator norm; the operator norm (ℓ_2-matrix norm) $|||\mathbf{U}|||_2^2 \equiv \Lambda_{\max}(\mathbf{U}\mathbf{U}')$ is equal to the maximal singular value of $\mathbf{U}$. Finally, $\|\mathbf{U}\|_\infty \equiv \max_{i,j} |u_{i,j}|$ denotes the element-wise maximum.

4.2 Graphical Models for Forecast Errors

This section briefly reviews a class of models, called graphical models, that search for the estimator of the precision matrix. In graphical models, each vertex represents a random variable, and the graph visualizes the joint distribution of the entire set of random variables. *Sparse graphs* have a relatively small number of edges.

Suppose we have p competing forecasts of the univariate series y_t, $t = 1, \ldots, T$. Let $\mathbf{e}_t = (e_{1t}, \ldots, e_{pt})' \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ be a $p \times 1$ vector of forecast errors. Assume they follow a Gaussian distribution. The precision matrix $\mathbf{\Sigma}^{-1} \equiv \mathbf{\Theta}$ contains information about partial covariances between the variables. For instance, if θ_{ij}, which is the ij-th element of the precision matrix, is zero, then the variables i and j are conditionally independent, given the other variables.

Let $\mathbf{W}$ be the estimate of $\mathbf{\Sigma}$. Given a sample $\{\mathbf{e}_t\}_{t=1}^{T}$, let $\mathbf{S} = (1/T)\sum_{t=1}^{T}(\mathbf{e}_t)(\mathbf{e}_t)'$ denote the sample covariance matrix, which can be used as a choice for $\mathbf{W}$. Also, let $\widehat{\mathbf{D}}^2 \equiv \text{diag}(\mathbf{W})$. We can write down the Gaussian log-likelihood (up to constants) $l(\mathbf{\Theta}) = \log\det(\mathbf{\Theta}) - \text{trace}(\mathbf{W\Theta})$. When $\mathbf{W} = \mathbf{S}$, the maximum likelihood estimator of $\mathbf{\Theta}$ is $\widehat{\mathbf{\Theta}} = \mathbf{S}^{-1}$.

In the high-dimensional settings it is necessary to regularize the precision matrix, which means that some edges will be zero. In the following subsections we discuss two most widely used techniques to estimate sparse high-dimensional precision matrices.

4.2.1 Graphical Lasso

The first approach to induce sparsity in the estimation of precision matrix is to add penalty to the maximum likelihood and use the connection between the precision matrix and regression coefficients to maximize the following *weighted penalized log-likelihood* ([76]):

$$\widehat{\mathbf{\Theta}}_\lambda = \arg\min_{\mathbf{\Theta}=\mathbf{\Theta}'} \text{trace}(\mathbf{W\Theta}) - \log\det(\mathbf{\Theta}) + \lambda\sum_{i\neq j}\widehat{d}_{ii}\widehat{d}_{jj}|\theta_{ij}|, \tag{4.1}$$

over positive definite symmetric matrices, where $\lambda \geq 0$ is a penalty parameter. The subscript λ in $\widehat{\mathbf{\Theta}}_\lambda$ means that the solution of the optimization problem in (4.1) will depend upon the choice of the tuning parameter. More details on the latter are provided in Subsection 4.1 that describes how to choose the shrinkage intensity in practice. In order to simplify notation, we will omit the subscript.

One of the most popular and fast algorithms to solve the optimization problem in (4.1) is called the Graphical Lasso (GLASSO), which was introduced by [65]. Define the following partitions of $\mathbf{W}$, $\mathbf{S}$ and $\boldsymbol{\Theta}$:

$$\mathbf{W} = \begin{pmatrix} \underbrace{\mathbf{W}_{11}}_{(p-1)\times(p-1)} & \underbrace{\mathbf{w}_{12}}_{(p-1)\times 1} \\ \mathbf{w}'_{12} & w_{22} \end{pmatrix}, \mathbf{S} = \begin{pmatrix} \underbrace{\mathbf{S}_{11}}_{(p-1)\times(p-1)} & \underbrace{\mathbf{s}_{12}}_{(p-1)\times 1} \\ \mathbf{s}'_{12} & s_{22} \end{pmatrix}, \boldsymbol{\Theta} = \begin{pmatrix} \underbrace{\boldsymbol{\Theta}_{11}}_{(p-1)\times(p-1)} & \underbrace{\boldsymbol{\theta}_{12}}_{(p-1)\times 1} \\ \boldsymbol{\theta}'_{12} & \theta_{22} \end{pmatrix}. \tag{4.2}$$

Let $\boldsymbol{\beta} \equiv -\boldsymbol{\theta}_{12}/\theta_{22}$. The idea of GLASSO is to set $\mathbf{W} = \mathbf{S} + \lambda\mathbf{I}$ in (4.1) and combine the gradient of (4.1) with the formula for partitioned inverses to obtain the following ℓ_1-regularized quadratic program

$$\widehat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \Big\{ \frac{1}{2}\boldsymbol{\beta}'\mathbf{W}_{11}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{s}_{12} + \lambda\|\boldsymbol{\beta}\|_1 \Big\}, \tag{4.3}$$

As shown by [65], (4.3) can be viewed as a LASSO regression, where the LASSO estimates are functions of the inner products of $\mathbf{W}_{11}$ and s_{12}. Hence, (4.1) is equivalent to p coupled LASSO problems. Once we obtain $\widehat{\boldsymbol{\beta}}$, we can estimate the entries $\boldsymbol{\Theta}$ using the formula for partitioned inverses. GLASSO procedure is summarized in Algorithm 7.

Algorithm 7: Graphical Lasso ([65])

```
Initialize $\mathbf{W} = \mathbf{S} + \lambda\mathbf{I}$. The diagonal of $\mathbf{W}$ remains the same in what follows.
Repeat for $j = 1, \ldots, p, 1, \ldots, p, \ldots$ until convergence:
    • Partition $\mathbf{W}$ into part 1: all but the $j$-th row and column, and part 2: the $j$-th row and column.
    • Solve the score equations using the cyclical coordinate descent:
          $\mathbf{W}_{11}\boldsymbol{\beta} - \mathbf{s}_{12} + \lambda \cdot \mathrm{Sign}(\boldsymbol{\beta}) = \mathbf{0}.$
      This gives a $(p-1) \times 1$ vector solution $\widehat{\boldsymbol{\beta}}$.
    • Update $\widehat{\mathbf{w}}_{12} = \mathbf{W}_{11}\widehat{\boldsymbol{\beta}}$.
In the final cycle (for $i = 1, \ldots, p$) solve for
      $\frac{1}{\widehat{\theta}_{22}} = w_{22} - \widehat{\boldsymbol{\beta}}'\widehat{\mathbf{w}}_{12}, \quad \widehat{\boldsymbol{\theta}}_{12} = -\widehat{\theta}_{22}\widehat{\boldsymbol{\beta}}.$
```

As was shown in [65], the estimator produced by Algorithm 7 is guaranteed to be positive definite. Furthermore, [76] showed that Algorithm 7 is guaranteed to converge and produces consistent estimator of precision matrix under certain sparsity conditions.

4.2.2 Nodewise Regression

An alternative approach to induce sparsity in the estimation of precision matrix in equation (4.1) is to solve for $\widehat{\mathbf{\Theta}}$ one column at a time via linear regressions, replacing population moments by their sample counterparts $\mathbf{S}$. When we repeat this procedure for each variable $j = 1, \dots, p$, we will estimate the elements of $\widehat{\mathbf{\Theta}}$ column by column using $\{\mathbf{e}_t\}_{t=1}^{T}$ via p linear regressions. [108] use this approach (which we will refer to as MB) to incorporate sparsity into the estimation of the precision matrix. Instead of running p coupled LASSO problems as in GLASSO, they fit p separate LASSO regressions using each variable (node) as the response and the others as predictors to estimate $\widehat{\mathbf{\Theta}}$. This method is known as the "nodewise" regression and it is reviewed below based on [135] and [24].

Let $\mathbf{e}_j$ be a $T \times 1$ vector of observations for the j-th regressor, the remaining covariates are collected in a $T \times p$ matrix $\mathbf{E}_{-j}$. For each $j = 1, \dots, p$ we run the following Lasso regressions:

$$\widehat{\boldsymbol{\gamma}}_j = \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^{p-1}} \Big(\|\mathbf{e}_j - \mathbf{E}_{-j}\boldsymbol{\gamma}\|_2^2 / T + 2\lambda_j \|\boldsymbol{\gamma}\|_1 \Big), \tag{4.4}$$

where $\widehat{\boldsymbol{\gamma}}_j = \{\widehat{\gamma}_{j,k}; j = 1, \dots, p, k \neq j\}$ is a $(p-1) \times 1$ vector of the estimated regression coefficients that will be used to construct the estimate of the precision matrix, $\widehat{\mathbf{\Theta}}$. Define

$$\widehat{\mathbf{C}} = \begin{pmatrix} 1 & -\widehat{\gamma}_{1,2} & \cdots & -\widehat{\gamma}_{1,p} \\ -\widehat{\gamma}_{2,1} & 1 & \cdots & -\widehat{\gamma}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\widehat{\gamma}_{p,1} & -\widehat{\gamma}_{p,2} & \cdots & 1 \end{pmatrix}. \tag{4.5}$$

For $j = 1, \ldots, p$, define

$$\hat{\tau}_j^2 = \|\mathbf{e}_j - \mathbf{E}_{-j}\widehat{\boldsymbol{\gamma}}_j\|_2^2/T + \lambda_j\|\widehat{\boldsymbol{\gamma}}_j\|_1 \tag{4.6}$$

and write

$$\widehat{\mathbf{T}}^2 = \text{diag}(\hat{\tau}_1^2, \ldots, \hat{\tau}_p^2). \tag{4.7}$$

The approximate inverse is defined as

$$\widehat{\boldsymbol{\Theta}}_{\lambda_j} = \widehat{\mathbf{T}}^{-2}\widehat{\mathbf{C}}. \tag{4.8}$$

Similarly to GLASSO, the subscript λ_j in $\widehat{\boldsymbol{\Theta}}_{\lambda_j}$ means that the estimated $\boldsymbol{\Theta}$ will depend upon the choice of the tuning parameter: more details are provided in Subsection 4.1 which discusses how to choose shrinkage intensity in practice. The subscript is omitted to simplify the notation. The procedure to estimate the precision matrix using nodewise regression is summarized in Algorithm 8.

Algorithm 8: Nodewise regression by [108] (MB)

```
1: Repeat for $j = 1, \ldots, p$ :
     • Estimate $\widehat{\boldsymbol{\gamma}}_j$ using (4.4) for a given $\lambda_j$.
     • Select $\lambda_j$ using a suitable information criterion (see section 4.1 for the possible options).
2: Calculate $\widehat{\mathbf{C}}$ and $\widehat{\mathbf{T}}^2$ .
3: Return $\widehat{\boldsymbol{\Theta}} = \widehat{\mathbf{T}}^{-2}\widehat{\mathbf{C}}$.
```

One of the caveats to keep in mind when using the MB method is that the estimator in (4.8) is not self-adjoint. [24] show (see their Lemma A.1) that $\widehat{\mathbf{\Theta}}$ in (4.8) is positive definite with high probability, however, it could still occur that $\widehat{\mathbf{\Theta}}$ is not positive definite in finite samples. In such cases we use the matrix symmetrization procedure as in [56] and then use eigenvalue cleaning as in [25] and [72].

4.3 Approximate Factor Models for Forecast Errors

The approximate factor models for the forecasts were first considered by [29]. They modeled a panel of ex-ante forecasts of a single time-series as a dynamic factor model and found out that the combined forecasts improved on individual ones when all forecasts have the same information set (up to difference in lags). This result emphasizes the benefit of forecast combination even when the individual forecasts are not based on different information and, therefore, do not broaden the information set used by any one forecaster.

In this paper, we are interested in finding the combination of forecasts which yields the best out-of-sample performance in terms of the mean-squared forecast error. We claim that the forecasters use the same set of public information to make forecasts and hence they tend to make common mistakes. Figure 4.3.1 illustrates this statement: it shows quarterly forecasts of Euro-area real GDP growth produced by the European Central Bank's Survey of Professional Forecasters from 1999Q3 to 2019Q3. As described in [42], forecasts are solicited for one year ahead of the latest available outcome: e.g., the 2007Q1 survey asked the respondents to forecast the GDP growth over 2006Q3-2007Q3. As evidenced from Figure 4.3.1, forecasters tend to jointly understate or overstate GDP growth, meaning that their forecast errors include common and idiosyncratic parts. Therefore, we can model the tendency of the forecast errors to move together via factor decomposition.

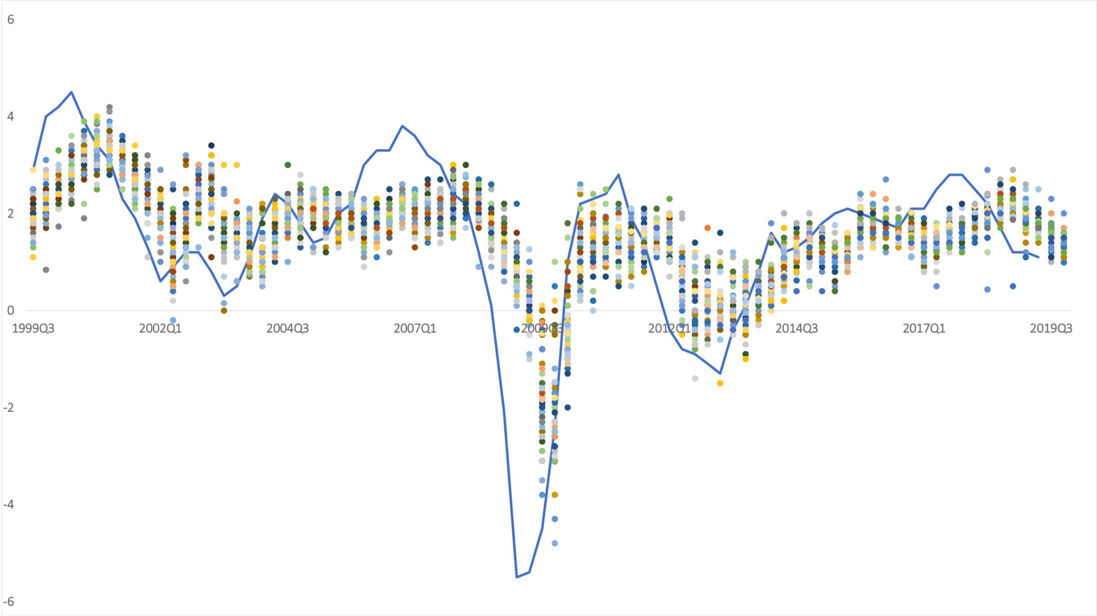

Figure 4.3.1: **The European Central Bank's (ECB) Survey of Professional Forecasters (SPF)**. Each circle denotes the forecast of each professional forecaster in the SPF for the quarterly 1-year-ahead forecasts of Euro-area real GDP growth, year-on-year percentage change. Actual series is the blue line. *Source: European Central Bank.*

Recall that we have p competing forecasts of the univariate series y_t, $t = 1, \ldots, T$ and $\mathbf{e}_t = (e_{1t}, \ldots, e_{pt})' \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ is a $p \times 1$ vector of forecast errors. Assume that the generating process for the forecast errors follows a q-factor model:

$$\underbrace{\mathbf{e}_t}_{p\times 1} = \mathbf{B} \underbrace{\mathbf{f}_t}_{q\times 1} + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T \tag{4.9}$$

where $\mathbf{f}_t = (f_{1t}, \ldots, f_{qt})'$ are the common factors of the forecast errors for p models, $\mathbf{B}$ is a $p \times q$ matrix of factor loadings, and $\boldsymbol{\varepsilon}_t$ is the idiosyncratic component that cannot be explained by the common factors. Unobservable factors, $\mathbf{f}_t$, and loadings, $\mathbf{B}$, are usually estimated by the principal component analysis (PCA), studied in [5,6,35,126]. Strict factor structure assumes that the idiosyncratic forecast error terms, $\boldsymbol{\varepsilon}_t$, are uncorrelated with each other, whereas approximate factor structure allows correlation of the idiosyncratic components ([28]).

We use the following notations: $\mathbb{E}[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'] = \boldsymbol{\Sigma}_\varepsilon$, $\mathbb{E}[\mathbf{f}_t\mathbf{f}_t'] = \boldsymbol{\Sigma}_f$, $\mathbb{E}[\mathbf{e}_t\mathbf{e}_t'] = \boldsymbol{\Sigma} = \mathbf{B}\boldsymbol{\Sigma}_f\mathbf{B}' + \boldsymbol{\Sigma}_\varepsilon$, and $\mathbb{E}[\boldsymbol{\varepsilon}_t|\mathbf{f}_t] = 0$. Let $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Theta}_\varepsilon = \boldsymbol{\Sigma}_\varepsilon^{-1}$ and $\boldsymbol{\Theta}_f = \boldsymbol{\Sigma}_f^{-1}$ be the precision matrices of forecast errors, idiosyncratic and common components respectively. The objective function to recover factors and loadings from (4.9) is:

$$\min_{\mathbf{f}_1,\ldots,\mathbf{f}_T,\mathbf{B}} \frac{1}{T}\sum_{t=1}^{T}(\mathbf{e}_t - \mathbf{B}\mathbf{f}_t)'(\mathbf{e}_t - \mathbf{B}\mathbf{f}_t) \tag{4.10}$$

$$\text{s.t. } \mathbf{B}'\mathbf{B} = \mathbf{I}_q, \tag{4.11}$$

where (4.11) is the assumption necessary for the unique identification of factors. Fixing the value of $\mathbf{B}$, we can project forecast errors $\mathbf{e}_t$ into the space spanned by $\mathbf{B}$: $\mathbf{f}_t = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{e}_t = \mathbf{B}'\mathbf{e}_t$. When combined with (4.10), this yields a concentrated objective function for $\mathbf{B}$:

$$\max_{\mathbf{B}} \ \operatorname{tr}\Big[\mathbf{B}'\Big(\frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_t\mathbf{e}_t'\Big)\mathbf{B}\Big]. \tag{4.12}$$

It is well-known (see [126] among others) that $\widehat{\mathbf{B}}$ estimated from the first q eigenvectors of $\frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_t\mathbf{e}_t'$ is the solution to (4.12). Given a sample of the estimated residuals $\{\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{e}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t\}_{t=1}^{T}$ and the estimated factors $\{\widehat{\mathbf{f}}_t\}_{t=1}^{T}$, let $\widehat{\boldsymbol{\Sigma}}_\varepsilon = (1/T)\sum_{t=1}^{T}\widehat{\boldsymbol{\varepsilon}}_t\widehat{\boldsymbol{\varepsilon}}_t'$ and $\widehat{\boldsymbol{\Sigma}}_f = (1/T)\sum_{t=1}^{T}\widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t'$ be the sample counterparts of the covariance matrices.

Moving forward to the forecast combination exercise, suppose we have p competing forecasts, $\widehat{\mathbf{y}}_t = (\hat{y}_{1,t}, \ldots, \hat{y}_{p,t})'$, of the variable y_t, $t = 1, \ldots, T$. The forecast combination is defined as follows:

$$\widehat{y}_t^c = \mathbf{w}'\widehat{\mathbf{y}}_t \tag{4.13}$$

where $\mathbf{w}$ is a $p \times 1$ vector of weights. Define a measure of risk $\text{MSFE}(\mathbf{w}, \mathbf{\Sigma}) = \mathbf{w}'\mathbf{\Sigma}\mathbf{w}$. As shown in [12], the *optimal* forecast combination minimizes the variance of the combined forecast error:

$$\min_{\mathbf{w}} \text{MSFE} = \min_{\mathbf{w}} \mathbb{E}\left[\mathbf{w}'\mathbf{e}_t\mathbf{e}_t^{'}\mathbf{w}\right] = \min_{\mathbf{w}} \mathbf{w}'\mathbf{\Sigma}\mathbf{w}, \text{ s.t. } \mathbf{w}'\boldsymbol{\iota}_p = 1, \tag{4.14}$$

where $\boldsymbol{\iota}_p$ is a $p \times 1$ vector of ones. The solution to (4.14) yields a $p \times 1$ vector of the optimal forecast combination weights:

$$\mathbf{w} = \frac{\mathbf{\Theta}\boldsymbol{\iota}_p}{\boldsymbol{\iota}_p'\mathbf{\Theta}\boldsymbol{\iota}_p}. \tag{4.15}$$

If the true precision matrix is known, the equation (4.15) guarantees to yield the optimal forecast combination. In reality, one has to estimate $\mathbf{\Theta}$. Hence, the out-of-sample performance of the combined forecast is affected by the estimation error. As pointed out by [125], when the estimation uncertainty of the weights is taken into account, there is no guarantee that the "optimal" forecast combination will be better than the equal weights or even improve the individual forecasts. Define $a = \boldsymbol{\iota}_p' \boldsymbol{\Theta} \boldsymbol{\iota}_p / p$, and $\widehat{a} = \boldsymbol{\iota}_p' \widehat{\boldsymbol{\Theta}} \boldsymbol{\iota}_p / p$. We can write

$$\left| \frac{\text{MSFE}(\widehat{\mathbf{w}}, \widehat{\boldsymbol{\Sigma}})}{\text{MSFE}(\mathbf{w}, \boldsymbol{\Sigma})} - 1 \right| = \left| \frac{\hat{a}^{-1}}{a^{-1}} - 1 \right| = \frac{|a - \hat{a}|}{|\hat{a}|}, \tag{4.16}$$

and

$$\left\| \widehat{\mathbf{w}} - \mathbf{w} \right\|_1 \leq \frac{a \frac{\left\| (\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) \boldsymbol{\iota}_p \right\|_1}{p} + |a - \widehat{a}| \frac{\left\| \boldsymbol{\Theta} \boldsymbol{\iota}_p \right\|_1}{p}}{|\widehat{a}| a}. \tag{4.17}$$

Therefore, in order to control the estimation uncertainty in the MSFE and combination weights, one needs to obtain a consistent estimator of the precision matrix $\boldsymbol{\Theta}$. More details are discussed in Subsection 5.2 and Theorems 12 and 13.

4.4 Factor Graphical Models for Forecast Errors

Since our interest is in constructing weights for the forecast combination, our goal is to estimate a precision matrix of the forecast errors. However, as pointed out by [87], when common factors are present across the forecast errors, the precision matrix cannot be sparse because all pairs of the forecast errors are partially correlated given other forecast errors through the common factors. To illustrate this point, we generated forecast errors that follow (4.9) with $q = 2$ and $\varepsilon_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$, where $\sigma_{\varepsilon,ij} = 0.4^{|i-j|}$ is the i, j-th element of $\boldsymbol{\Sigma}_\varepsilon$. The vector of factors $\mathbf{f}_t$ is drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I}_q/10)$, and the entries of the matrix of factor loadings for forecast error $j = 1, \ldots, p$, $\mathbf{b}_j$, are drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I}_q/100)$. The full loading matrix is given by $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$. Let $\widehat{q}$ denote the number of factors estimated by the PCA. We set $(T, p) = (1000, 50)$ and plot the heatmap and histogram of population partial correlations of forecast errors $\mathbf{e}_t$, which are the entries of a precision matrix, in Figure 4.4.1. We now examine the performance of graphical models for estimating partial correlations under the factor structure. Figure 4.4.2 shows the partial correlations estimated by GLASSO that does not take into account factors: due to strict sparsity imposed by graphical models almost all partial correlations are shrunk to zero which degenerates the histogram in Figure 4.4.2. This means that strong sparsity assumption on $\mathbf{\Theta}$ imposed by classical graphical models (such as GLASSO and nodewise regression from Algorithms 7-8) is not realistic under the factor structure.

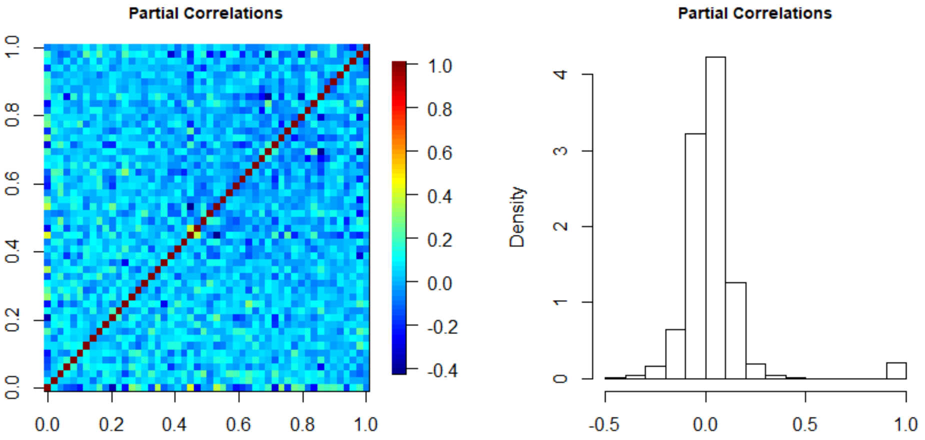

Figure 4.4.1: **Heatmap and histogram of population partial correlations.** $T = 1000$, $p = 50$, $q = 2$.

In order to avoid the aforementioned problem, instead of imposing sparsity assumption on the precision of forecast errors, $\mathbf{\Theta}$, we require sparsity of the precision matrix of the idiosyncratic errors, $\mathbf{\Theta}_\varepsilon$. The latter is obtained using the estimated residuals after removing the co-movements induced by the factors (see [11, 18, 87]). Naturally, once we

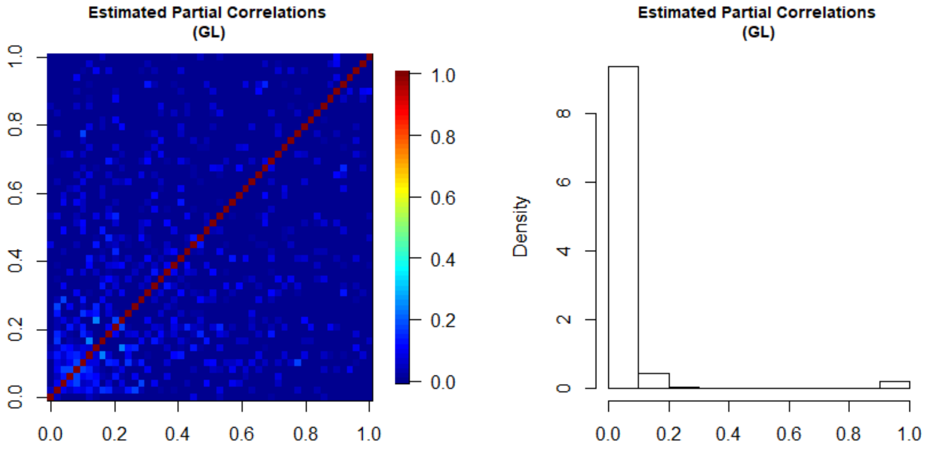

Figure 4.4.2: **Heatmap and histogram of sample partial correlations estimated using GLASSO with no factors.** $T = 1000$, $p = 50$, $q = 2$, $\hat{q} = 0$.

condition on the common components, it is sensible to assume that many remaining partial correlations of $\boldsymbol{\varepsilon}_t$ will be negligible and thus $\boldsymbol{\Theta}_\varepsilon$ is sparse.

We use the weighted Graphical Lasso and nodewise regression as shrinkage techniques to estimate the precision matrix of residuals. Once the precision of the low-rank component is obtained, we use the Sherman-Morrison-Woodbury formula to estimate the precision of forecast errors:

$$\boldsymbol{\Theta} = \boldsymbol{\Theta}_\varepsilon - \boldsymbol{\Theta}_\varepsilon \mathbf{B}[\boldsymbol{\Theta}_f + \mathbf{B}'\boldsymbol{\Theta}_\varepsilon \mathbf{B}]^{-1}\mathbf{B}'\boldsymbol{\Theta}_\varepsilon. \tag{4.18}$$

To obtain $\widehat{\boldsymbol{\Theta}}_f = \widehat{\boldsymbol{\Sigma}}_f^{-1}$, we use $\widehat{\boldsymbol{\Sigma}}_f = \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t'$. To get $\widehat{\boldsymbol{\Theta}}_\varepsilon$, we develop two approaches: the first uses the weighted GLASSO Algorithm 7, with the initial estimate of the covariance matrix of the idiosyncratic errors calculated as $\widehat{\boldsymbol{\Sigma}}_\varepsilon = \frac{1}{T}\sum_{t=1}^{T}\widehat{\boldsymbol{\varepsilon}}_t\widehat{\boldsymbol{\varepsilon}}_t'$, where $\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{e}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t$. The second uses nodewise regression and applies Algorithm 8 to $\widehat{\boldsymbol{\varepsilon}}_t$. Once we estimate $\widehat{\boldsymbol{\Theta}}_f$

and $\widehat{\mathbf{\Theta}}_{\varepsilon}$, we can get $\widehat{\mathbf{\Theta}}$ using a sample analogue of (4.18). We call the proposed procedures *Factor Graphical Lasso* and *Factor nodewise regression* and summarize them in Algorithm 9 and Algorithm 10 respectively.

```
Algorithm 9: Factor Graphical Lasso (Factor GLASSO)
1: Estimate factors, $\widehat{\mathbf{f}}_t$, and factor loadings, $\widehat{\mathbf{B}}$, using PCA. Obtain $\widehat{\mathbf{\Sigma}}_f = \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t'$, $\widehat{\mathbf{\Theta}}_f = \widehat{\mathbf{\Sigma}}_f^{-1}$, $\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{e}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t$, and $\widehat{\mathbf{\Sigma}}_{\varepsilon} = \frac{1}{T}\sum_{t=1}^{T}\widehat{\boldsymbol{\varepsilon}}_t\widehat{\boldsymbol{\varepsilon}}_t'$.
2: Estimate a sparse $\mathbf{\Theta}_{\varepsilon}$ using the weighted Graphical Lasso in (4.1) initialized with $\mathbf{W}_{\varepsilon} = \widehat{\mathbf{\Sigma}}_{\varepsilon} + \lambda\mathbf{I}$:
       $\widehat{\mathbf{\Theta}}_{\varepsilon,\lambda} = \arg\min_{\mathbf{\Theta}_{\varepsilon}=\mathbf{\Theta}_{\varepsilon}'} \operatorname{trace}(\mathbf{W}_{\varepsilon}\mathbf{\Theta}_{\varepsilon}) - \log\det(\mathbf{\Theta}_{\varepsilon}) + \lambda\sum_{i\neq j}\widehat{d}_{\varepsilon,ii}\widehat{d}_{\varepsilon,jj}|\theta_{\varepsilon,ij}|. \qquad (4.19)$
   to get $\widehat{\mathbf{\Theta}}_{\varepsilon}$.
3: Use $\widehat{\mathbf{\Theta}}_f$ from Step 1 and $\widehat{\mathbf{\Theta}}_{\varepsilon}$ from Step 2 to estimate $\mathbf{\Theta}$ using the sample counterpart of the Sherman-Morrison-Woodbury formula in (4.18):
       $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{\Theta}}_{\varepsilon} - \widehat{\mathbf{\Theta}}_{\varepsilon}\widehat{\mathbf{B}}[\widehat{\mathbf{\Theta}}_f + \widehat{\mathbf{B}}'\widehat{\mathbf{\Theta}}_{\varepsilon}\widehat{\mathbf{B}}]^{-1}\widehat{\mathbf{B}}'\widehat{\mathbf{\Theta}}_{\varepsilon}. \qquad (4.20)$
```

Algorithm 10: Factor nodewise regression [108] (Factor MB)

```
1: Estimate factors, $\widehat{\mathbf{f}}_t$, and factor loadings, $\widehat{\mathbf{B}}$, using PCA. Obtain $\widehat{\mathbf{\Sigma}}_f = \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t'$, $\widehat{\mathbf{\Theta}}_f = \widehat{\mathbf{\Sigma}}_f^{-1}$, and $\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{e}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t$.
2: Estimate a sparse $\mathbf{\Theta}_\varepsilon$ using nodewise regression: let $\widehat{\boldsymbol{\varepsilon}}_j$ be a $T \times 1$ vector of observations for the $j$-th regressor, and $\widehat{\mathbf{\Upsilon}}_{-j}$ is a $T \times p$ matrix that collects the remaining covariates. Run LASSO regressions in (4.4) for $\widehat{\boldsymbol{\varepsilon}}_t$:
       $$\widehat{\boldsymbol{\gamma}}_{\varepsilon,j} = \arg\min_{\boldsymbol{\gamma}_\varepsilon \in \mathbb{R}^{p-1}} \left( \left\| \widehat{\boldsymbol{\varepsilon}}_j - \widehat{\mathbf{\Upsilon}}_{-j}\boldsymbol{\gamma}_\varepsilon \right\|_2^2 / T + 2\lambda_j \|\boldsymbol{\gamma}_\varepsilon\|_1 \right), \qquad (4.21)$$
   to get $\widehat{\mathbf{\Theta}}_\varepsilon$.
3: Use $\widehat{\mathbf{\Theta}}_f$ from Step 1 and $\widehat{\mathbf{\Theta}}_\varepsilon$ from Step 2 to estimate $\mathbf{\Theta}$ using the sample counterpart of the Sherman-Morrison-Woodbury formula in (4.18):
       $$\widehat{\mathbf{\Theta}} = \widehat{\mathbf{\Theta}}_\varepsilon - \widehat{\mathbf{\Theta}}_\varepsilon\widehat{\mathbf{B}}[\widehat{\mathbf{\Theta}}_f + \widehat{\mathbf{B}}'\widehat{\mathbf{\Theta}}_\varepsilon\widehat{\mathbf{B}}]^{-1}\widehat{\mathbf{B}}'\widehat{\mathbf{\Theta}}_\varepsilon. \qquad (4.22)$$
```

Note that Algorithms 9 and 10 involve the tuning parameters λ and λ_j, the procedure on how to choose the shrinkage intensity coefficients is described in more detail in Subsection 4.1 that describes how to choose the shrinkage intensity in practice, and Section 5 that establishes sparsity requirements that guarantee convergence of (4.19), (4.20), (4.21), and (4.22).

We can use $\widehat{\boldsymbol{\Theta}}$ to estimate the forecast combination weights $\widehat{\mathbf{w}}$

$$\widehat{\mathbf{w}} = \frac{\widehat{\boldsymbol{\Theta}}\boldsymbol{\iota}_p}{\boldsymbol{\iota}_p'\widehat{\boldsymbol{\Theta}}\boldsymbol{\iota}_p}, \tag{4.23}$$

where $\widehat{\boldsymbol{\Theta}}$ is obtained from Algorithm 9 or Algorithm 10. Let us now revisit the motivating example at the beginning of this section: Figures 4.4.3-4.4.5 plot the heatmaps and the estimated partial correlations when precision matrix is computed using Factor GLASSO in Algorithm 2 with $\widehat{q} \in \{1, 2, 3\}$ statistical factors. The heatmaps and histograms closely resemble population counterparts in Figure 4.4.1, and the result is not very sensitive to over- or under-estimating the number of factors $\widehat{q}$. This demonstrates that using a combination of classical graphical models and factor structure via Factor Graphical Models in Algorithms 9-10 improves upon the performance of classical graphical models: our approach allows to extract the benefits of modeling common movements in forecast errors, captured by a factor model, and the benefits of using many competing forecasting models that give rise to a high-dimensional precision matrix, captured by a graphical model.

4.4.1 The Choice of the Tuning Parameters for FGM

Algorithms 9-10 require the tuning parameters λ (from Algorithm 7) and λ_j (from Algorithm 8) respectively. We now comment on the choices for both tuning parameters.

To motivate the choice of the tuning parameter for GLASSO and Factor GLASSO, we first briefly discuss some of the existing options to motivate our choice of λ in (4.1) in simulations and the empirical application. Usually λ is selected from a grid of values

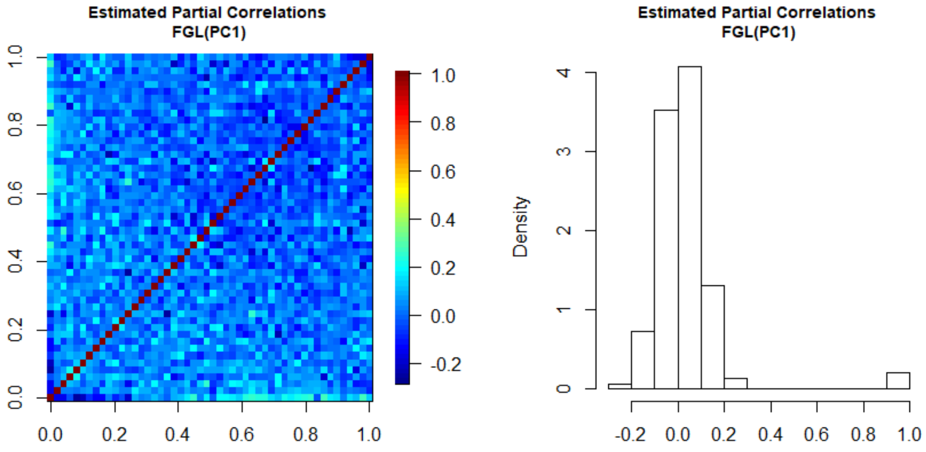

Figure 4.4.3: **Heatmap and histogram of sample partial correlations estimated using Factor GLASSO with 1 statistical factor.** $T = 1000$, $p = 50$, $q = 2$, $\hat{q} = 1$.

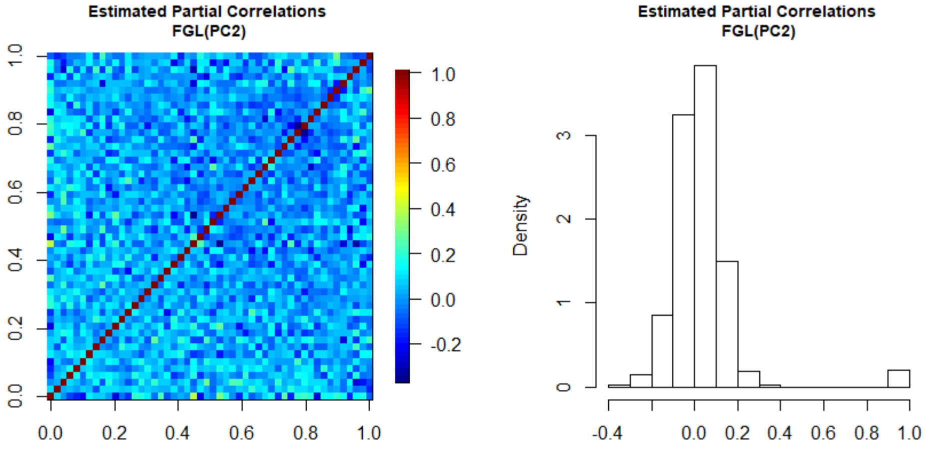

Figure 4.4.4: **Heatmap and histogram of sample partial correlations estimated using Factor GLASSO with 2 statistical factors.** $T = 1000$, $p = 50$, $q = 2$, $\hat{q} = 2$.

$F_{\lambda} = (\lambda_{\min}, \ldots, \lambda_{\max})$ which minimizes the score measuring the goodness-of-fit. Some popular examples include multifold cross-validation (CV), Stability Approach to Regularization Selection (STARS, [101]), and the Extended Bayesian Information Criteria (EBIC, [64]).

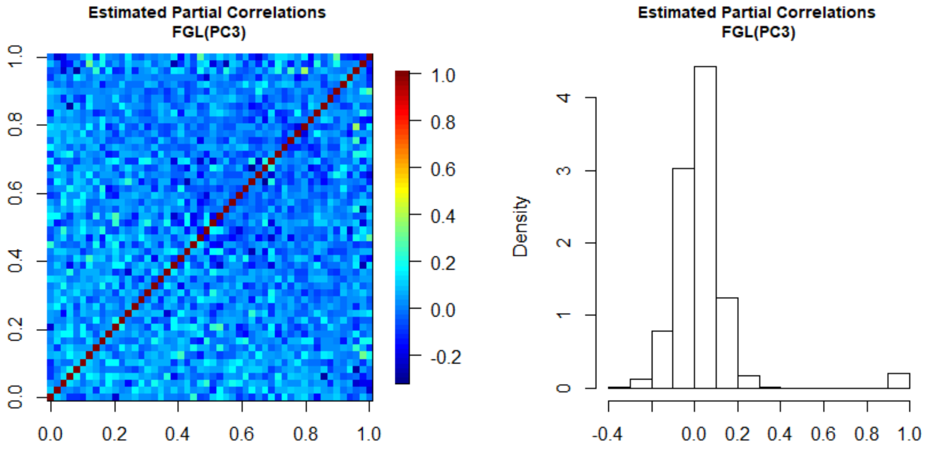

Figure 4.4.5: **Heatmap and histogram of sample partial correlations estimated using Factor GLASSO with 3 statistical factors.** $T = 1000$, $p = 50$, $q = 2$, $\hat{q} = 3$.

Since we are interested in estimating a sparse high-dimensional precision matrix, we need to choose a method for selecting the tuning parameter which is consistent in high-dimensions. [109] suggest that CV performs poorly for high-dimensional data, it overfits ([101]), and it does not consistently select models. [143] pointed out that the STARS is not computationally efficient. It is consistent under certain conditions, but suffers from the problem of overselection in estimating Gaussian graphical models. In contrast, EBIC is computationally efficient and is considered to be the state-of-the-art technique for choosing the tuning parameter for the undirected graphs. The score measuring the goodness of fit for EBIC can be written as:

$$\lambda_{\text{EBIC}} = \arg\min_{\lambda \in F_\lambda} \{-2l(\boldsymbol{\Theta}_{\varepsilon,\lambda}) + \log(T)\text{df}(\boldsymbol{\Theta}_{\varepsilon,\lambda}) + 4\text{df}(\boldsymbol{\Theta}_{\varepsilon,\lambda}) \log(p)\eta\}, \tag{4.24}$$

where $\eta \in [0, 1]$, $\boldsymbol{\Theta}_{\varepsilon,\lambda}$ is the precision matrix estimated for the tuning parameter $\lambda \in F_\lambda$, and the log-likelihood is $l(\mathbf{\Theta}_{\varepsilon,\lambda}) = \log\det(\mathbf{\Theta}_{\varepsilon,\lambda}) - \text{trace}(\mathbf{W}_\varepsilon \mathbf{\Theta}_\varepsilon)$. For the estimation of graphical models, the degrees of freedom are usually defined as the number of unique nonzero elements in the estimated precision matrix, $\text{df}(\mathbf{\Theta}_{\varepsilon,\lambda}) = \sum_{i \leq j} I_{\mathbf{\Theta}_{\varepsilon,\lambda,i,j} \neq 0}$. [31] showed that when $\eta = 1$, EBIC is consistent as long as the dimension p does not grow exponentially with the sample size T. Hence, in our simulations and the empirical exercise we use EBIC with $\eta = 1$ for GLASSO and Factor GLASSO in Algorithms 7 and 9.

For Algorithms 8 and 10, we follow [24] to choose λ_j in (4.4) by minimizing the generalized information criterion (GIC). Let $\left|\widehat{S}_j(\lambda_j)\right|$ denote the estimated number of nonzero parameters in the vector $\widehat{\boldsymbol{\gamma}}_{\varepsilon,j}$:

$$\text{GIC}(\lambda_j) = \log\left(\left\|\widehat{\boldsymbol{\varepsilon}}_j - \widehat{\mathbf{\Upsilon}}_{-j}\boldsymbol{\gamma}_\varepsilon\right\|_2^2 / T\right) + \left|\widehat{S}_j(\lambda_j)\right| \frac{\log(p)}{T} \log(\log(T)). \tag{4.25}$$

As pointed out by [24], the GIC selects the true model with probability approaching one both when $p > T$ and when $p \leq T$.

4.5 Asymptotic Properties

We first introduce some terminology and notations. Let $A \in \mathcal{S}_p$. Define the following set for $j = 1, \ldots, p$:

$$D_j(A) \equiv \{i : A_{ij} \neq 0, \ i \neq j\}, \quad d_j(A) \equiv \text{card}(D_j(A)), \quad d(A) \equiv \max_{j=1,\ldots,p} d_j(A), \tag{4.26}$$

where $d_j(A)$ is the number of edges adjacent to the vertex j (i.e., the *degree* of vertex j), and $d(A)$ measures the maximum vertex degree. Define $S(A) \equiv \bigcup_{j=1}^{p} D_j(A)$ to be the overall off-diagonal sparsity pattern, and $s(A) \equiv \sum_{j=1}^{p} d_j(A)$ is the overall number of edges contained in the graph. Note that $\mathrm{card}(S(A)) \leq s(A)$: when $s(A) = p(p-1)/2$ this would give a fully connected graph.

For the nodewise regression in (4.21), denote $D_j \equiv \{k; \gamma_{j,k} \neq 0\}$ to be the active set for row $\boldsymbol{\gamma}_j$, and let $d_j \equiv |D_j|$. Define $\bar{d} \equiv \max_{1 \leq j \leq p} d_j$.

4.5.1 Assumptions

We now list the assumptions on the model (4.9):

(A.1) (Spiked covariance model) As $p \to \infty$, $\Lambda_1(\boldsymbol{\Sigma}) > \Lambda_2(\boldsymbol{\Sigma}) + \ldots > \Lambda_q(\boldsymbol{\Sigma}) \gg \Lambda_{q+1}(\boldsymbol{\Sigma}) \geq \ldots \geq \Lambda_p(\boldsymbol{\Sigma}) > 0$, where $\Lambda_j(\boldsymbol{\Sigma}) = \mathcal{O}(p)$ for $j \leq q$, while the non-spiked eigenvalues are bounded, $\Lambda_j(\boldsymbol{\Sigma}) = o(p)$ for $j > q$. We further require that $\Lambda_1(\boldsymbol{\Sigma})$ is uniformly bounded away from infinity.

(A.2) (Pervasive factors) There exists a positive definite $q \times q$ matrix $\breve{\mathbf{B}}$ such that $\left|\left|\left| p^{-1}\mathbf{B}'\mathbf{B} - \breve{\mathbf{B}} \right|\right|\right|_2 \to 0$ and $\Lambda_{\min}(\breve{\mathbf{B}})^{-1} = \mathcal{O}(1)$ as $p \to \infty$.

We also impose strong mixing condition. Let $\mathcal{F}_{-\infty}^{0}$ and $\mathcal{F}_{T}^{\infty}$ denote the σ-algebras that are generated by $\{(\mathbf{f}_t, \boldsymbol{\varepsilon}_t) : t \leq 0\}$ and $\{(\mathbf{f}_t, \boldsymbol{\varepsilon}_t) : t \geq T\}$ respectively. Define the mixing coefficient

$$\alpha(T) = \sup_{A \in \mathcal{F}_{-\infty}^{0}, B \in \mathcal{F}_{T}^{\infty}} |\Pr A \Pr B - \Pr AB|. \tag{4.27}$$

(A.3) (Strong mixing) There exists $r_3 > 0$ such that $3r_1^{-1} + 1.5r_2^{-1} + 3r_3^{-1} > 1$, and $C > 0$ satisfying, for all $T \in \mathbb{Z}^+$, $\alpha(T) \leq \exp(-CT^{r_3})$.

Assumption **(A.1)** divides the eigenvalues into the diverging and bounded ones. This assumption is satisfied by the factor model with pervasive factors, which is stated in Assumption **(A.2)**. We say that a factor is pervasive in the sense that it has non-negligible effect on a non-vanishing proportion of individual time-series. Assumptions **(A.1)**-**(A.2)** are crucial for estimating a high-dimensional factor model: they ensure that the space spanned by the principal components in the population level $\mathbf{\Sigma}$ is close to the space spanned by the columns of the factor loading matrix $\mathbf{B}$. Assumption **(A.3)** is a technical condition which is needed to consistently estimate the factors and loadings.

Let $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^{'}$, where $\mathbf{\Sigma}$ is the covariance matrix of returns that follow factor structure described in equation (4.9). Define $\widehat{\mathbf{\Sigma}}, \widehat{\mathbf{\Lambda}}_q, \widehat{\mathbf{\Gamma}}_q$ to be the estimators of $\mathbf{\Sigma}, \mathbf{\Lambda}, \mathbf{\Gamma}$. We further let $\widehat{\mathbf{\Lambda}}_q = \text{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_q)$ and $\widehat{\mathbf{\Gamma}}_q = (\hat{v}_1, \dots, \hat{v}_q)$ to be constructed by the first q leading empirical eigenvalues and the corresponding eigenvectors of $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{B}}\widehat{\mathbf{B}}' = \widehat{\mathbf{\Gamma}}_q\widehat{\mathbf{\Lambda}}_q\widehat{\mathbf{\Gamma}}_q^{'}$. Similarly to [56], we require the following bounds on the componentwise maximums of the estimators:

(B.1) $\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_{\max} = \mathcal{O}_P(\sqrt{\log p/T})$,

(B.2) $\left\|(\widehat{\mathbf{\Lambda}}_q - \mathbf{\Lambda})\mathbf{\Lambda}^{-1}\right\|_{\max} = \mathcal{O}_P(\sqrt{\log p/T})$,

(B.3) $\left\|\widehat{\mathbf{\Gamma}}_q - \mathbf{\Gamma}\right\|_{\max} = \mathcal{O}_P(\sqrt{\log p/(Tp)})$.

Assumptions **(B.1)**-**(B.3)** are needed in order to ensure that the first q principal components are approximately the same as the columns of the factor loadings. The estimator $\widehat{\mathbf{\Sigma}}$ can be thought of as any “pilot” estimator that satisfies **(B.1)**. For sub-Gaussian distributions, sample covariance matrix, its eigenvectors and eigenvalues satisfy **(B.1)**-**(B.3)**.

In addition, the following structural assumption on the model is imposed:

(C.1) $\|\mathbf{\Sigma}\|_{\max} = \mathcal{O}(1)$ and $\|\mathbf{B}\|_{\max} = \mathcal{O}(1)$.

4.5.2 Convergence of Forecast Combination Weights and MSFE

To study the properties of the combination weights in (4.23) and MSFE, we first need to establish the convergence properties of precision matrix produced by Algorithms 9-10. Let $\omega_T \equiv \sqrt{\log p/T} + 1/\sqrt{p}$. Also, let $s(\mathbf{\Theta}_\varepsilon) = \mathcal{O}_P(s_T)$ for some sequence $s_T \in (0, \infty)$ and $d(\mathbf{\Theta}_\varepsilon) = \mathcal{O}_P(d_T)$ for some sequence $d_T \in (0, \infty)$. The deterministic sequences s_T and d_T will control the sparsity $\mathbf{\Theta}_\varepsilon$ for Factor GLASSO. Note that d_T can be smaller than or equal to s_T. The reason why we distinguish between these two sequences is to juxtapose it with the sparsity conditions for the Factor MB, where we will only use the analogue of d_T which was defined as $\bar{d}$ at the beginning of this section.

Let ϱ_{1T} be a sequence of positive-valued random variables such that $\varrho_{1T}^{-1}\omega_T \xrightarrow{p} 0$ and $\varrho_{1T} d_T s_T \xrightarrow{\text{p}} 0$, with $\lambda \asymp \omega_T$ (where λ is the tuning parameter for the Factor GLASSO in (4.19)). [97] show that under the Assumptions **(A.1)-(A.3)**, **(B.1)-(B.3)** and **(C.1)**, $\left|\!\left|\!\left|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\right|\!\right|\!\right|_1 = \mathcal{O}_P(\varrho_{1T} d_T s_T)$ for Factor GLASSO. Furthermore, let ϱ_{2T} be a sequence of positive-valued random variables such that $\varrho_{2T}^{-1}\omega_T \xrightarrow{p} 0$ and $\varrho_{2T}\bar{d}^2 \xrightarrow{\text{p}} 0$, with $\lambda_j \asymp \omega_T$ (where λ_j is the tuning parameter for Factor nodewise regression in (4.21)). [122] shows that under the Assumptions **(A.1)-(A.3)**, **(B.1)-(B.3)**, and **(C.1)**, we have $\left|\!\left|\!\left|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\right|\!\right|\!\right|_1 = \mathcal{O}_P(\varrho_{2T}\bar{d}^2)$.

It is interesting to compare the rates for precision matrix obtained by two factor graphical models: if $d_T = s_T$, the rates are similar, whereas if $d_T < s_T$ Factor MB is expected to converge faster. In fact, in high dimensions when $p > T$ and $\omega_T \simeq \sqrt{\log p/T}$, Factor MB achieves the minimax rate for this problem (see [21] for the rate expression).

Having established the convergence rates for precision matrix, we now study the properties of the combination weights and MSFE.

Theorem 12 *Assume* ***(A.1)-(A.3)***, ***(B.1)-(B.3)***, *and* ***(C.1)*** *hold.*

(i) If $\varrho_{1T}d_T^2 s_T \xrightarrow{p} 0$, *Algorithm 9 consistently estimates forecast combination weights in* (4.23): $\|\widehat{\mathbf{w}} - \mathbf{w}\|_1 = \mathcal{O}_P\Big(\varrho_{1T}d_T^2 s_T\Big) = o_P(1)$.

(ii) If $\varrho_{2T}\bar{d}^3 \xrightarrow{p} 0$, *Algorithm 10 consistently estimates forecast combination weights in* (4.23): $\|\widehat{\mathbf{w}} - \mathbf{w}\|_1 = \mathcal{O}_P\Big(\varrho_{2T}\bar{d}^3\Big) = o_P(1)$.

Theorem 13 *Assume* ***(A.1)-(A.3)***, ***(B.1)-(B.3)***, *and* ***(C.1)*** *hold.*

(i) If $\varrho_{1T}d_T s_T \xrightarrow{p} 0$, *Algorithm 9 consistently estimates* $MSFE(\mathbf{w}, \mathbf{\Sigma})$: $\left|\frac{MSFE(\widehat{\mathbf{w}},\widehat{\mathbf{\Sigma}})}{MSFE(\mathbf{w},\mathbf{\Sigma})} - 1\right| = \mathcal{O}_P(\varrho_{1T}d_T s_T) = o_P(1)$.

(ii) If $\varrho_{2T}\bar{d}^2 \xrightarrow{p} 0$, *Algorithm 10 consistently estimates* $MSFE(\mathbf{w}, \mathbf{\Sigma})$: $\left|\frac{MSFE(\widehat{\mathbf{w}},\widehat{\mathbf{\Sigma}})}{MSFE(\mathbf{w},\mathbf{\Sigma})} - 1\right| = \mathcal{O}_P(\varrho_{2T}\bar{d}^2) = o_P(1)$.

Proofs of Theorems 12-13 can be found in Section 9. Note that the rates of convergence for MSFE and precision matrix $\mathbf{\Theta}$ are the same and both are faster than the combination weight rates in Theorem 12. In contrast to classical graphical models in Algorithms 7-8, the convergence properties of which were examined by [76] among others, the rates in Theorems 12-13 depend on the sparsity of $\mathbf{\Theta}_\varepsilon$ rather than of $\mathbf{\Theta}$. This means that instead of assuming that many partial correlations of forecast errors $\mathbf{e}_t$ are negligible, which is not realistic under the factor structure, we impose a milder restriction requiring many partial correlations of $\boldsymbol{\varepsilon}_t$ to be negligible once the common components have been taken into account. Similarly to the comparison of precision matrix $\boldsymbol{\Theta}$ obtained by two graphical models, if $d_T < s_T$ Factor MB is expected to converge faster for combination weights and MSFE. In our simulations the rates of Factor Graphical models are comparable, whereas an empirical application shows that for most macroeconomic series that we studied Factor GLASSO outperforms Factor MB. This suggests that for macroeconomic forecasting using weighted penalized log-likelihood and running p coupled LASSO problems for estimating precision matrix is preferable to fitting p separate LASSO regressions using each variable as the response and the others as predictors.

4.6 Monte Carlo

We divide the simulation results into two subsections. In the first subsection we study the consistency of the Factor GLASSO and Factor MB for estimating precision matrix and the combination weights. In the second subsection we evaluate the out-of-sample forecasting performance of combined forecasts based on the Factor Graphical models from Algorithms 9-10 in terms of the mean-squared forecast error. We compare the performance of forecast combinations based on the factor models with equal-weighted (EW) forecast combination, forecast combinations using GLASSO and nodewise regression from Algorithms 7-8. Similarly to the literature on graphical models, all exercises use 100 Monte Carlo simulations.

4.6.1 Consistent Estimation of forecast combination weights based on FGM

We consider sparse Gaussian graphical models which may be fully specified by a precision matrix $\mathbf{\Theta}_0$. Therefore, the random sample is distributed as $\mathbf{e}_t = (e_{1t}, \dots, e_{pt})' \sim \mathcal{N}(0, \mathbf{\Sigma}_0)$, where $\mathbf{\Theta}_0 = (\mathbf{\Sigma}_0)^{-1}$ for $t = 1, \dots, T,\ j = 1, \dots, p$. Let $\widehat{\mathbf{\Theta}}$ be the precision matrix estimator. We show consistency of the Factor GLASSO (Algorithm 9) and Factor MB (Algorithm 10), in (i) the operator norm, $\left\|\left\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0\right\|\right\|_2$, (ii) ℓ_1/ℓ_1-matrix norm, $\left\|\left\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0\right\|\right\|_1$, and (iii) in ℓ_1-vector norm for the combination weights, $\|\widehat{\mathbf{w}} - \mathbf{w}\|_1$, where $\mathbf{w}$ is given by (4.15).

The forecast errors are assumed to have the following structure:

$$\underbrace{\mathbf{e}_t}_{p\times 1} = \mathbf{B} \underbrace{\mathbf{f}_t}_{q\times 1} + \boldsymbol{\varepsilon}_t, \quad t = 1, \dots, T \tag{4.28}$$

$$\mathbf{f}_t = \phi_f \mathbf{f}_{t-1} + \boldsymbol{\zeta}_t, \tag{4.29}$$

where $\mathbf{e}_t$ is a $p \times 1$ vector of forecast errors following $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, $\mathbf{f}_t$ is a $q \times 1$ vector of factors, $\mathbf{B}$ is a $p \times q$ matrix of factor loadings, ϕ_f is an autoregressive parameter in the factors which is a scalar for simplicity, $\boldsymbol{\zeta}_t$ is a $q \times 1$ random vector with each component independently following $\mathcal{N}(0, \sigma^2_\zeta)$, $\boldsymbol{\varepsilon}_t$ is a $p \times 1$ random vector following $\mathcal{N}(0, \mathbf{\Sigma}_\varepsilon)$, with sparse $\mathbf{\Theta}_\varepsilon$ that has a random graph structure described below. To create $\mathbf{B}$ in (4.28) we take the first q columns of an upper triangular matrix from a Cholesky decomposition of the $p \times p$ Toeplitz matrix parameterized by ρ: that is, $\mathbf{B} = (b)_{ij}$, where $(b)_{ij} = \rho^{|i-j|}$, $i, j \in \{1, \dots, p\}$. We set $\rho = 0.2$, $\phi_f = 0.2$ and $\sigma^2_\zeta = 1$.

The specification in (4.28) leads to the low-rank plus sparse decomposition of the covariance matrix:

$$\mathbb{E}\left[\mathbf{e}_t\mathbf{e}_t'\right] = \mathbf{\Sigma} = \mathbf{B}\mathbf{\Sigma}_f\mathbf{B}' + \mathbf{\Sigma}_\varepsilon. \tag{4.30}$$

When $\mathbf{\Sigma}_\varepsilon$ has a sparse inverse $\mathbf{\Theta}_\varepsilon$, it leads to the low-rank plus sparse decomposition of the precision matrix $\mathbf{\Theta}$, such that $\mathbf{\Theta}$ can be expressed as a function of the low-rank $\mathbf{\Theta}_f$ plus sparse $\mathbf{\Theta}_\varepsilon$.

We consider the following setup: let $p = T^\delta$, $\delta = 0.85$, $q = 2(\log(T))^{0.5}$ and $T = [2^\kappa]$, for $\kappa = 7, 7.5, 8, \ldots, 9.5$. Our setup allows the number of individual forecasts, p, and the number of common factors in the forecast errors, q, to increase with the sample size, T.

A sparse precision matrix of the idiosyncratic components $\mathbf{\Theta}_\varepsilon$ is constructed as follows: we first generate the adjacency matrix using a random graph structure. Define a $p \times p$ adjacency matrix $\mathbf{A}_\varepsilon$ which represents the structure of the graph:

$$a_{\varepsilon,ij} = \begin{cases} 1, & \text{for } i \neq j \ \text{ with probability } \pi, \\ 0, & \text{otherwise,} \end{cases} \tag{4.31}$$

where $a_{\varepsilon,ij}$ denotes the i, j-th element of the adjacency matrix $\mathbf{A}_\varepsilon$. We set $a_{\varepsilon,ij} = a_{\varepsilon,ji} = 1$, for $i \neq j$ with probability π, and 0 otherwise. Such structure results in $s_T = p(p-1)\pi/2$ edges in the graph. To control sparsity, we set $\pi = 1/(pT^{0.8})$, which makes $s_T = \mathcal{O}(T^{0.05})$. The adjacency matrix has all diagonal elements equal to zero. Hence, to obtain a positive definite precision matrix we apply the procedure described in [142]: using their notation, $\mathbf{\Theta}_\varepsilon = \mathbf{A}_\varepsilon \cdot v + \mathbf{I}(|\tau| + 0.1 + u)$, where $u > 0$ is a positive number added to the diagonal of the precision matrix to control the magnitude of partial correlations, v controls the magnitude of partial correlations with u, and τ is the smallest eigenvalue of $\mathbf{A}_\varepsilon \cdot v$. In our simulations we use $u = 0.1$ and $v = 0.3$.

Figures 4.6.1-4.6.2 show the averaged (over Monte Carlo simulations) errors of the estimators of the precision matrix $\mathbf{\Theta}$ and the optimal combination weight versus the sample size T in the logarithmic scale (base 2). The estimate of the precision matrix of the EW forecast combination is obtained using the fact that diagonal covariance and precision matrices imply equal weights. To determine the values of the diagonal elements we use the shrinkage intensity coefficient calculated as the average of the eigenvalues of the sample covariance matrix of the forecast errors (see [91]). As evidenced by Figures 4.6.1-4.6.2, Factor GLASSO and Factor MB demonstrate superior performance over EW and non-factor based models (GLASSO and MB). Furthermore, our method achieves lower estimation error in the combination weights (4.17), which leads to lower risk of the combined forecast as shown in (4.16). Interestingly, even though the precision matrix estimated using Factor MB has faster convergence rate in $|||\cdot|||_2$ and $|||\cdot|||_1$ norms as compared to Factor GLASSO, the weights estimated using Factor GLASSO converge faster. Also, note that the precision matrix estimated using the EW method also shows good convergence properties. However, in terms of estimating the combination weight, the performance of EW does not exhibit convergence properties. This is in agreement with previously reported findings ([125]) that equal weights are not theoretically optimal, however, as demonstrated in the next subsection, the EW combination still leads to a relatively good performance in terms of MSFE although the FGM-based combinations outperform it.

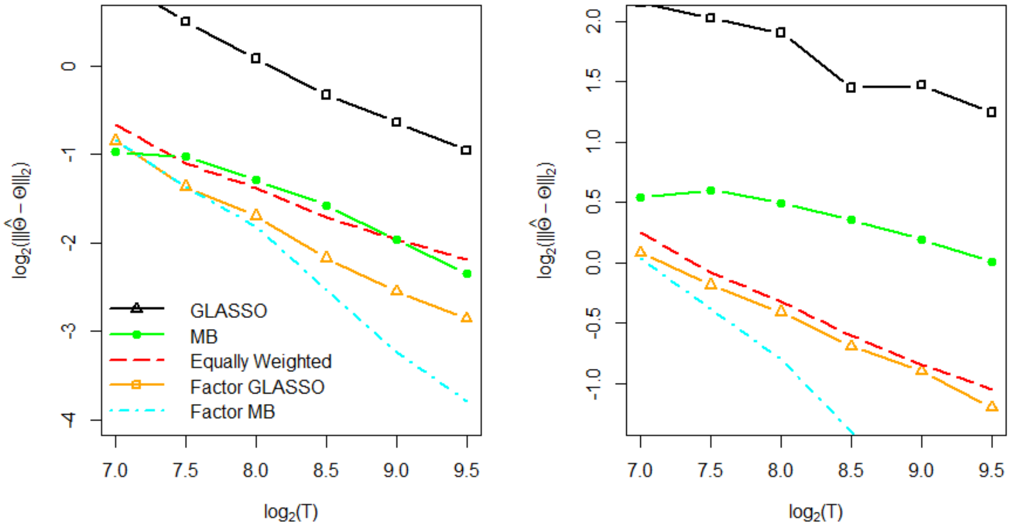

Figure 4.6.1: **Averaged errors of the estimators of Θ on logarithmic scale (base 2).** $p = T^{0.85}$**,** $q = 2(\log(T))^{0.5}$**,** $s_T = \mathcal{O}(T^{0.05})$**.**

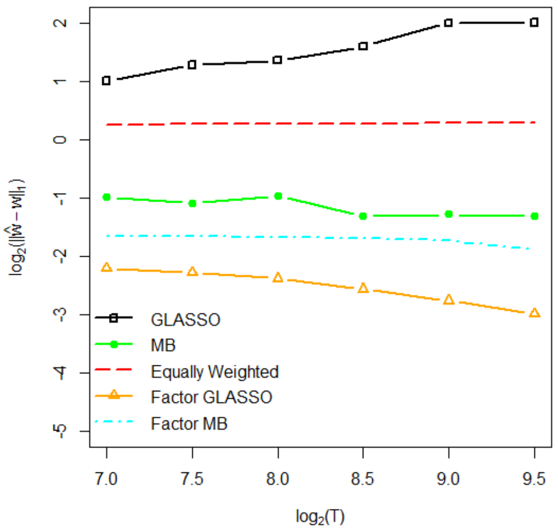

Figure 4.6.2: **Averaged errors of the estimator of w (base 2) on logarithmic scale.** $p = T^{0.85}$**,** $q = 2(\log(T))^{0.5}$**,** $s_T = \mathcal{O}(T^{0.05})$**.**

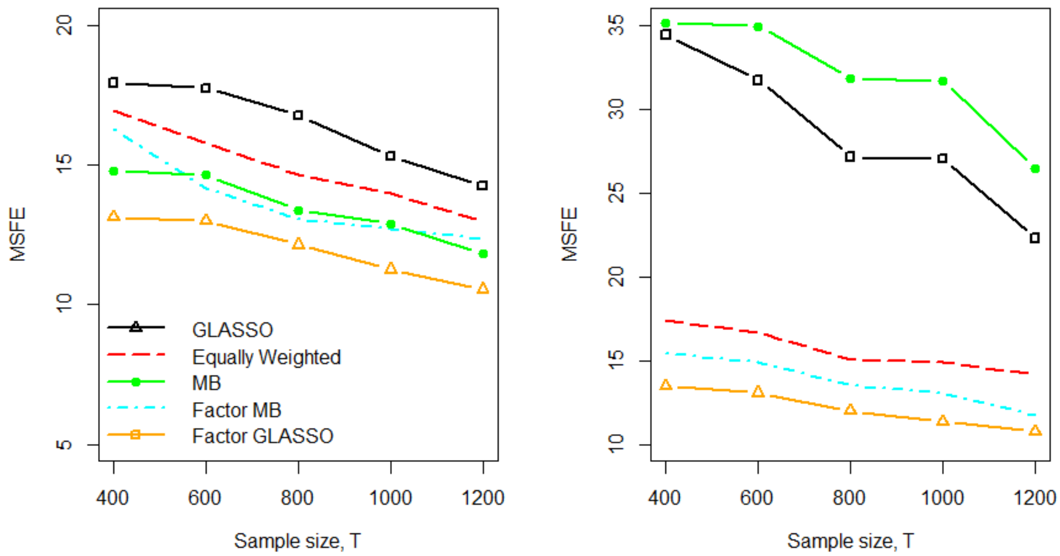

Figure 4.6.3: **Plots of the MSFE over the sample size** $\boldsymbol{T}$**.** $c_1 = 0$ (left), $c_1 = 0.75$ (right), $c_2 = 0.9,\ N = 100,\ r = 5, \sigma_\xi = 1,\ L = 7,\ K = 2,\ p = 24,\ q = 5,\ \rho = 0.9,\ \phi = 0.8.$

4.6.2 Comparing Performance of forecast combinations based on FGM

We consider the standard forecasting model in the literature (e.g., [126]), which uses the factor structure of the high dimensional predictors. Suppose the data is generated from the following data generating process (DGP):

$$\mathbf{x}_t = \mathbf{\Lambda} \mathbf{g}_t + \mathbf{v}_t, \tag{4.32}$$

$$\mathbf{g}_t = \phi \mathbf{g}_{t-1} + \boldsymbol{\xi}_t, \tag{4.33}$$

$$y_{t+1} = \mathbf{g}_t' \boldsymbol{\alpha} + \sum_{s=1}^{\infty} \theta_s \epsilon_{t+1-s} + \epsilon_{t+1}, \tag{4.34}$$

where y_{t+1} is a univariate series of our interest in forecasting, $\mathbf{x}_t$ is an $N \times 1$ vector of regressors (predictors), $\boldsymbol{\beta}$ is an $N \times 1$ parameter vector, $\mathbf{g}_t$ is an $r \times 1$ vector of factors, $\mathbf{\Lambda}$ is an $N \times r$ matrix of factor loadings, $\mathbf{v}_t$ is an $N \times 1$ random vector following $\mathcal{N}(0, \sigma_v^2)$, ϕ is an autoregressive parameter in the factors which is a scalar for simplicity, $\boldsymbol{\xi}_t$ is an $r \times 1$ random vector with each component independently following $\mathcal{N}(0, \sigma_\xi^2)$, ϵ_{t+1} is a random error following $\mathcal{N}(0, \sigma_\epsilon^2)$, and $\boldsymbol{\alpha}$ is an $r \times 1$ parameter vector which is drawn randomly from $\mathcal{N}(1, 1)$. We set $\sigma_\epsilon = 1$. The coefficients θ_s are set according to the rule

$$\theta_s = (1+s)^{c_1} c_2^s, \tag{4.35}$$

as in [70]. We set $c_1 \in \{0, 0.75\}$ and $c_2 \in \{0.6, 0.7, 0.8, 0.9\}$. We generate r factors using (4.33) with a grid of 10 different AR(1) coefficients ϕ equidistant between 0 and 0.9. To create $\mathbf{\Lambda}$ in (4.32) we take the first r rows of an upper triangular matrix from a Cholesky decomposition of the $N \times N$ Toeplitz matrix parameterized by ρ. We consider a grid of 10 different values of ρ equidistant between 0 and 0.9. One-step ahead forecasts are estimated from the factor-augmented autoregressive (FAR) models of orders k, l, denoted as FAR(k, l):

$$\hat{y}_{t+1} = \hat{\mu} + \hat{\kappa}_1 \hat{g}_{1,t} + \cdots + \hat{\kappa}_k \hat{g}_{k,t} + \hat{\psi}_1 y_t + \cdots + \hat{\psi}_l y_{t+1-l}, \tag{4.36}$$

where the factors $(\hat{g}_{1,t}, \ldots, \hat{g}_{k,t})$ are estimated from equation (4.32). We consider the FAR models of various orders, with $k = 1, \ldots, K$ and $l = 1, \ldots, L$. We also consider the models without any lagged y or any factors. Therefore, the total number of forecasting models is $p \equiv (1 + K) \times (1 + L)$, which includes the forecasting models using naive average or no factors.

The total number of observations is T, and the number of observations in the regression period (the train sample) is set to be the first half of the sample, $t = 1, \ldots, m \equiv T/2$, to leave the second half of the sample, $t = m+1, \ldots, T$, for the out-of-sample evaluation (the test sample). We roll the estimation window over the test sample of the size $n \equiv T - m$, to update all the estimates in each point of time $t = 1, \ldots, m$. Recall that q denotes the number of factors in the forecast errors as in equation (3.31). We first examine the properties of the combined forecasts based on the Factor Graphical models when T and p vary and compare their performance with the combined forecasts based on the GLASSO, MB and EW forecasts.

We consider a low-dimensional setup to demonstrate the advantage of using FGM even when the number of forecasts, p, is small relative to the sample size, T: (1) in such scenario EW has an advantage since there are not many models to combine and assigning equal weights should produce satisfactory performance, and (2) non-factor based models have the advantage over the models that estimate factors due to the estimation errors. As a result, this framework with the low-dimensional setup is favorable to EW and non-factor based models. Figure 4.6.3 shows the MSFE for different sample sizes and fixed parameters: we report the results for two values of $c_1 \in \{0, 0.75\}$. As evidenced from Figure 4.6.3, the models that use the factor structure outperform EW combination and non-factor based counterparts for both values of c_1. We see that Factor GLASSO, in general, has lower MSFE than Factor MB. This finding is further supported by our empirical application in Section 7.

In Appendix 4.A we examine the sensitivity of the competing models with respect to variation in the DGP parameters such as number of predictors N, values of c_2, ϕ, the strength of factor loadings ρ, and the number of factors q. We conclude that Factor Graphical Models outperform equally-weighted combinations and the graphical models without factors.

4.7 Application of FGM for Macroeconomic Forecasting

An empirical application to forecasting macroeconomic time series in big data environment highlights the advantage of both Factor Graphical models described in Algorithms 9-10 in comparison with the existing methods of forecast combination. We use a large monthly frequency macroeconomic database of [107], who provide a comprehensive description of the dataset and 128 macroeconomic series. We consider the time period 1960:01-2020:07 with the total number of observations $T = 726$, the training sample consists of $m = 120$ observations, and the test sample $n \equiv T - m - h + 1$, where h is the forecast horizon. We roll the estimation window over the test sample to update all the estimates in each point of time $t = m, \ldots, T - h$. We estimate h-step ahead forecasts from FAR(k, l) which were defined in (4.36) with $k = 0, 1, \ldots, K = 9$, and $l = 0, 1, \ldots, L = 11$. The total number of forecasting models is $p = 120$. The optimal number of factors in the forecast errors (denoted as q in equation (4.9)) is chosen using the standard data-driven method that uses the information criterion IC1 described in [6]. We note that in the majority of the cases the optimal number of factors was estimated to be equal to 1.

Table 4.7.1 compares the performance of the Factor GLASSO and Factor MB with the competitors for predicting seven representative macroeconomic indicators of the US economy: monthly industrial production (INDPRO), S&P500 composite index (S&P500), Consumer Price Index (CPIAUCSL), real personal consumption (DPCERA3MO86SBEA), M1 money stock (M1SL), civilian unemployment rate (UNRATE), and the effective federal funds rate (FEDFUNDS) using 127 remaining macroeconomic series. Let $\{Y_t\}_{t=1}^{T}$ be the series of interest for forecasting. Similarly to [37], for INDPROD, S&P500, CPI, Real Personal Consumption and M1 Money Stock we forecast the average growth rate (with logs):

$$y_{t+h} = \frac{1}{h} \ln(Y_{t+h}/Y_t). \tag{4.37}$$

For UNRATE we forecast the average change (without logs):

$$y_{t+h} = \frac{1}{h}(Y_{t+h}/Y_t). \tag{4.38}$$

And for FEDFUNDS we forecast the log of the series:

$$y_{t+h} = \ln(Y_{t+h}). \tag{4.39}$$

Table 4.7.1 reports MSFEs of the competing methods with the smallest MSFE in each row in bold font. As evidenced from Table 4.7.1, our methods outperform EW, GLASSO and nodewise regression: accounting for the factor structure results in lower MSFE. Therefore, the FGM framework developed in this paper leads to the superior performance of the combined forecast as compared to EW model even when the models/experts do not contain a lot of unique information. Our empirical application demonstrates that this finding does not originate from the difference in the performance of EW vs graphical models: as evidenced from Table 4.7.1, the performance of GLASSO is worse than that of EW for the FEDFUNDS series, whereas Factor GLASSO outperforms EW. A similar pattern is observed in the performance of nodewise regression for M1 Money Stock. Therefore, the improvement in the combined forecast comes from incorporating the factor structure of the forecast errors into the graphical models. Note that in contrast with EW and non-factor based methods, the performance of Factor GLASSO and Factor MB does not deteriorate significantly when the forecast horizon, h, increases. Notice, however, that Factor Graphical Models tend to perform better for $h \geq 2$. In other words, accounting for common factors in forecast errors has greater benefit for longer horizons. Finally, for most series Factor GLASSO outperforms Factor MB, suggesting that for macroeconomic forecasting using weighted penalized log-likelihood and running p coupled LASSO problems for estimating precision matrix is

INDPRO					
h	EW	GLASSO	Factor GLASSO	MB	Factor MB
1	2.77E-04	1.51E-04	**1.24E-04**	2.23E-04	1.28E-04
2	3.26E-04	1.79E-04	**5.59E-05**	1.61E-04	1.38E-04
3	1.55E-04	9.77E-05	**3.81E-05**	1.17E-04	6.54E-05
4	1.18E-04	7.60E-05	**2.38E-05**	1.03E-04	2.65E-05
S&P500					
1	1.40E-03	1.39E-03	1.37E-03	**1.34E-03**	9.57E-03
2	1.71E-03	1.44E-03	**8.95E-04**	1.55E-03	1.01E-03
3	1.66E-03	1.34E-03	**3.48E-04**	1.43E-03	6.69E-04
4	1.27E-03	1.06E-03	**3.95E-04**	9.55E-04	7.91E-04
CPI: All Items					
1	6.88E-06	6.75E-06	**5.84E-06**	6.46E-06	8.98E-06
2	1.05E-05	1.06E-05	**8.39E-06**	9.93E-06	9.93E-06
3	1.52E-05	1.47E-05	**9.36E-06**	1.56E-05	1.34E-05
4	1.63E-05	1.63E-05	**7.00E-06**	1.60E-05	1.14E-05
Real Personal Consumption					
1	3.05E-05	**2.70E-05**	4.18E-05	2.88E-05	2.74E-05
2	2.65E-04	8.52E-05	2.79E-05	8.11E-05	**2.39E-05**
3	7.94E-04	1.41E-04	2.91E-05	6.42E-05	**2.84E-05**
4	8.65E-04	7.87E-04	**2.61E-05**	6.42E-05	2.63E-05
M1 Money Stock					
1	5.42E-05	5.18E-05	**4.99E-05**	5.40E-05	5.47E-05
2	5.82E-05	1.58E-04	7.27E-05	5.86E-05	**5.40E-05**
3	5.97E-05	1.56E-04	7.44E-05	5.96E-05	**5.64E-05**
4	5.97E-05	1.63E-04	6.97E-05	5.94E-05	**5.78E-05**
UNRATE					
1	0.2531	0.0858	0.0109	0.0557	**0.0107**
2	0.3758	0.1334	**0.0066**	0.0448	0.0081
3	0.0743	0.0651	0.0066	0.0532	**0.0051**
4	2.1999	0.6871	**0.1578**	1.0973	0.2510
FEDFUNDS					
1	0.0609	0.1813	**0.0205**	0.0424	0.0448
2	0.1426	1.2230	**0.0288**	0.0675	0.0416
3	0.2354	1.2710	**0.0508**	0.1217	0.1038
4	0.3702	1.4672	**0.0592**	0.2470	0.1962

Table 4.7.1: Prediction of Monthly Macroeconomic Variables. The numbers are MSFEs with the smallest MSFE in each row in bold font. *h* indicates the forecast horizon, EW stands for the “Equal-Weighted” forecast, GLASSO and MB are the models that do not use the factor structure in the forecast errors.

preferable to fitting p separate LASSO regressions using each variable as the response and the others as predictors.

4.8 Conclusions

In this paper we overcome the challenge of using graphical models under the factor structure and provide a simple framework that allows practitioners to combine a large number of forecasts when experts tend to make common mistakes. Our new approach to forecast combinations breaks down forecast errors into common and unique parts which improves the accuracy of the combined forecast. The proposed algorithms, Factor Graphical Models, are shown to consistently estimate forecast combination weights and MSFE. Extensive simulations and empirical applications to macroeconomic forecasting in big data environment reveal that FGM outperforms equal-weighted forecasts and combined forecasts produced using graphical models without factors. With the superior performance observed at all forecast horizons, we find that the greater benefit from accounting for the common factors is evidenced at longer horizons.

4.9 Proofs of Theorems

In this section we collected the proofs of Theorems 12-13. We first present a Lemma which is used in the theoretical derivations.

Lemma 14 *Let* $l \in \{1,2\} \equiv \{\textit{Factor GLASSO}, \textit{Factor MB}\}$.

(a) $\vert\vert\vert \mathbf{\Theta} \vert\vert\vert_1 = \mathcal{O}(\kappa_{1l})$*, where* $\kappa_{1l} = d_T$ *if* $l = 1$ *which corresponds to Factor GLASSO, and* $\kappa_{1l} = \bar{d}$ *if* $l = 2$ *which corresponds to Factor MB. This will be further abbreviated as* $\kappa_{1l} \in \{d_T, \bar{d}\}_{l=1,2}$.

(b) $a \geq C_0 > 0$*, where* a *was defined in Section 3 and* C_0 *is a positive constant representing the minimal eigenvalue of* $\mathbf{\Theta}$.

(c) $|\widehat{a} - a| = \mathcal{O}_P(\kappa_{2l})$*, where* $\widehat{a}$ *was defined in Section 3 and* $\kappa_{2l} \in \{\varrho_{1T} d_T s_T, \varrho_{2T} \bar{d}^2\}_{l=1,2}$.

Proof.

(a) To prove part (a) we use the following matrix inequality which holds for any $\mathbf{A} \in \mathcal{S}_p$:

$$\vert\vert\vert \mathbf{A} \vert\vert\vert_1 = \vert\vert\vert \mathbf{A} \vert\vert\vert_\infty \leq \sqrt{d(\mathbf{A})} \vert\vert\vert \mathbf{A} \vert\vert\vert_2, \tag{4.40}$$

where $d(\mathbf{A})$ was defined at the beginning of Section 5. The proof of (4.40) is a straightforward consequence of the Schwarz inequality.

Sherman-Morrison-Woodbury formula together with (4.40) and Assumptions **(B.1)**-**(B.3)** yield:

$$\begin{aligned}\vert\vert\vert \mathbf{\Theta} \vert\vert\vert_1 &\leq \vert\vert\vert \mathbf{\Theta}_\varepsilon \vert\vert\vert_1 + \left\vert\left\vert\left\vert \mathbf{\Theta}_\varepsilon \mathbf{B}[\mathbf{\Theta}_f + \mathbf{B}'\mathbf{\Theta}_\varepsilon \mathbf{B}]^{-1} \mathbf{B}' \mathbf{\Theta}_\varepsilon \right\vert\right\vert\right\vert_1 \\ &= \mathcal{O}(\sqrt{\kappa_{1l}}) + \mathcal{O}\Big(\sqrt{\kappa_{1l}} \cdot p \cdot \frac{1}{p} \cdot \sqrt{\kappa_{1l}}\Big) = \mathcal{O}(\kappa_{1l}). \end{aligned} \tag{4.41}$$

(b) Assumption **(A.1)** states that the minimal eigenvalue of $\mathbf{\Theta}$ is bounded away from zero, hence,

$$a = \boldsymbol{\iota}_p' \mathbf{\Theta} \boldsymbol{\iota}_p / p \geq C_0 > 0.$$

(c) Using the Hölders inequality, we have

$$|\widehat{a}-a| = \left|\frac{\boldsymbol{\iota}_p'(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta})\boldsymbol{\iota}_p}{p}\right| \leq \frac{\left\|(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta})\boldsymbol{\iota}_p\right\|_1 \|\boldsymbol{\iota}_p\|_\infty}{p} \leq \left\|\left|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\right|\right\|_1$$

$$= \mathcal{O}_P(\kappa_{2l}) = o_P(1),$$

where the last rate is obtained using the assumptions of Theorem 12.

■

4.9.1 Proof of Theorem 12

First, note that the forecast combination weight can be written as

$$\begin{aligned}\widehat{\mathbf{w}}-\mathbf{w} &= \frac{\Big((a\widehat{\boldsymbol{\Theta}}\boldsymbol{\iota}_p)-(\hat{a}\boldsymbol{\Theta}\boldsymbol{\iota}_p)\Big)/p}{\hat{a}a} \\ &= \frac{\Big((a\widehat{\boldsymbol{\Theta}}\boldsymbol{\iota}_p)-(a\boldsymbol{\Theta}\boldsymbol{\iota}_p)+(a\boldsymbol{\Theta}\boldsymbol{\iota}_p)-(\hat{a}\boldsymbol{\Theta}\boldsymbol{\iota}_p)\Big)/p}{\hat{a}a}.\end{aligned}$$

As shown in [24], the above can be rewritten as

$$\|\widehat{\mathbf{w}}-\mathbf{w}\|_1 \leq \frac{a\frac{\left\|(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta})\boldsymbol{\iota}_p\right\|_1}{p}+|a-\widehat{a}|\frac{\|\boldsymbol{\Theta}\boldsymbol{\iota}_p\|_1}{p}}{|\widehat{a}|a}. \tag{4.42}$$

Prior to bounding the terms in (4.42), we first present an inequality which is used in the derivations. Let $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\mathbf{v} \in \mathbb{R}^{p\times 1}$. Also, let $\mathbf{A}_j$ and $\mathbf{A}'_j$ be a $p \times 1$ and $1 \times p$ row and column vectors in $\mathbf{A}$, respectively.

$$\begin{aligned} \|\mathbf{A}\mathbf{v}\|_1 &= \left|\mathbf{A}'_1\mathbf{v}\right| + \ldots + \left|\mathbf{A}'_p\mathbf{v}\right| \leq \|\mathbf{A}_1\|_1 \|\mathbf{v}\|_\infty + \ldots + \|\mathbf{A}_p\|_1 \|\mathbf{v}\|_\infty \\ &= \left(\sum_{j=1}^{p} \|\mathbf{A}_j\|_1\right) \|\mathbf{v}\|_\infty \leq p \max_j |\mathbf{A}_j|_1 \|\mathbf{v}\|_\infty. \end{aligned} \tag{4.43}$$

Hölders inequality was used to obtain each inequality in (4.43). If $\mathbf{A} \in \mathcal{S}_p$, then the last expression can be further reduced to $p|||\mathbf{A}|||_1 \|\mathbf{v}\|_\infty$.

Let us now bound the right-hand side of (4.42). In the numerator we have:

$$\frac{\left\|(\widehat{\mathbf{\Theta}} - \mathbf{\Theta})\boldsymbol{\iota}_p\right\|_1}{p} \leq |||\mathbf{\Theta}|||_1 = \mathcal{O}_P(\kappa_{3l}), \tag{4.44}$$

where $\kappa_{3l} \in \{\varrho_{1T} d_T s_T, \varrho_{2T} \bar{d}^2\}_{l=1,2}$, the rates were derived in [97, 122] as discussed at the beginning of Section 5, and the inequality follows from (4.43).

$$\frac{\|\mathbf{\Theta}\boldsymbol{\iota}_p\|_1}{p} \leq |||\mathbf{\Theta}|||_1 = \mathcal{O}(\kappa_{1l}), \tag{4.45}$$

where the rate follows from Lemma 14 (a) and the inequality is obtained from (4.43). Combining (4.44), (4.45), and Lemma 14 (c) we get:

$$a\frac{\left\|(\widehat{\mathbf{\Theta}} - \mathbf{\Theta})\boldsymbol{\iota}_p\right\|_1}{p} + |a - \widehat{a}|\frac{\|\mathbf{\Theta}\boldsymbol{\iota}_p\|_1}{p} = \mathcal{O}(1) \cdot \mathcal{O}_P(\kappa_{3l}) + \mathcal{O}_P(\kappa_{2l}) \cdot \mathcal{O}(\kappa_{1l}) = \mathcal{O}_P(\kappa_{4l}) = o_P(1), \tag{4.46}$$

where $\kappa_{4l} \in \{\varrho_{1T} d_T^2 s_T, \varrho_{2T} \bar{d}^3\}_{l=1,2}$ and the last equality holds under the assumptions of Theorem 12.

For the denominator of (4.42) it easy to see that $|\widehat{a}|a = \mathcal{O}_P(1)$ using the results of Lemma 14 (b).

4.9.2 Proof of Theorem 13

Using Lemma 14 (b)-(c), we get

$$\left| \frac{\hat{a}^{-1}}{a^{-1}} - 1 \right| = \frac{|a - \hat{a}|}{|\hat{a}|} = \mathcal{O}_P(\kappa_{2l}) = o_P(1),$$

where the last rate is obtained using the assumptions of Theorem 13.

Appendices

4.A Additional Simulations

Figures 4.A.1-4.A.5 show the performance in terms of MSFE for different number of predictors N, different values of c_2, ϕ, ρ and q: Factor-based models (Factor GLASSO and Factor MB) outperform the equal-weighted forecast combination and the standard GLASSO and nodewise regression without any factor structure. As evidenced from the figures, these findings are robust to the changes in the model parameters. Importantly, Figure 4.A.5 shows the scenario when the true number of principal components, r, is equal to 5, whereas none of the forecasters use PCA for prediction: in this case including at least 2 common components of the forecasting errors reduces MSFE, such that Factor GLASSO and Factor MB outperform EW forecast combination.

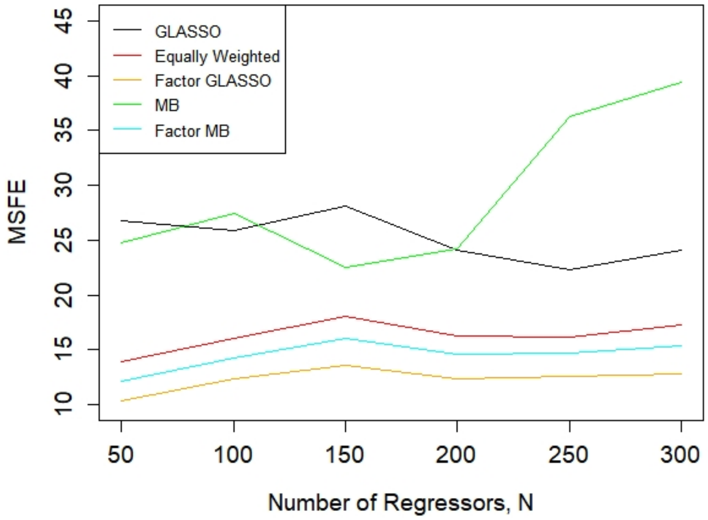

Figure 4.A.1: **Plots of the MSFE over the number of predictors $\boldsymbol{N}$.** $c_1 = 0.75,\ c_2 = 0.9,$
$T = 800,\ r = 5,\ \sigma_\xi = 1,\ L = 7,\ K = 2,\ p = 24,\ q = 5,\ \rho = 0.9,\ \phi = 0.8.$

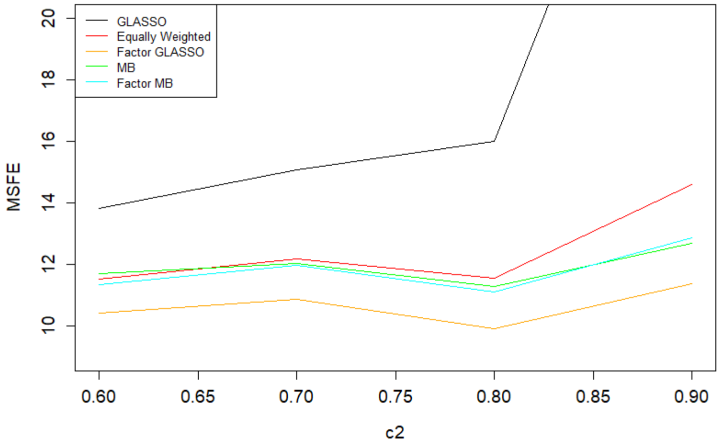

Figure 4.A.2: **Plots of the MSFE over the values of $\boldsymbol{c_2}$.** $c_1 = 0.75,\ c_2 \in \{0.6, 0.7, 0.8, 0.9\},$
$T = 800,\ N = 100,\ r = 5,\ \sigma_\xi = 1,\ L = 7,\ K = 2,\ p = 24,\ q = 5,\ \rho = 0.9,\ \phi = 0.8.$

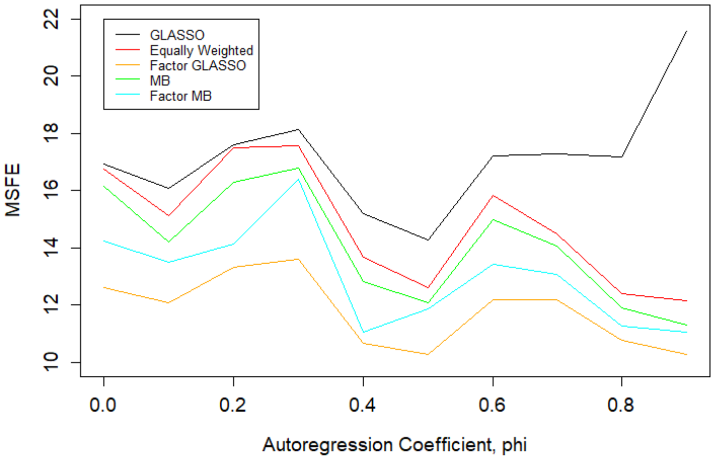

Figure 4.A.3: **Plots of the MSFE over the values of ϕ.** $c_1 = 0.75,\ c_2 = 0.8,\ T = 800,\ N = 100,\ r = 5,\ \sigma_\xi = 1,\ L = 7,\ K = 2,\ p = 24,\ q = 5,\ \rho = 0.9,\ \phi \in \{0, 0.1, \ldots, 0.9\}$.

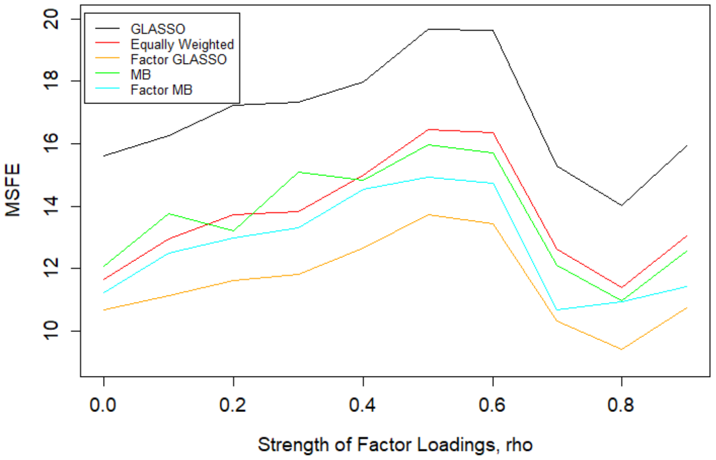

Figure 4.A.4: **Plots of the MSFE over the values of ρ.** $c_1 = 0.75,\ c_2 = 0.8,\ T = 800,\ N = 100,\ r = 5,\ \sigma_\xi = 1,\ L = 7,\ K = 2,\ p = 24,\ q = 5,\ \rho \in \{0, 0.1, \ldots, 0.9\},\ \phi = 0.7$.

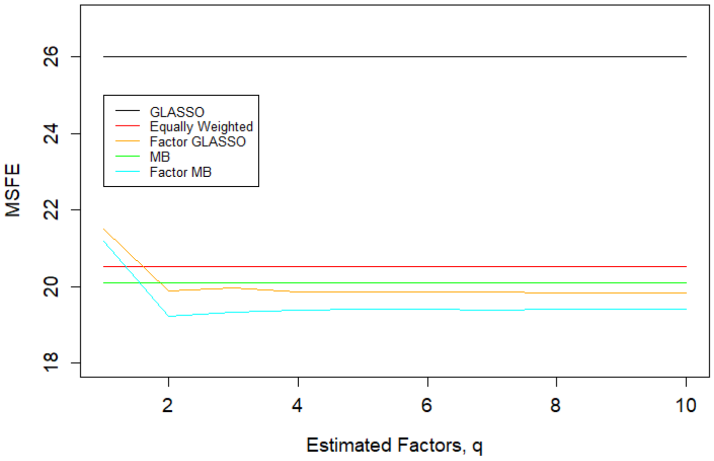

Figure 4.A.5: **Plots of the MSFE over the values of q.** $c_1 = 0.75,\ c_2 = 0.9,\ T = 800,$ $N = 100,\ r = 5,\ \sigma_\xi = 1,\ L = 12,\ K = 0,\ p = 13,\ q \in \{0, 1, \ldots, 10\},\ \rho = 0.9,\ \phi = 0.8.$

Bibliography

[1] Yacine Ait-Sahalia and Dacheng Xiu. Using principal component analysis to estimate a high dimensional factor model with high-frequency data. *Journal of Econometrics*, 201(2):384–399, 2017.

[2] Mengmeng Ao, Li Yingying, and Xinghua Zheng. Approaching mean-variance efficiency for large portfolios. *The Review of Financial Studies*, 32(7):2890–2919, 2019.

[3] Amir F. Atiya. Why does forecast combination work so well? *International Journal of Forecasting*, 36(1):197–200, 2020.

[4] Oluwatoyin Abimbola Awoye. *Markowitz Minimum Variance Portfolio Optimization Using New Machine Learning Methods*. PhD thesis, University College London, 2016.

[5] Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171, 2003.

[6] Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

[7] Jushan Bai and Serena Ng. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150, 2006.

[8] Natalia Bailey, George Kapetanios, and M Hashem Pesaran. Measurement of factor strength: Theory and practice. 2020.

[9] Gah-Yi Ban, Noureddine El Karoui, and Andrew EB Lim. Machine learning and portfolio optimization. *Management Science*, 64(3):1136–1154, 2018.

[10] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, June 2008.

[11] Matteo Barigozzi, Christian Brownlees, and Gábor Lugosi. Power-law partial correlation network models. *Electronic Journal of Statistics*, 12(2):2905–2929, 2018.

[12] J. M. Bates and C. W. J. Granger. The combination of forecasts. *OR*, 20(4):451–468, 1969.

[13] Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 05 2013.

[14] Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102(1):77–94, 2015.

[15] Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 12 2008.

[16] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag, Berlin, Heidelberg, 2006.

[17] Joshua Brodie, Ingrid Daubechies, Christine De Mol, Domenico Giannone, and Ignace Loris. Sparse and stable markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272, 2009.

[18] Christian Brownlees, Eulàlia Nualart, and Yucheng Sun. Realized networks. *Journal of Applied Econometrics*, 33(7):986–1006, 2018.

[19] Fabio Caccioli, Imre Kondor, Matteo Marsili, and Susanne Still. Liquidity risk and instabilities in portfolio optimization. *International Journal of Theoretical and Applied Finance*, 19(05):1650035, 2016.

[20] T. Tony Cai, Jianchang Hu, Yingying Li, and Xinghua Zheng. High-dimensional minimum variance portfolio estimation based on high-frequency data. *Journal of Econometrics*, 214(2):482–494, 2020.

[21] T Tony Cai, Weidong Liu, Harrison H Zhou, et al. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Annals of Statistics*, 44(2):455–488, 2016.

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

[23] Tony Cai, Weidong Liu, and Xi Luo. A constrained l1-minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[24] Laurent Callot, Mehmet Caner, A. Özlem Önder, and Esra Ulaşan. A nodewise regression approach to estimating large portfolios. *Journal of Business & Economic Statistics*, 0(0):1–12, 2019.

[25] Laurent A. F. Callot, Anders B. Kock, and Marcelo C. Medeiros. Modeling and forecasting large realized covariance matrices and portfolio choice. *Journal of Applied Econometrics*, 32(1):140–158, 2017.

[26] John Y Campbell, Andrew W Lo, and A Craig MacKinlay. *The Econometrics of Financial Markets.* Princeton University Press, 1997.

[27] Mehmet Caner and Anders Bredahl Kock. Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative lasso. *Journal of Econometrics*, 203(1):143–168, 2018.

[28] Gary Chamberlain and Michael Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304, 1983.

[29] Yeung Lewis Chan, James H. Stock, and Mark W. Watson. A dynamic factor model framework for forecast combination. *Spanish Economic Review*, 1(2):91–121, Jul 1999.

[30] Jinyuan Chang, Yumou Qiu, Qiwei Yao, and Tao Zou. Confidence regions for entries of a large precision matrix. *Journal of Econometrics*, 206(1):57–82, 2018.

[31] Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

[32] Gerda Claeskens, Jan R Magnus, Andrey L Vasnev, and Wendun Wang. The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762, 2016.

[33] Robert T. Clemen. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4):559 – 583, 1989.

[34] Gregory Connor, Matthias Hagmann, and Oliver Linton. Efficient semiparametric estimation of the fama–french model and extensions. *Econometrica*, 80(2):713–754, 2012.

[35] Gregory Connor and Robert A. Korajczyk. Risk and return in an equilibrium APT: Application of a new test methodology. *Journal of Financial Economics*, 21(2):255–289, 1988.

[36] Gregory Connor and Oliver Linton. Semiparametric estimation of a characteristic-based factor model of common stock returns. *Journal of Empirical Finance*, 14(5):694 – 717, 2007.

[37] Philippe Goulet Coulombe, Maxime Leroux, Dalibor Stevanovic, and Stéphane Surprenant. How is machine learning useful for macroeconomic forecasting? *arXiv:2008.12477*, 2020.

[38] Chaoxing Dai, Kun Lu, and Dacheng Xiu. Knowing factors or factor loadings, or neither? Evaluating estimators of large covariance matrices with noisy and asynchronous data. *Journal of Econometrics*, 208(1):43–79, 2019. Special Issue on Financial Engineering and Risk Management.

[39] Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953, 2009.

[40] Francis X. Diebold. Forecast combination and encompassing: Reconciling two divergent literatures. *International Journal of Forecasting*, 5(4):589 – 592, 1989.

[41] Francis X. Diebold and Peter Pauly. The use of prior information in forecast combination. *International Journal of Forecasting*, 6(4):503 – 508, 1990.

[42] Francis X. Diebold and Minchul Shin. Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. *International Journal of Forecasting*, 2018.

[43] Yi Ding, Yingying Li, and Xinghua Zheng. High dimensional minimum variance portfolio estimation under statistical factor models. *Journal of Econometrics*, 222(1, Part B):502–515, 2021.

[44] Graham Elliott, Antonio Gargano, and Allan Timmermann. Complete subset regressions. *Journal of Econometrics*, 177(2):357–373, 2013. Dynamic Econometric Modeling and Forecasting.

[45] Graham Elliott, Antonio Gargano, and Allan Timmermann. Complete subset regressions with large-dimensional sets of predictors. *Journal of Economic Dynamics and Control*, 54:86–110, 2015.

[46] Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

[47] Eugene F. Fama and Kenneth R. French. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22, 2015.

[48] Jianqing Fan, Yingying Fan, and Jinchi Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186 – 197, 2008.

[49] Jianqing Fan, Yingying Fan, and Jinchi Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147:186–197, 11 2008.

[50] Jianqing Fan, Yang Feng, and Yichao Wu. Network exploration via the adaptive Lasso and SCAD penalties. *The Annals of Applied Statistics*, 3(2):521–541, 06 2009.

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

[52] Jianqing Fan, Yuan Liao, and Martina Mincheva. High-dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39(6):3320–3356, 12 2011.

[53] Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B*, 75(4):603–680, 2013.

[54] Jianqing Fan, Yuan Liao, and Xiaofeng Shi. Risks of large portfolios. *Journal of Econometrics*, 186(2):367–387, 2015. High Dimensional Problems in Econometrics.

[55] Jianqing Fan, Yuan Liao, and Weichen Wang. Projected principal component analysis in factor models. *Ann. Statist.*, 44(1):219–254, 02 2016.

[56] Jianqing Fan, Han Liu, and Weichen Wang. Large covariance estimation through elliptical factor models. *The Annals of Statistics*, 46(4):1383–1414, 08 2018.

[57] Jianqing Fan, Weichen Wang, and Yiqiao Zhong. Robust covariance estimation for approximate factor models. *Journal of Econometrics*, 208(1):5–22, 2019.

[58] Jianqing Fan, Haolei Weng, and Yifeng Zhou. Optimal estimation of functionals of high-dimensional mean and covariance matrix. *arXiv:1908.07460*, 2019.

[59] Jianqing Fan, Lingzhou Xue, and Jiawei Yao. Sufficient forecasting using factor models. *Journal of Econometrics*, 201(2):292–306, 2017.

[60] Jianqing Fan, Jingjin Zhang, and Ke Yu. Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606, 2012. PMID: 23293404.

[61] Yingying Fan and Jinchi Lv. Innovated scalable efficient estimation in ultra-large gaussian graphical models. *The Annals of Statistics*, 44(5):2098–2126, 10 2016.

[62] Yingying Fan and Cheng Yong Tang. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B*, 75(3):531–552, 2013.

[63] Wayne E. Ferson. Tests of multifactor pricing models, volatility bounds and portfolio performance (Chapter 12). In *Financial Markets and Asset Pricing*, volume 1 of *Handbook of the Economics of Finance*, pages 743 – 802. Elsevier, 2003.

[64] Rina Foygel and Mathias Drton. Extended bayesian information criteria for gaussian graphical models. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS, pages 604–612, USA, 2010. Curran Associates Inc.

[65] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the Graphical Lasso. *Biostatistics*, 9(3):432–441, 12 2008.

[66] Márcio G.P. Garcia, Marcelo C. Medeiros, and Gabriel F.R. Vasconcelos. Real-time inflation forecasting with high-dimensional models: The case of Brazil. *International Journal of Forecasting*, 33(3):679 – 693, 2017.

[67] Raffaella Giacomini and Ivana Komunjer. Evaluation and combination of conditional quantile forecasts. *Journal of Business & Economic Statistics*, 23(4):416–431, 2005.

[68] Shingo Goto and Yan Xu. Improving mean variance optimization through sparse hedging restrictions. *Journal of Financial and Quantitative Analysis*, 50(6):1415–1441, 2015.

[69] C. W. J. Granger. Invited review combining forecasts—twenty years later. *Journal of Forecasting*, 8(3):167–173, 1989.

[70] Bruce E. Hansen. Least-squares forecast averaging. *Journal of Econometrics*, 146(2):342–350, 2008.

[71] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[72] Nikolaus Hautsch, Lada M. Kyj, and Roel C. A. Oomen. A blocking and regularization approach to high-dimensional realized covariance estimation. *Journal of Applied Econometrics*, 27(4):625–645, 2012.

[73] David F. Hendry and Michael P. Clements. Pooling of forecasts. *The Econometrics Journal*, 7(1):1–31, 2004.

[74] Eric Hillebrand, Huiyu Huang, Tae-Hwy Lee, and Canlin Li. Using the entire yield curve in forecasting output and inflation. *Econometrics*, 6(3), 2018.

[75] Huiyu Huang and Tae-Hwy Lee. To combine forecasts or to combine information? *Econometric Reviews*, 29(5-6):534–570, 2010.

[76] Jana Janková and Sara van de Geer. Inference in high-dimensional graphical models. *Handbook of Graphical Models*, Chapter 14, pages 325–351. CRC Press, 2018.

[77] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[78] Adel Javanmard and Andrea Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014.

[79] Adel Javanmard, Andrea Montanari, et al. Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622, 2018.

[80] Narasimhan Jegadeesh. Evidence of predictable behavior of security returns. *The Journal of Finance*, 45(3):881–898, 1990.

[81] Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91, 1993.

[82] Iain M Johnstone and Arthur Yu Lu. Sparse principal components analysis. *arXiv:0901.4392*, 2009.

[83] Raymond Kan and Xiaolu Wang. On the economic value of alphas. *Available at SSRN 1785161*, 2019.

[84] Raymond Kan and Guofu Zhou. Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis*, 42(3):621–656, 2007.

[85] George Kapetanios. A testing procedure for determining the number of factors in approximate factor models with large datasets. *Journal of Business & Economic Statistics*, 28(3):397–409, 2010.

[86] Soohun Kim, Robert A. Korajczyk, and Andreas Neuhierl. Arbitrage portfolios. Technical Report 18-43, Georgia Tech Scheller College of Business Research Paper, April 2019.

[87] Yuta Koike. De-biased graphical lasso for high-frequency data. *Entropy*, 22(4):456, 2020.

[88] Philipp Kremer, Sangkyun Lee, Malgorzata Bogdan, and Sandra Paterlini. Sparse portfolio selection via the sorted l1 - norm. *Journal of Banking & Finance*, 110:105687, 11 2019.

[89] Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621, 2003.

[90] Olivier Ledoit and Michael Wolf. Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119, 2004.

[91] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[92] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, 04 2012.

[93] Olivier Ledoit and Michael Wolf. Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *Journal of Multivariate Analysis*, 139:360–384, 2015.

[94] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. *The Review of Financial Studies*, 30(12):4349–4388, 2017.

[95] Olivier Ledoit and Michael Wolf. Nonlinear Shrinkage of the Covariance Matrix for Portfolio Selection: Markowitz Meets Goldilocks. *The Review of Financial Studies*, 30(12):4349–4388, 06 2017.

[96] Tae-Hwy Lee, Millie Yi Mao, and Aman Ullah. Estimation of high-dimensional dynamic conditional precision matrices with an application to forecast combination. *Forthcoming in Econometric Reviews*, 2020.

[97] Tae-Hwy Lee and Ekaterina Seregina. Optimal portfolio using factor graphical lasso. *arXiv:2011.00435*, 2020.

[98] Bruce N. Lehmann. Fads, martingales, and market efficiency. *The Quarterly Journal of Economics*, 105(1):1–28, 1990.

[99] Hongjun Li, Qi Li, and Yutang Shi. Determining the number of factors when the number of factors can increase with sample size. *Journal of Econometrics*, 197(1):76–86, 2017.

[100] Jiahan Li. Sparse and stable portfolio selection with parameter uncertainty. *Journal of Business & Economic Statistics*, 33(3):381–392, 2015.

[101] Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, NIPS'10, pages 1432–1440, USA, 2010. Curran Associates Inc.

[102] Matthew R Lyle and Teri Lombardi Yohn. Fundamental analysis and mean-variance optimal portfolios. *Kelley School of Business Research Paper*, 2020.

[103] A.Craig MacKinlay. Multifactor models do not explain deviations from the capm. *Journal of Financial Economics*, 38(1):3 – 28, 1995.

[104] Ross A Maller and Darrell A Turkington. New light on the portfolio allocation problem. *Mathematical Methods of Operations Research*, 56(3):501–511, 2003.

[105] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[106] Rahul Mazumder and Trevor Hastie. The Graphical Lasso: new insights and alternatives. *Electronic Journal of Statistics*, 6:2125–2149, 2012.

[107] Michael W. McCracken and Serena Ng. FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.

[108] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 06 2006.

[109] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society, Series B*, 72:417–473, 2010.

[110] Tristan Millington and Mahesan Niranjan. Robust portfolio risk minimization using the graphical lasso. In *Neural Information Processing*, pages 863–872, Cham, 2017. Springer International Publishing.

[111] Serena Ng. Chapter 14 - variable selection in predictive regressions. In Graham Elliott and Allan Timmermann, editors, *Handbook of Economic Forecasting*, volume 2 of *Handbook of Economic Forecasting*, pages 752 – 789. Elsevier, 2013.

[112] Alexei Onatski. Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016, 2010.

[113] Alexei Onatski. Discussion on the paper by Fan J., Liao Y., and Mincheva M. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B*, 75(4):650–652, 2013.

[114] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[115] M. Pourahmadi. *High-Dimensional Covariance Estimation: With High-Dimensional Data*. Wiley Series in Probability and Statistics. John Wiley and Sons, 2013, 2013.

[116] Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and B Yu. High-dimensional covariance estimation by minimizing -penalized log-determinant divergence. *Electronic Journal of Statistics*, 5, 01 2011.

[117] Roger, Byoungwook Jang, Yuekai Sun, and Shuheng Zhou. Precision matrix estimation with noisy and missing data. In *AISTATS*, 2019.

[118] Barr Rosenberg. Extra-market components of covariance in security returns. *The Journal of Financial and Quantitative Analysis*, 9(2):263–274, 1974.

[119] Stephen A Ross. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341–360, 1976.

[120] Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[121] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 03 1978.

[122] Ekaterina Seregina. A basket half full: Sparse portfolios. *arXiv preprint arXiv:2011.04278*, 2020.

[123] Jun Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–494, 1993.

[124] Dong X. Shaw, Shucheng Liu, and Leonid Kopman. Lagrangian relaxation procedure for cardinality-constrained portfolio optimization. *Optimization Methods and Software*, 23(3):411–420, 2008.

[125] Jeremy Smith and Kenneth F Wallis. A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355, 2009.

[126] James H. Stock and Mark W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002.

[127] James H. Stock and Mark W. Watson. Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430, 2004.

[128] James H. Stock and Mark W. Watson. Chapter 10 forecasting with many predictors. volume 1 of *Handbook of Economic Forecasting*, pages 515 – 554. Elsevier, 2006.

[129] Mary E. Thomson, Andrew C. Pollock, Dilek Önkal, and M. Sinan Gönül. Combining forecasts: performance and coherence. *International Journal of Forecasting*, 35(2):474–484, 2019.

[130] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[131] Chris Tidmore, Francis M. Kinniry, Giulio Renzi-Ricci, and Edoardo Cilla. How to increase the odds of owning the few stocks that drive returns. *The Journal of Investing*, 2019.

[132] Allan Timmermann. Forecast combinations. *Handbook of Economic Forecasting*, Vol. 1, Chapter 4, pages 135–196. Elsevier, 2006.

[133] J. Tobin. Liquidity preference as behavior towards risk. *The Review of Economic Studies*, 25(2):65–86, 02 1958.

[134] Federico Tomasi, Veronica Tozzo, Saverio Salzo, and Alessandro Verri. Latent variable time-varying network inference. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2338–2346, 2018.

[135] Sara van de Geer, Peter Buhlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 06 2014.

[136] Lieven Vandenberghe, Stephen Boyd, and Shao-Po Wu. Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications*, 19(2):499–533, April 1998.

[137] Thomas Verma and Judea Pearl. Causal networks: Semantics and expressiveness. In *Uncertainty in Artificial Intelligence*, volume 9 of *Machine Intelligence and Pattern Recognition*, pages 69–76. North-Holland, 1990.

[138] Ivan Vujačić, Antonino Abbruzzo, and Ernst Wit. A computationally fast alternative to cross-validation in penalized gaussian graphical models. *Journal of Statistical Computation and Simulation*, 85(18):3628–3640, 2015.

[139] John Wishart. The generalized product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A(1-2):32–52, 12 1928.

[140] Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 03 2007.

[141] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[142] Tuo Zhao, Han Liu, Kathryn Roeder, John Lafferty, and Larry Wasserman. The HUGE package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*, 13(1):1059–1062, April 2012.

[143] Yunan Zhu and Ivor Cribben. Sparse graphical models for functional connectivity networks: Best methods and the autocorrelation issue. *Brain Connectivity*, 8(3):139–165, 2018. PMID: 29634321.