Optimal Portfolio Using Factor Graphical Lasso

Tae-Hwy Lee[*]   and   Ekaterina Seregina[†]

First version: September 8, 2020
Second version: November 5, 2020[‡]

Abstract

Graphical models are a powerful tool to estimate a high-dimensional inverse covariance (*precision*) matrix, which has been applied for portfolio allocation problem. The assumption made by these models is a sparsity of the precision matrix. However, when the stock returns are driven by the common factors, this assumption does not hold. Our paper develops a framework for estimating a high-dimensional precision matrix which combines the benefits of exploring the factor structure of the stock returns and the sparsity of the precision matrix of the factor-adjusted returns. The proposed algorithm is called *Factor Graphical Lasso* (FGL). We study a high-dimensional portfolio allocation problem when the asset returns admit the approximate factor model. In high dimensions, when the number of assets is large relative to the sample size, the sample covariance matrix of the excess returns is subject to the large estimation uncertainty, which leads to unstable solutions for portfolio weights. To resolve this issue, we consider the decomposition of low-rank and sparse components. This strategy allows us to consistently estimate the optimal portfolio in high dimensions, even when the covariance matrix is ill-behaved. We establish consistency of the portfolio weights in a high-dimensional setting *without assuming sparsity on the covariance or precision matrix of stock returns*. Our theoretical results and simulations demonstrate that FGL is robust to heavy-tailed distributions, which makes our method suitable for financial applications. The empirical application uses daily and monthly data for the constituents of the S&P500 to demonstrate superior performance of FGL compared to the equal-weighted portfolio, index and some prominent precision and covariance-based estimators.

Keywords: High-dimensionality, Portfolio optimization, Graphical Lasso, Approximate Factor Model, Sharpe Ratio, Elliptical Distributions

JEL Classifications: C13, C55, C58, G11, G17

---

[*]Department of Economics, University of California, Riverside. Email: tae.lee@ucr.edu.

[†]Department of Economics, University of California, Riverside. Email: ekaterina.seregina@email.ucr.edu.

[‡]In the second version, a part of Theorem 4 is corrected, with Lemma 11(e) added. Figure 1 is revised.

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 (Cai et al. (2011); Friedman et al. (2008); Meinshausen and Bühlmann (2006)). Goto and Xu (2015) 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. Awoye (2016) used Graphical Lasso (Friedman et al. (2008)) to estimate a sparse covariance matrix for the Markowitz mean-variance portfolio problem to improve covariance estimation in terms of lower realized portfolio risk. Millington and Niranjan (2017) 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. Millington and Niranjan (2017) 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. Callot et al. (2019) use the nodewise-regression method of Meinshausen and Bühlmann (2006) 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) (Fan et al. (2013)) and linear shrinkage (Ledoit and Wolf (2004)). Cai et al. (2020) use constrained ℓ_1-minimization for inverse matrix estimation (CLIME) of the precision matrix (Cai et al. (2011)) 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 Campbell et al. (1997) 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 (Fan et al. (2011, 2013, 2016b, 2018)). 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 Fan et al. (2016a) and Ait-Sahalia and Xiu (2017), but they did not study a high-dimensional case. *Factor models are generally treated as competitors to graphical models*: as an example, Callot et al. (2019) find evidence of superior performance of nodewise-regression estimator of precision matrix over a factor-based estimator POET (Fan et al. (2013)) 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 Koike (2020), *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.*

In this paper we develop a new 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 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, portfolios based on other estimators of precision matrix (CLIME, Cai et al. (2011)) and covariance matrix (POET, Fan et al. (2013) and the shrinkage estimator adjusted to allow for the factor structure (Ledoit and Wolf (2004))) 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 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 (Fan et al. (2013)) 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.

This paper is organized as follows: Section 2 reviews the basics of the Markowitz mean-variance portfolio theory and provides several formulations of the optimal portfolio allocation. 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, $\mathcal{S}_p^+$ denotes the set of all $p \times p$ positive semi-definite matrices, and $\mathcal{S}_p^{++}$ denotes the set of all $p \times p$ positive definite matrices. 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 $\|\mathbf{U}\|_{a,b}$ denote the induced matrix-operator norm $\max_{\|\mathbf{y}\|_a = 1} \|\mathbf{U}\mathbf{y}\|_b$. The special cases are $\|\mathbf{U}\|_1 \equiv \max_{1 \leq j \leq N} \sum_{i=1}^N |\mathbf{U}_{i,j}|$ 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 $\mathbf{U}$; $\|\mathbf{U}\|_\infty \equiv \max_{1 \leq j \leq N} \sum_{i=1}^N |\mathbf{U}_{j,i}|$ for the ℓ_∞/ℓ_∞-operator norm. Finally, $\|\mathbf{U}\|_{\max}$ denotes the element-wise maximum $\max_{i,j} |\mathbf{U}_{i,j}|$, and $\|\mathbf{U}\|_F^2 = \sum_{i,j} u_{i,j}^2$ denotes the Frobenius matrix norm.

2 Optimal Portfolio Allocation

The importance of the minimum-variance portfolio introduced by Markowitz (1952) 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 $\mathbf{r}_t = (r_{1t}, r_{2t}, \ldots, r_{pt})' \sim \mathcal{D}(\mathbf{m}, \boldsymbol{\Sigma})$ be a $p \times 1$ vector of *excess* returns drawn from a distribution $\mathcal{D}$. 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.[1] This can be formulated as the following quadratic optimization problem:

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

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 $\boldsymbol{\Theta} \equiv \boldsymbol{\Sigma}^{-1}$ be the *precision matrix*.

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

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

---

[1]If, in addition to the constraint that weights sum up to unity, short-sales are not allowed, then the combination of portfolio weights forms a convex hull. We do not impose any short-selling constraints in this paper.

If $\mathbf{m}'\mathbf{w} = \mu$, the solution to (2.1) is a well-known two-fund separation theorem introduced by Tobin (1958):

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

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

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

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:

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

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

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

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:

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

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

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

To get the second equality in (2.9) we use the definition of μ from (2.1) and (2.6). It follows that if $\mu = \sigma\sqrt{\theta}$, where $\theta \equiv \mathbf{m}'\boldsymbol{\Theta}\mathbf{m}$ is the squared Sharpe Ratio of the portfolio, then the solution to (2.6) and (2.8) admits the following expression:

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

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 $\boldsymbol{\Theta}$. In this paper we develop a novel method for estimating precision matrix for the above-mentioned financial portfolios which account for the fact that the returns follow approximate factor structure. The next section reviews Graphical methods for estimating the precision matrix, and introduces a *Factor Graphical Lasso* for constructing financial portfolios.

3   Factor Graphical Model

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 propose an estimator of the precision matrix which accounts for the common factors in the excess returns.

The review of the Gaussian graphical models is based on Hastie et al. (2001) and Bishop (2006). 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.

3.1   Using Penalized Bregman Divergence to Estimate Precision Matrix

Define $\mathbf{x}_t$ to be a $p \times 1$ vector at time $t = 1, \ldots, T$. Let $\mathbf{x}_t \sim \mathcal{D}(\mathbf{m}, \boldsymbol{\Sigma})$, where $\mathcal{D}$ belongs to either sub-Gaussian or elliptical families. When $\mathcal{D} = \mathcal{N}$, the precision matrix $\boldsymbol{\Sigma}^{-1} \equiv \boldsymbol{\Theta}$ contains information about conditional dependence between the variables. For instance, if $\boldsymbol{\Theta}_{ij}$, 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.

Given a sample $\{\mathbf{x}_t\}_{t=1}^T$, let $\mathbf{S} = (1/T) \sum_{t=1}^T (\mathbf{x}_t - \bar{\mathbf{x}}_t)(\mathbf{x}_t - \bar{\mathbf{x}}_t)'$ denote the sample covariance matrix. We can write down the Gaussian log-likelihood (up to constants): $l(\boldsymbol{\Theta}) = \log \det(\boldsymbol{\Theta}) -$

trace($\mathbf{S\Theta}$). The maximum likelihood (ML) estimate 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.

One of the approaches 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. Let $\widehat{\mathbf{D}}^2 \equiv \mathrm{diag}(\mathbf{S})$. Janková and van de Geer (2018) propose to use the *weighted Graphical Lasso* to maximize the following *weighted penalized log-likelihood*:

$$\widehat{\mathbf{\Theta}} = \arg \min_{\mathbf{\Theta} \in \mathcal{S}_p^{++}} \mathrm{trace}(\mathbf{S\Theta}) - \log \det(\mathbf{\Theta}) + \lambda \sum_{i \neq j} \widehat{\mathbf{D}}_{ii} \widehat{\mathbf{D}}_{jj} |\mathbf{\Theta}_{ij}|, \tag{3.1}$$

over symmetric positive definite matrices, where $\lambda \geq 0$ is a penalty parameter. When $\lambda = 0$, the MLE for $\mathbf{\Sigma}$ and $\mathbf{\Theta}$ in (3.1) are the sample covariance matrix $\mathbf{S}$ and its inverse $\mathbf{S}^{-1}$ respectively. When $\lambda > 0$, the solution to (3.1) yields *penalized MLE* of the covariance and precision matrices, denoted as $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{\Sigma}}^{-1}$. Ravikumar et al. (2011) showed that solving $\min_{\mathbf{\Theta} \in \mathcal{S}_p^{++}} \mathrm{trace}(\widehat{\mathbf{S}}\mathbf{\Theta}) - \log \det(\mathbf{\Theta}) + \sum_{i=1}^{p} \sum_{j=1}^{p} p_\lambda(|\mathbf{\Theta}_{ij}|)$, where $p(\cdot)$ is a generic penalty function, corresponds to minimizing penalized log-determinant Bregman divergence.

One of the most popular and fast algorithms to solve the optimization problem in (3.1) is called the Graphical Lasso (GLasso), it was introduced by Friedman et al. (2008). Graphical Lasso combines the neighborhood method by Meinshausen and Bühlmann (2006) and block-coordinate descent by Banerjee et al. (2008). A brief summary of the procedure to estimate the precision matrix using GLasso is presented in Algorithm 1.

Algorithm 1 Graphical Lasso (Friedman et al. (2008))

1: Let $\mathbf{W}$ be the estimate of $\mathbf{\Sigma}$. Initialize $\mathbf{W} = \mathbf{S} + \lambda \mathbf{I}$. The diagonal of $\mathbf{W}$ remains the same in what follows.

2: 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 $\widehat{\mathbf{w}}_{12} = \mathbf{W}_{11}\widehat{\boldsymbol{\beta}}$.

3: In the final cycle (for $i = 1, \ldots, p$) solve for $\frac{1}{\widehat{\theta}_{22}} = w_{22} - \widehat{\boldsymbol{\beta}}' \widehat{\mathbf{w}}_{12}$ and $\widehat{\boldsymbol{\theta}}_{12} = -\widehat{\theta}_{22}\widehat{\boldsymbol{\beta}}$.

3.2 Factor Graphical Lasso

The arbitrage pricing theory (APT), developed by Ross (1976), 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. Let $\mathbf{r}_t = (r_{1t}, r_{2t}, \ldots, r_{pt})' \sim \mathcal{D}(\mathbf{m}, \boldsymbol{\Sigma})$ be a $p \times 1$ vector of excess returns drawn from a distribution $\mathcal{D}$, where $\mathbf{m}$ is the unconditional mean of the returns. Assume that the return generating process $(\mathbf{r}_t)$ follows a K-factor model:

$$\underbrace{\mathbf{r}_t}_{p \times 1} = \mathbf{B} \underbrace{\mathbf{f}_t}_{K \times 1} + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T \tag{3.2}$$

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 (3.2) can be either observable, such as in Fama and French (1993, 2015), or can be estimated using statistical factor models. Unobservable factors and loadings are usually estimated by the principal component analysis (PCA), as studied in Bai (2003); Bai and Ng (2002); Connor and Korajczyk (1988); Stock and Watson (2002). 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 Bai (2003); Chamberlain and Rothschild (1983) 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 precision matrix and portfolio weights 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 Fan et al. (2013): 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 (3.2) in matrix form:

$$\underbrace{\mathbf{R}}_{p \times T} = \underbrace{\mathbf{B}}_{p \times K} \mathbf{F} + \mathbf{E}. \tag{3.3}$$

Recall that the factors and loadings in (3.3) are estimated by solving the following minimization problem: $(\widehat{\mathbf{B}}, \widehat{\mathbf{F}}) = \mathrm{argmin}_{\mathbf{B},\mathbf{F}} \|\mathbf{R} - \mathbf{B}\mathbf{F}\|_F^2$ 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 (Fan et al. (2018)). It was shown (Stock and Watson (2002)) that $\widehat{\mathbf{F}} = \sqrt{T}\mathrm{eig}_K(\mathbf{R}'\mathbf{R})$ and $\widehat{\mathbf{B}} = T^{-1}\mathbf{R}\widehat{\mathbf{F}}'$. Given $\widehat{\mathbf{F}}, \widehat{\mathbf{B}}$, define $\widehat{\mathbf{E}} = \mathbf{R} - \widehat{\mathbf{B}}\widehat{\mathbf{F}}$.

Having introduced the generating process for stock returns, we move to portfolio construction exercise. 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 Koike (2020), *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 Barigozzi et al. (2018); Brownlees et al. (2018); Koike (2020)).

We use the weighted Graphical Lasso as a shrinkage technique to estimate the precision matrix $\mathbf{\Theta}_\varepsilon$ of the idiosyncratic errors. Once the precision $\mathbf{\Theta}_f$ of the low-rank component is also obtained, similarly to Fan et al. (2011), we use the Sherman-Morrison-Woodbury formula to estimate the precision of excess returns:

$$\mathbf{\Theta} = \mathbf{\Theta}_\varepsilon - \mathbf{\Theta}_\varepsilon \mathbf{B}[\mathbf{\Theta}_f + \mathbf{B}'\mathbf{\Theta}_\varepsilon \mathbf{B}]^{-1}\mathbf{B}'\mathbf{\Theta}_\varepsilon. \tag{3.4}$$

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 first use the weighted GLasso Algorithm 1, with the initial estimate of the covariance matrix of the idiosyncratic errors calculated as $\widehat{\mathbf{\Sigma}}_\varepsilon = T^{-1}\widehat{\mathbf{E}}\widehat{\mathbf{E}}'$. Once we have estimated $\widehat{\mathbf{\Theta}}_f$ and $\widehat{\mathbf{\Theta}}_\varepsilon$, we can get $\widehat{\mathbf{\Theta}}$ using a sample analogue of (3.4).

We call the proposed procedure *Factor Graphical Lasso* and summarize it in Algorithm 2.

Algorithm 2 Factor Graphical Lasso (FGL)

1: Estimate the residuals: $\widehat{\varepsilon}_t = \mathbf{r}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t$ using PCA.
  Get $\widehat{\mathbf{\Sigma}}_\varepsilon = \frac{1}{T}\sum_{t=1}^{T}(\widehat{\varepsilon}_t - \bar{\varepsilon})(\widehat{\varepsilon}_t - \bar{\varepsilon})'$.
2: Estimate a sparse $\mathbf{\Theta}_\varepsilon$ using the weighted Graphical Lasso: initialize Algorithm 1 with $\mathbf{W} = \widehat{\mathbf{\Sigma}}_\varepsilon + \lambda\mathbf{I}$.
3: Estimate $\mathbf{\Theta}$ using the Sherman-Morrison-Woodbury formula in (3.4).

Now we can use $\widehat{\boldsymbol{\Theta}}$ obtained from (3.4) using 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 (Bai and Ng (2002); Kapetanios (2010)). As an example, in their paper Fan et al. (2013) adopt the approach from Bai and Ng (2002).*

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 Li et al. (2017): following their notation, let

$$V(K) = \min_{\mathbf{B}_K, \mathbf{F}_K} \frac{1}{pT} \sum_{i=1}^{p} \sum_{t=1}^{T} \left(r_{it} - \frac{1}{\sqrt{K}} \mathbf{b}'_{i,K} \mathbf{f}_{t,K} \right)^2, \tag{3.5}$$

where the minimum is taken over $1 \le K \le K_{max}$, subject to normalization $\mathbf{B}'_K \mathbf{B}_K / p = \mathbf{I}_K$. Hence, $\bar{\mathbf{F}}'_K = \sqrt{K} \mathbf{R}' \bar{\mathbf{B}}_K / p$. Define $\bar{\mathbf{F}}'_K = \bar{\mathbf{F}}'_K (\bar{\mathbf{F}}_K \bar{\mathbf{F}}'_K / T)^{1/2}$, 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 Li et al. (2017) to estimate K:

$$\widehat{K} = \arg \min_{1 \le K \le K_{max}} \ln(V(K, \bar{\mathbf{F}}_K)) + K g(p, T), \tag{3.6}$$

where $1 \le K \le K_{max} = o(\min\{p^{1/17}, T^{1/16}\})$ and $g(p, T)$ is a penalty function of (p, T) such that (i) $K_{max} \cdot g(p, T) \to 0$ and (ii) $C_{p,T,K_{max}}^{-1} \cdot g(p, T) \to \infty$ with $C_{p,T,K_{max}} = \mathcal{O}_p\left(\max \left[\frac{K_{max}^3}{\sqrt{p}}, \frac{K_{max}^{5/2}}{\sqrt{T}} \right] \right)$. The choice of the penalty functions is similar to Bai and Ng (2002). Throughout the paper we let $\widehat{K}$ be the solution to (3.6).

4   Asymptotic Properties

In this section we establish consistency of the Factor Graphical Lasso in Algorithm 2. After that we 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.

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

$$D_j(A) \equiv \{i : A_{ij} \ne 0, \ i \ne j\}, \quad d_j(A) \equiv \text{card}(D_j(A)), \quad d(A) \equiv \max_{j=1,\ldots,p} d_j(A), \tag{4.1}$$

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.

4.1 Assumptions

We now list the assumptions on the model (3.2):

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

(A.2) (Pervasive factors) There exists a positive definite $K \times K$ matrix $\check{\mathbf{B}}$ such that $\left\| \left\| p^{-1}\mathbf{B}'\mathbf{B} - \check{\mathbf{B}} \right\| \right\|_2 \to 0$ and $\Lambda_{\min}(\check{\mathbf{B}})^{-1} = \mathcal{O}(1)$ 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 $\Lambda_{min}(\boldsymbol{\Sigma}_\varepsilon) > c_1$, $\left\| \left\| \boldsymbol{\Sigma}_\varepsilon \right\| \right\|_1 < c_2$ and $\min_{i \leq p, j \leq p} \mathrm{var}(\varepsilon_{it}\varepsilon_{jt}) > c_1$.

(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 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

$$\alpha(T) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_T^\infty} |\Pr A \Pr B - \Pr AB|. \tag{4.2}$$

(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

11

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

Some comments regarding the aforementioned assumptions are in order. Assumptions **(A.1)**-**(A.4)** are the same as in Fan et al. (2013), 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 Fan et al. (2018), **(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. Assumption **(A.3)**(a) is slightly stronger than in Bai (2003), 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 Fan et al. (2011); Koike (2020), 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, at the end of section 4 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 Bai (2003) since our goal is to estimate a precision matrix, and **(A.5)**(c) differs from Bai (2003) and Bai and Ng (2006) in that the number of factors is assumed to slowly grow with p.

In addition, the following structural assumptions on the model are imposed:

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

(B.2) $s(\boldsymbol{\Theta}_\varepsilon) = \mathcal{O}_p(s_T)$ for some sequence $s_T \in (0, \infty)$, $T = 1, 2, \ldots$.

(B.3) $d(\boldsymbol{\Theta}_\varepsilon) = \mathcal{O}_p(d_T)$ for some sequence $d_T \in (0, \infty)$, $T = 1, 2, \ldots$.

(B.1) is a natural structural assumption on the population quantities, and **(B.2)**-**(B.3)** are sparsity assumptions on the precision matrix of the residual process. Specifically, **(B.2)** states that the sparsity of $\mathbf{\Theta}_\varepsilon$ is controlled by the deterministic sequence s_T; we will impose restrictions on the growth rate of s_T. **(B.3)** is another sparsity assumption on $\mathbf{\Theta}_\varepsilon$: it is weaker than **(B.2)** since it is always satisfied when $s_T = d_T$. However, d_T can generally be smaller than s_T. Note that in contrast to Fan et al. (2013) we do not impose sparsity on the covariance matrix of the idiosyncratic component, instead, it is more realistic to impose conditional sparsity on the precision matrix after the common factors are accounted for.

4.2 The FGL Procedure

Recall the definition of the Weighted Graphical Lasso estimator in (3.1) for the precision matrix of the idiosyncratic components:

$$\widehat{\mathbf{\Theta}}_\varepsilon = \arg \min_{\mathbf{\Theta} \in \mathcal{S}_p^{++}} \operatorname{trace}(\widehat{\mathbf{\Sigma}}_\varepsilon \mathbf{\Theta}_\varepsilon) - \log \det(\mathbf{\Theta}_\varepsilon) + \lambda \sum_{i \neq j} \widehat{\mathbf{D}}_{\varepsilon,ii} \widehat{\mathbf{D}}_{\varepsilon,jj} |\mathbf{\Theta}_{\varepsilon,ij}|, \tag{4.3}$$

Also, recall that to estimate $\mathbf{\Theta}$ we used equation (3.4). 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 (4.3). **(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 $\widehat{\mathbf{\Theta}}$.

Subsection 4.3 examines the theoretical foundations of the first step, and Subsection 4.4 is devoted to steps 2 and 3.

4.3 Convergence of Unknown Factors and Loadings

As pointed out in Bai (2003) and Fan et al. (2013), $K \times 1$-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 Li et al. (2017) 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 $\mathbf{H} = (1/T)\mathbf{V}^{-1}\widehat{\mathbf{F}}'\mathbf{FB}'\mathbf{B}$. 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 Fan et al. (2013) for the case when the number of factors is unknown and is allowed to grow.

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) 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 $\max_{i \leq p}\left\|\widehat{\mathbf{b}}_i - \mathbf{Hb}_i\right\| = \mathcal{O}_p(\omega_{1T})$ and $\max_{t \leq T}\left\|\widehat{\mathbf{f}}_t - \mathbf{Hf}_t\right\| = \mathcal{O}_p(\omega_{2T})$.*

The conditions $K^3 \log(p) = o(T^{\gamma/6})$, $KT = o(p^2)$ are similar to Fan et al. (2013), 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, $\widehat{\boldsymbol{\Sigma}}_\varepsilon$, and precision matrix $\widehat{\boldsymbol{\Theta}}_\varepsilon$ 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, the estimator $\widehat{\boldsymbol{\Sigma}}_\varepsilon$ obtained by estimating factor model in (3.3) satisfies $\left\|\widehat{\boldsymbol{\Sigma}}_\varepsilon - \boldsymbol{\Sigma}_\varepsilon\right\|_{max} = \mathcal{O}_p(\omega_{3T})$.*
We additionally assume (B.1)-(B.2). Let λ_T be a sequence of positive-valued random variables such that $\lambda_T^{-1}\omega_{3T} \xrightarrow{p} 0$. If $s_T\lambda_T \xrightarrow{p} 0$, then $\lambda_T^{-1}\left\|\widehat{\boldsymbol{\Theta}}_\varepsilon - \boldsymbol{\Theta}_\varepsilon\right\|_l = \mathcal{O}_p(s_T)$ 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: Fan et al. (2011) obtained a similar rate but for the case when factors are observable (in their work, $\omega_{2T} = K^{1/2}\sqrt{\log p/T}$). The second part of Theorem 2 is based on the relationship between the convergence rates of the estimated covariance and precision matrices established in Janková and van de Geer (2018) (Theorem 14.1.3). Koike (2020) 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 $\lambda_T^{-1}\sqrt{K \log p/T} \xrightarrow{p} 0$). We

now comment on the optimality of the rate in Theorem 2: as pointed out in Koike (2020), in the standard Gaussian setting without factor structure, the minimax optimal rate is $d(\boldsymbol{\Theta}_\varepsilon)\sqrt{\log p/T}$, which can be faster than the rate obtained in Theorem 2 if $d(\boldsymbol{\Theta}_\varepsilon) < s_T$. 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.

4.4 Convergence of Precision Matrix Estimator and Portfolio Weights by FGL

Having established the convergence properties of $\widehat{\boldsymbol{\Sigma}}_\varepsilon$ and $\widehat{\boldsymbol{\Theta}}_\varepsilon$, we now move to the estimation of the precision matrix of the factor-adjusted returns in equation (3.4).

Theorem 3. *Under the assumptions of Theorem 2, we additionally assume (B.3). If $d_T s_T \lambda_T \xrightarrow{p} 0$, then $\lambda_T^{-1}\left\|\!\left\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right\|\!\right\|_2 = \mathcal{O}_p(s_T + 1/(p\sqrt{K}))$ and $\lambda_T^{-1}\left\|\!\left\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right\|\!\right\|_1 = \mathcal{O}_p(d_T K^{3/2}(s_T + (1/p)))$.*

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

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

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

We now comment on the rates in Theorem 4: first, the rates obtained by Callot et al. (2019) for GMV and MWC formulations, when no factor structure of stock returns is assumed, require $s(\boldsymbol{\Theta})^{3/2}\sqrt{\log(p)/T}$, where the authors imposed sparsity on the precision matrix of stock returns, $\boldsymbol{\Theta}$. 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 $d_T^2 s_T K^3 \sqrt{\log(p)/T} = o(1)$ 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 $\boldsymbol{\Sigma}$ or $\boldsymbol{\Theta}$. 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.

4.5 Implications on Portfolio Risk, Return, and Sharpe Ratio

Having examined the properties of portfolio weights, it is natural to comment on the portfolio risk estimation error. It is determined by the errors in the two components: the estimated covariance matrix and the estimated portfolio weights. We focus on the effect of the second component and compare portfolio risk estimation error for three alternative portfolio formulations. First, we note that for any estimators of covariance matrix and portfolio weights, we have:

$$\left| \widehat{\mathbf{w}}' \widehat{\boldsymbol{\Sigma}} \widehat{\mathbf{w}} - \mathbf{w}' \widehat{\boldsymbol{\Sigma}} \mathbf{w} \right| \le \| \widehat{\mathbf{w}} - \mathbf{w} \|_1 \left\| \widehat{\boldsymbol{\Sigma}} \mathbf{w} \right\|_{\max}. \tag{4.4}$$

The estimation error in portfolio risk is bounded by the estimation error in portfolio weights. Hence, combining equation (4.4) and Theorem 4, we conclude that FGL consistently estimates risk of a financial portfolio. 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.

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 $(\mathbf{A.3})(c)$. Since this tail-behavior may be too restrictive for financial portfolio, we comment on the possibility to relax it. First, recall where $(\mathbf{A.3})(c)$ 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 $\widehat{\boldsymbol{\Sigma}}_\varepsilon$ in Theorem 2. Hence, when Assumption $(\mathbf{A.3})(c)$ is relaxed, one needs to find another way to consistently estimate $\boldsymbol{\Sigma}_\varepsilon$. We achieve it using the tools developed in Fan et al. (2018). Specifically, let $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}_p \boldsymbol{\Lambda}_p \boldsymbol{\Gamma}_p'$,

where $\boldsymbol{\Sigma}$ is the covariance matrix of returns that follow a factor structure described in equation (3.2). Define $\widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\Lambda}}_K, \widehat{\boldsymbol{\Gamma}}_K$ to be the estimators of $\boldsymbol{\Sigma}, \boldsymbol{\Lambda}_p, \boldsymbol{\Gamma}_p$. We further let $\widehat{\boldsymbol{\Lambda}}_K = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_K)$ and $\widehat{\boldsymbol{\Gamma}}_K = (\hat{v}_1, \ldots, \hat{v}_K)$ to be constructed by the first K leading empirical eigenvalues and the corresponding eigenvectors of $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\mathbf{B}}\widehat{\mathbf{B}}' = \widehat{\boldsymbol{\Gamma}}_K \widehat{\boldsymbol{\Lambda}}_K \widehat{\boldsymbol{\Gamma}}'_K$. Similarly to Fan et al. (2018), we require the following bounds on the componentwise maximums of the estimators:

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

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

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

Let $\widehat{\boldsymbol{\Sigma}}^{SG}$ be the sample covariance matrix, with $\widehat{\boldsymbol{\Lambda}}_K^{SG}$ and $\widehat{\boldsymbol{\Gamma}}_K^{SG}$ constructed with the first K leading empirical eigenvalues and eigenvectors of $\widehat{\boldsymbol{\Sigma}}^{SG}$ respectively. Also, let $\widehat{\boldsymbol{\Sigma}}^{EL1} = \widehat{\mathbf{D}}\widehat{\mathbf{R}}_1\widehat{\mathbf{D}}$, 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 $\widehat{\boldsymbol{\Sigma}}^{EL2} = \widehat{\mathbf{D}}\widehat{\mathbf{R}}_2\widehat{\mathbf{D}}$, where $\widehat{\mathbf{R}}_2$ is obtained using the spatial Kendall's tau estimator. Define $\widehat{\boldsymbol{\Lambda}}_K^{EL}$ to be the matrix of the first K leading empirical eigenvalues of $\widehat{\boldsymbol{\Sigma}}^{EL1}$, and $\widehat{\boldsymbol{\Gamma}}_K^{EL}$ is the matrix of the first K leading empirical eigenvectors of $\widehat{\boldsymbol{\Sigma}}^{EL2}$. For more details regarding constructing $\widehat{\boldsymbol{\Sigma}}^{SG}$, $\widehat{\boldsymbol{\Sigma}}^{EL1}$ and $\widehat{\boldsymbol{\Sigma}}^{EL2}$ see Fan et al. (2018), Sections 3 and 4.

Proposition 1. *For sub-Gaussian distributions, $\widehat{\boldsymbol{\Sigma}}^{SG}$, $\widehat{\boldsymbol{\Lambda}}_K^{SG}$ and $\widehat{\boldsymbol{\Gamma}}_K^{SG}$ satisfy (C.1)-(C.3). For elliptical distributions, $\widehat{\boldsymbol{\Sigma}}^{EL1}$, $\widehat{\boldsymbol{\Lambda}}_K^{EL}$ and $\widehat{\boldsymbol{\Gamma}}_K^{EL}$ satisfy (C.1)-(C.3). When (C.1)-(C.3) are satisfied, the bounds obtained in Theorems 2-4 continue to hold.*

Proposition 1 is essentially a rephrasing of the results obtained in Fan et al. (2018), 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, $\widehat{\boldsymbol{\Sigma}}^{EL2}$ is only used for estimating the eigenvectors. This is necessary due to the fact that, in contrast with $\widehat{\boldsymbol{\Sigma}}^{EL2}$, the theoretical properties of the eigenvectors of $\widehat{\boldsymbol{\Sigma}}^{EL}$ are mathematically involved because of the sin function. The FGL for the elliptical distributions will be called the Robust FGL.

5　Monte Carlo

In order to validate our theoretical results, we perform several simulation studies which are divided into three parts. The first set of results computes the empirical convergence rates and compares them with the theoretical expressions derived in Theorems 3-4. The second set of results compares the performance of the FGL with several alternative models for estimating covariance and precision matrix: linear shrinkage estimator of covariance that incorporates factor structure through the Sherman-Morrison inversion formula (Ledoit and Wolf (2004), further referred to as LW), POET (Fan et al. (2013)), CLIME (Cai et al. (2011)) and the standard Graphical Lasso without incorporating factor structure. 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. All three exercises use 100 Monte Carlo simulations.

We first consider the following low-dimensional setup: let $p = T^\delta$, $\delta = 0.85$, $K = 2(\log(T))^{0.5}$ and $T = [2^h]$, for $h = 7, 7.5, 8, \ldots, 9.5$. 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 $\mathbf{A}$ which is used to represent the structure of the graph:

$$\mathbf{A}^{ij} = \begin{cases} 1, & \text{for } i \neq j \ \text{ with probability } q, \\ 0, & \text{otherwise.} \end{cases} \tag{5.1}$$

Let $\mathbf{A}_{\varepsilon,ij}$ denote the i,j-th element of the adjacency matrix $\mathbf{A}_\varepsilon$. We set $\mathbf{A}_{\varepsilon,ij} = \mathbf{A}_{\varepsilon,ji} = 1$, 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 Zhao et al. (2012): using their notation, $\boldsymbol{\Theta}_\varepsilon = \mathbf{A} \cdot v + \mathbf{I}(|\tau| + 0.1 + u)$, 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 $\mathbf{A} \cdot v$.[2] In our simulations we use $u = 0.1$ and $v = 0.3$.

Factors are assumed to have the following structure:

$$\mathbf{f}_t = \phi_f \mathbf{f}_{t-1} + \boldsymbol{\zeta}_t \tag{5.2}$$

$$\underbrace{\mathbf{r}_t}_{p \times 1} = \mathbf{B} \underbrace{\mathbf{f}_t}_{K \times 1} + \varepsilon_t, \quad t = 1, \ldots, T \tag{5.3}$$

---

[2]See Zhao et al. (2012) for further details on the generation process.

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 (5.3) 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 (5.3) leads to the low-rank plus sparse decomposition of the covariance matrix of stock returns $\mathbf{r}_t$.

To compare the empirical rate with the theoretical expressions derived in Theorems 3-4, we use the facts from Theorem 2 that $\omega_{3T} \equiv K^2\sqrt{\log p/T} + K^3/\sqrt{p}$ and $\lambda_T^{-1}\omega_{3T} \xrightarrow{p} 0$ to introduce the following functions that correspond to the theoretical rates for the choice of parameters in the empirical setting:

$$f_{\|\cdot\|_2} = C_1 + C_2 \cdot \log_2((\delta \cdot x)^{0.5 \cdot 2}\sqrt{\delta \cdot x/2^x} + (\delta \cdot x)^{0.5 \cdot 3}/(2^x)^{\delta/2}) + 0.05 \cdot x, \tag{5.4}$$

$$g_{\|\cdot\|_1} = C_3 + C_2 \cdot \log_2((\delta \cdot x)^{0.5 \cdot 2}\sqrt{\delta \cdot x/2^x} + (\delta \cdot x)^{0.5 \cdot 3}/(2^x)^{\delta/2}) + 0.1 \cdot x + 1.5\log_2 x, \tag{5.5}$$

$$h_1 = C_4 + C_2 \cdot \log_2((\delta \cdot x)^{0.5 \cdot 2}\sqrt{\delta \cdot x/2^x} + (\delta \cdot x)^{0.5 \cdot 3}/(2^x)^{\delta/2}) + 0.15 \cdot x + 3\log_2 x, \tag{5.6}$$

$$h_2 = C_5 + C_6 \cdot \log_2((\delta \cdot x)^{0.5 \cdot 2}\sqrt{\delta \cdot x/2^x} + (\delta \cdot x)^{0.5 \cdot 3}/(2^x)^{\delta/2}) + 0.15 \cdot x + 3\log_2 x. \tag{5.7}$$

where $C_1, \ldots, C_6$ are constants with $C_6 < C_2$ (by Theorem 4), and $x = \log_2 T$.

Figure 1 shows the averaged (over Monte Carlo simulations) errors of the estimators of the precision matrix $\boldsymbol{\Theta}$ and the portfolio weights versus the sample size T in the logarithmic scale (base 2). In order to confirm the theoretical findings from Theorems 3-4, we also plot the theoretical rates of convergence given by the functions in (5.4)-(5.7). Figure 1 verifies that the empirical and theoretical rates are matched. Since the convergence rates for GMV and MWC portfolio weights are very similar, we only report the former. Note that as predicted by Theorem 3, the rate of convergence of the precision matrix in $\|\cdot\|_2$-norm is faster than the rate in $\|\cdot\|_1$-norm. Furthermore, the convergence rate of the GMV, MWC and MRC portfolio weights are close to the rate of the precision matrix $\boldsymbol{\Theta}$ in $\|\cdot\|_1$-norm, which is confirmed by Theorem 4. Finally, as evidenced by Figure 1, the convergence rate of the MRC portfolio is faster than the rate of GMV and MWC. This finding is in accordance with Theorem 4.

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 ("low-dimensional"): $p = T^\delta$, $\delta = 0.85$, $K = 2(\log(T))^{0.5}$, $s_T = \mathcal{O}(T^{0.05})$. Case 2 is "high-dimensional" with $p = 3 \cdot T^\delta$, $\delta = 0.85$, all else equal. The results for cases 1 and 2 are reported in Figures 2-3 and Figures 4-5 respectively. As evidenced by the figures, FGL demonstrates superior performance in both cases, exhibiting consistency for both low-dimensional and high-dimensional settings. The only instance when FGL is strictly dominated occurs in Figure 2: POET outperforms FGL in terms of convergence of precision matrix in the spectral norm. However, this changes for case 2 in Figure 4.

As a final exercise, we examine the performance of FGL and Robust FGL (described in subsection 4.6) when the dependent variable follows elliptical distributions. The data generating process (DGP) is similar to Fan et al. (2018): let $(\mathbf{f}_t, \boldsymbol{\varepsilon}_t)$ from (3.2) 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 $\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\Sigma}_f, \boldsymbol{\Sigma}_\varepsilon)$, 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 $h = 7, 7.5, 8, \ldots, 9.5$. Figures 6-7 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 robust to heavy-tailed distributions even without additional modifications. Furthermore, FGL outperforms its Robust counterpart in terms of estimating portfolio weights, as evidenced by Figure 7. We further compare the performance of FGL and Robust FGL for different degrees of freedom: Figure 8 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 8 are consistent with the findings in Fan et al. (2018): Robust FGL outperforms the non-robust counterpart for thicker tails.

6 Empirical Application

In this section we examine the performance of the Factor Graphical Lasso for constructing a financial portfolio using daily and monthly data. 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.

6.1 Data

We use monthly and 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. The full sample for the monthly data 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. 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 a 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 15 or 2 years of historical stock return data for monthly and daily returns respectively.

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, index portfolio, CLIME, LW (as in the simulations, we use a linear shrinkage estimator of covariance that incorporates the factor structure through Sherman-Morrison inversion formula) and POET. The index is the composite S&P500 index listed as ^GSPC.[3] We take the risk-free rate and Fama/French factors from Kenneth R. French's data library.

6.2 Performance Measures

Similarly to Callot et al. (2019), 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

---

[3]Using ^SPX and ^GSPC yields very similar results, we use the latter due to better data availability.

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 $n = T - m$.

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

$$\hat{\mu}_{\text{test}} = \frac{1}{n} \sum_{t=m}^{T-1} \widehat{\mathbf{w}}_t' \mathbf{r}_{t+1}, \ \hat{\sigma}_{\text{test}}^2 = \frac{1}{n-1} \sum_{t=m}^{T-1} (\widehat{\mathbf{w}}_t' \mathbf{r}_{t+1} - \hat{\mu}_{\text{test}})^2, \ \text{SR} = \hat{\mu}_{\text{test}} / \hat{\sigma}_{\text{test}}. \tag{6.1}$$

When transaction costs are considered, we follow Ban et al. (2018); Callot et al. (2019); DeMiguel et al. (2009); Li (2015) 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

$$r_{t+1,\text{portfolio}} = \widehat{\mathbf{w}}_t' \mathbf{r}_{t+1} - \text{tc}(1 + \widehat{\mathbf{w}}_t' \mathbf{r}_{t+1}) \sum_{j=1}^{p} \left| \hat{w}_{t+1,j} - \hat{w}_{t,j}^+ \right|, \tag{6.2}$$

$$\text{where} \quad \hat{w}_{t,j}^+ = \hat{w}_{t,j} \frac{1 + r_{t+1,j} + r_{t+1}^f}{1 + r_{t+1,\text{portfolio}} + r_{t+1}^f}, \tag{6.3}$$

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, variance, Sharpe Ratio and turnover are defined accordingly:

$$\hat{\mu}_{\text{test,tc}} = \frac{1}{n} \sum_{t=m}^{T-1} r_{t,\text{portfolio}}, \ \hat{\sigma}_{\text{test,tc}}^2 = \frac{1}{n-1} \sum_{t=m}^{T-1} (r_{t,\text{portfolio}} - \hat{\mu}_{\text{test,tc}})^2, \ \text{SR}_{\text{tc}} = \hat{\mu}_{\text{test,tc}} / \hat{\sigma}_{\text{test,tc}}, \tag{6.4}$$

$$\text{Turnover} = \frac{1}{n} \sum_{t=m}^{T-1} \sum_{j=1}^{p} \left| \hat{w}_{t+1,j} - \hat{w}_{t,j}^+ \right|. \tag{6.5}$$

6.3 Results

This section explores the performance of the Factor Graphical Lasso for the financial portfolio using monthly and 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. For the statistical factors we consider up to three PCs. 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 Tables 1-2, we report the monthly and daily portfolio performance

for three alternative portfolio allocations in (2.2), (2.3) and (2.10). Following Callot et al. (2019), we set a return target $\mu \in \{0.7974\%, 0.0378\%\}$ for monthly and daily data respectively (both are equivalents of 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 $\sigma \in \{0.05, 0.013\}$ which is the standard deviation of the monthly and daily excess returns of the S&P500 index in the first training set. Following Ao et al. (2019) and Callot et al. (2019), transaction costs for each individual stock are set to be a constant 0.1%.

We first comment on the results for monthly data in Table 1: **(1):** 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, to the best of our knowledge, has not been previously well-studied in the empirical finance literature. The increase in the Sharpe Ratio and return, however, comes at the cost of higher risk and higher portfolio turnover: for MRC portfolios the risk-constraint is often violated. **(2):** FGL outperforms all the competitors, including equal-weighted portfolio (EW) and Index. Specifically, our method has the lowest risk and turnover (compared to CLIME, LW 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. One of the explanations for such behavior is an underestimation of the number of factors $(\widehat{K})$, however, adjusting this quantity by $\widehat{K} + 1$ did not fix the problem. **(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.

Table 2 reports the results for daily data. Some comments are in order: **(1):** MRC portfolios produce higher return and higher risk, compared to MWC and GMV, which is consistent with the monthly results from Table 1. 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):** similarly to the results from Table 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 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.

Table 3 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 "Surge". 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 ("Boom"). Table 3 reveals some interesting findings: **(1):** the conclusions from Tables 1-2 are supported: 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.

7   Conclusion and Discussion

In this paper, we propose a new 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 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.

The empirical exercise uses the constituents of the S&P500 and demonstrates superior performance of FGL compared to several alternative models for estimating precision (CLIME) and

covariance (LW, POET) matrices, Equal-Weighted (EW) portfolio and Index portfolio in terms of the out-of-sample Sharpe Ratio and risk. This finding 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. To the best of our knowledge, this finding has not been previously well-studied in the empirical finance literature.

There are several venues for potential extensions. First, having examined empirical performance of FGL we notice that some of the estimated portfolio weights are very close to zero. This means that 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 Fan et al. (2019), Ao et al. (2019), Li (2015), Brodie et al. (2009) among others). Therefore, our model can be extended to create a sparse portfolio. Second, it is possible to make FGL estimator of the precision matrix time-varying, such that the model could also capture the dynamic nature of the relationship between stock returns. Third, one can incorporate stock-specific characteristics (e.g. company fundamentals, such as current earnings, book value, growth in net operating assets and financing) in the FGL framework, which would integrate fundamental analysis with portfolio optimization (see Lyle and Yohn (2020)). We are currently working on all of these extensions.

References

Ait-Sahalia, Y. and Xiu, D. (2017). Using principal component analysis to estimate a high dimensional factor model with high-frequency data. *Journal of Econometrics*, 201(2):384–399.

Ao, M., Yingying, L., and Zheng, X. (2019). Approaching mean-variance efficiency for large portfolios. *The Review of Financial Studies*, 32(7):2890–2919.

Awoye, O. A. (2016). *Markowitz Minimum Variance Portfolio Optimization Using New Machine Learning Methods*. PhD thesis, University College London.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150.

Ban, G.-Y., El Karoui, N., and Lim, A. E. (2018). Machine learning and portfolio optimization. *Management Science*, 64(3):1136–1154.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516.

Barigozzi, M., Brownlees, C., and Lugosi, G. (2018). Power-law partial correlation network models. *Electronic Journal of Statistics*, 12(2):2905–2929.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Brodie, J., Daubechies, I., De Mol, C., Giannone, D., and Loris, I. (2009). Sparse and stable markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272.

Brownlees, C., Nualart, E., and Sun, Y. (2018). Realized networks. *Journal of Applied Econometrics*, 33(7):986–1006.

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

Cai, T. T., Hu, J., Li, Y., and Zheng, X. (2020). High-dimensional minimum variance portfolio estimation based on high-frequency data. *Journal of Econometrics*, 214(2):482–494.

Callot, L., Caner, M., Önder, A. O., and Ulaşan, E. (2019). A nodewise regression approach to estimating large portfolios. *Journal of Business & Economic Statistics*, 0(0):1–12.

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton University Press.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304.

Connor, G. and Korajczyk, R. A. (1988). Risk and return in an equilibrium APT: Application of a new test methodology. *Journal of Financial Economics*, 21(2):255–289.

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

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Fan, J., Furger, A., and Xiu, D. (2016a). 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.

Fan, J., Liao, Y., and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39(6):3320–3356.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680.

Fan, J., Liao, Y., and Wang, W. (2016b). Projected principal component analysis in factor models. *The Annals of Statistics*, 44(1):219–254.

Fan, J., Liu, H., and Wang, W. (2018). Large covariance estimation through elliptical factor models. *The Annals of Statistics*, 46(4):1383–1414.

Fan, J., Weng, H., and Zhou, Y. (2019). Optimal estimation of functionals of high-dimensional mean and covariance matrix. *arXiv:1908.07460*.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the Graphical Lasso. *Biostatistics*, 9(3):432–441.

Goto, S. and Xu, Y. (2015). Improving mean variance optimization through sparse hedging restrictions. *Journal of Financial and Quantitative Analysis*, 50(6):1415–1441.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Janková, J. and van de Geer, S. (2018). Inference in high-dimensional graphical models. *Handbook of Graphical Models*, Chapter 14, pages 325–351. CRC Press.

Kapetanios, G. (2010). 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.

Koike, Y. (2020). De-biased graphical lasso for high-frequency data. *Entropy*, 22(4):456.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Li, H., Li, Q., and Shi, Y. (2017). Determining the number of factors when the number of factors can increase with sample size. *Journal of Econometrics*, 197(1):76–86.

Li, J. (2015). Sparse and stable portfolio selection with parameter uncertainty. *Journal of Business & Economic Statistics*, 33(3):381–392.

Lyle, M. R. and Yohn, T. L. (2020). Fundamental analysis and mean-variance optimal portfolios. *Kelley School of Business Research Paper*.

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

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Millington, T. and Niranjan, M. (2017). Robust portfolio risk minimization using the graphical lasso. In *Neural Information Processing*, pages 863–872, Cham. Springer International Publishing.

Ravikumar, P., J. Wainwright, M., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing -penalized log-determinant divergence. *Electronic Journal of Statistics*, 5.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341–360.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Tobin, J. (1958). Liquidity preference as behavior towards risk. *The Review of Economic Studies*, 25(2):65–86.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The HUGE package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*, 13(1):1059–1062.
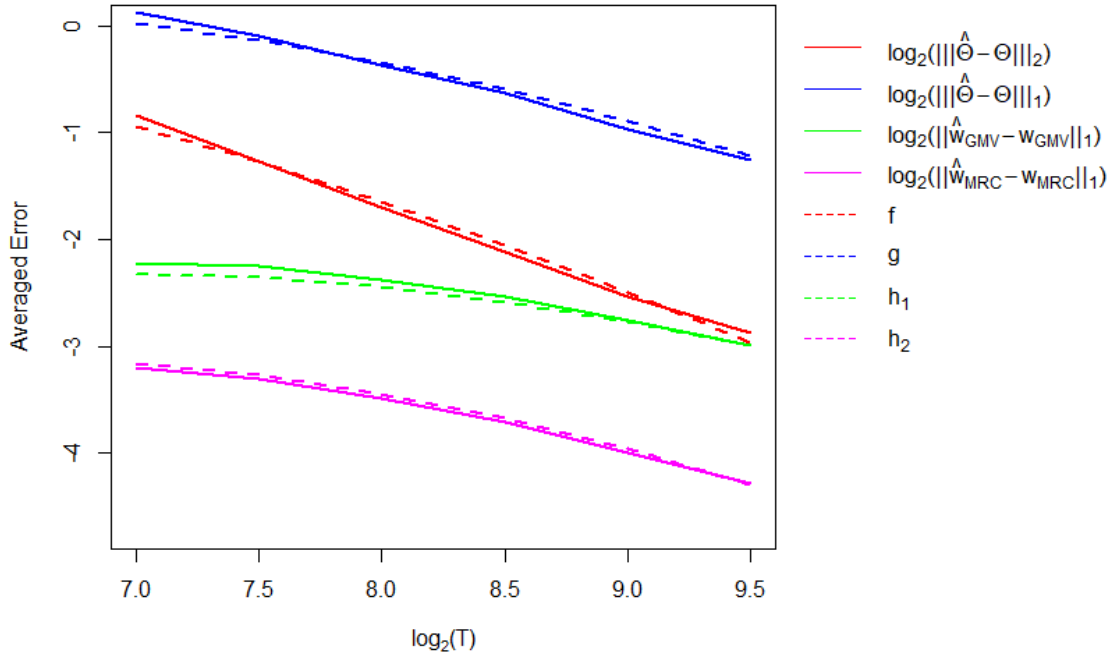
Figure 1: **Averaged empirical errors (solid lines) and theoretical rates of convergence (dashed lines) on logarithmic scale:** $p = T^{0.85}$, $K = 2(\log(T))^{0.5}$, $s_T = \mathcal{O}(T^{0.05})$.
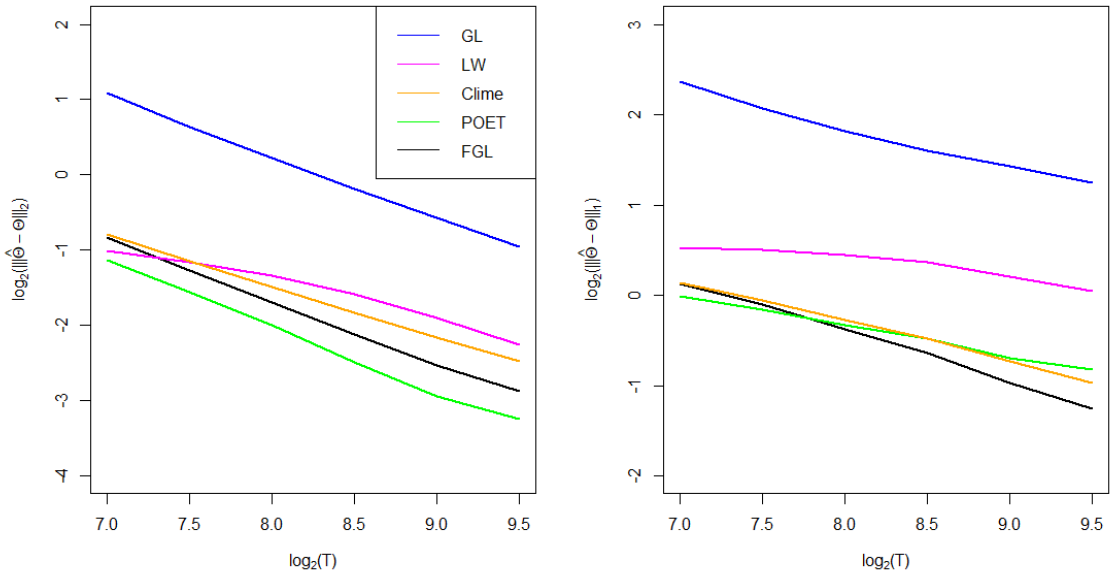

Figure 2: **Averaged errors of the estimators of Θ for Case 1 on logarithmic scale:** $p = T^{0.85}$, $K = 2(\log(T))^{0.5}$, $s_T = \mathcal{O}(T^{0.05})$.

Figure 3: **Averaged errors of the estimators of $\mathbf{w_{GMV}}$ (left) and $\mathbf{w_{MRC}}$ (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 4: **Averaged errors of the estimators of Θ for Case 2 on logarithmic scale:** $p = 3 \cdot T^{0.85}$, $K = 2(\log(T))^{0.5}$, $s_T = \mathcal{O}(T^{0.05})$.
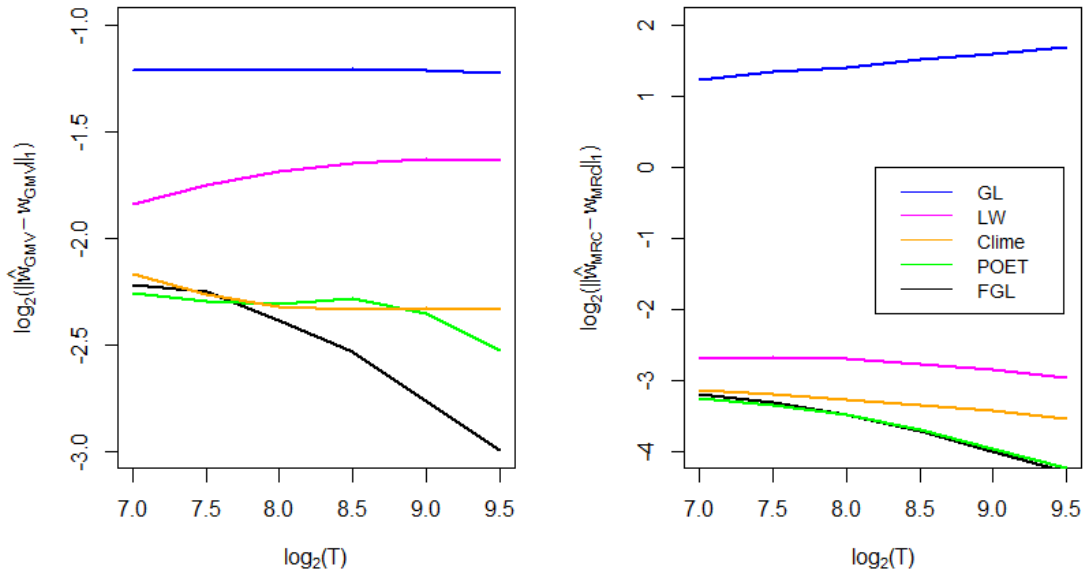
Figure 5: **Averaged errors of the estimators of w_{GMV} (left) and w_{MRC} (right) for Case 2 on logarithmic scale:** $p = 3 \cdot T^{0.85}$, $K = 2(\log(T))^{0.5}$, $s_T = \mathcal{O}(T^{0.05})$.
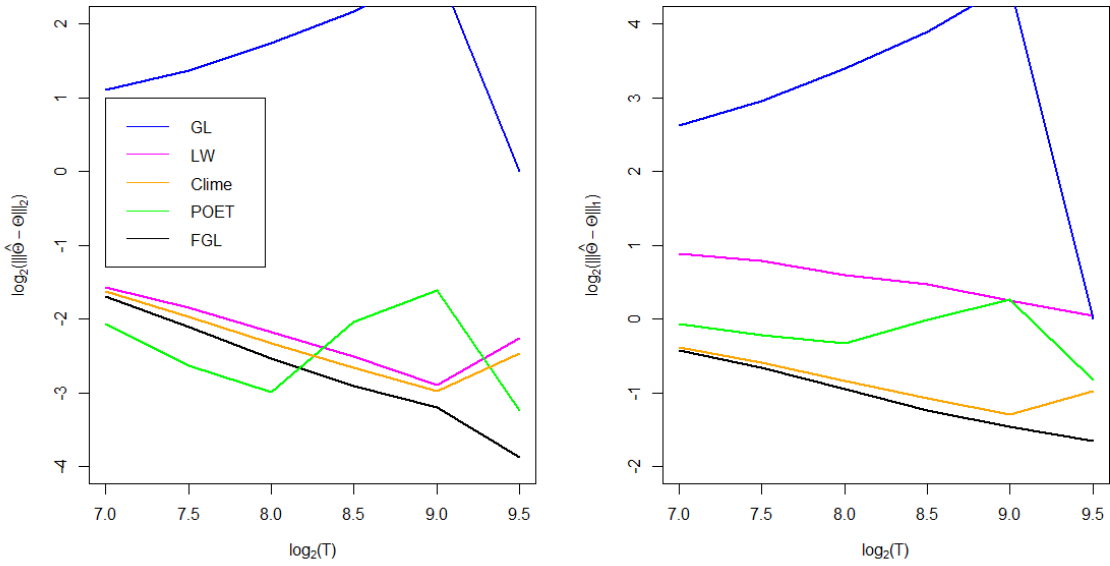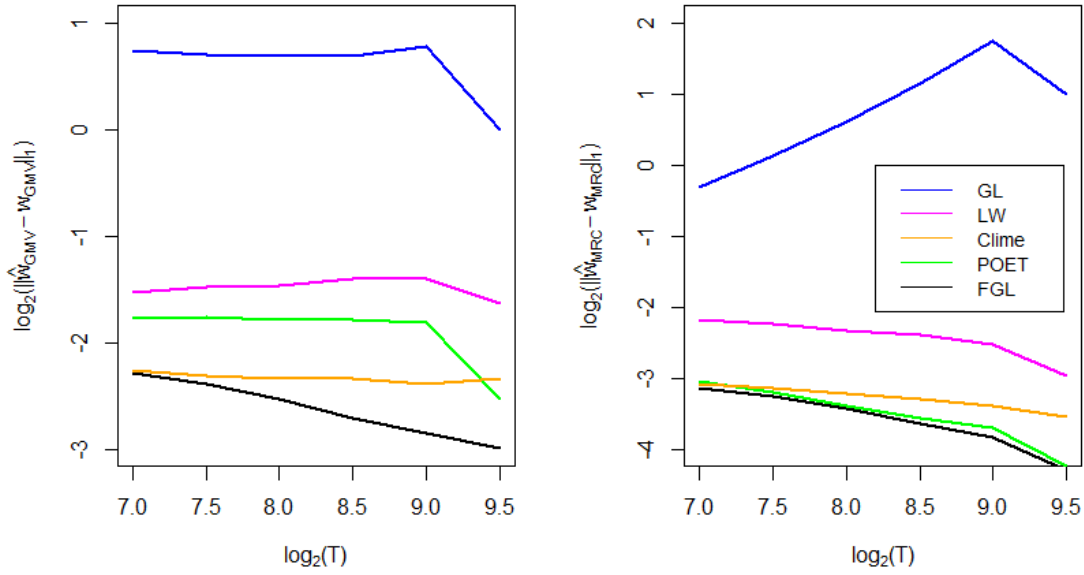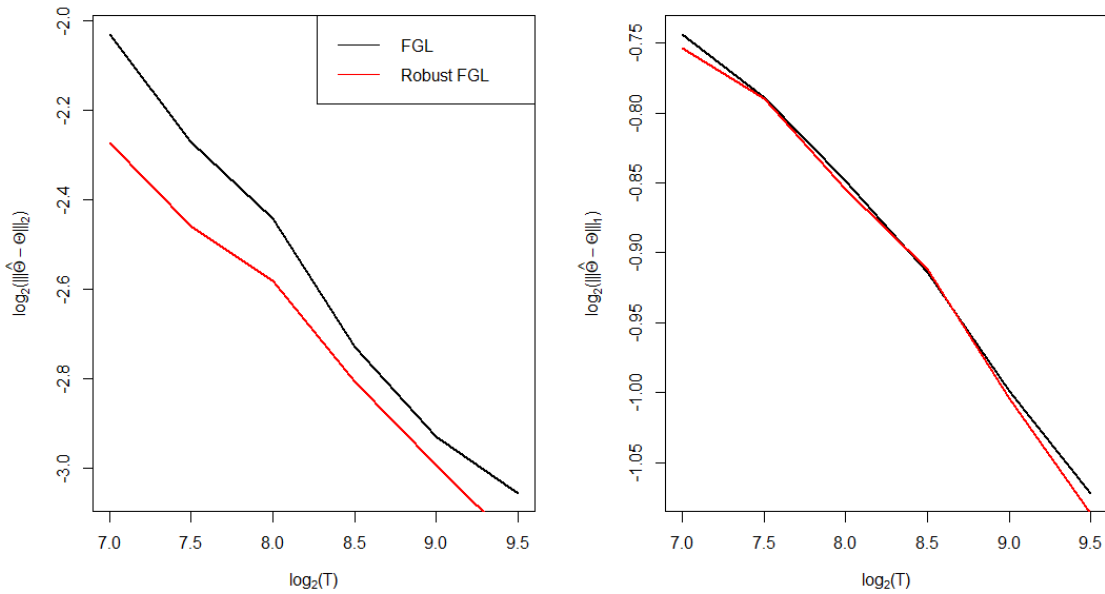

Figure 6: **Averaged errors of the estimators of Θ on logarithmic scale:** $p = T^{0.85}$, $K = 2(\log(T))^{0.5}$, $\nu = 4.2$.
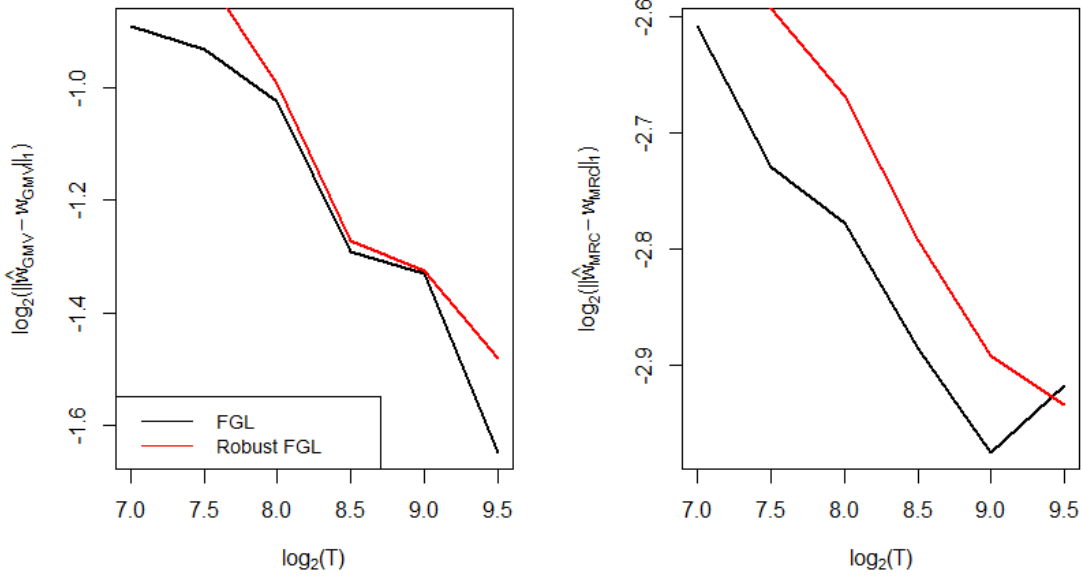
Figure 7: **Averaged errors of the estimators of $\mathbf{w}_{\mathrm{GMV}}$ (left) and $\mathbf{w}_{\mathrm{MRC}}$ (right) on logarithmic scale:** $p = T^{0.85}$, $K = 2(\log(T))^{0.5}$, $\nu = 4.2$.

Figure 8: **Log ratios (base 2) of the averaged errors of the FGL and the Robust FGL estimators of Θ:** $\log_2\left(\frac{\|\widehat{\Theta} - \Theta\|_2}{\|\widehat{\Theta}_{\mathbf{R}} - \Theta\|_2}\right)$ **(left)**, $\log_2\left(\frac{\|\widehat{\Theta} - \Theta\|_1}{\|\widehat{\Theta}_{\mathbf{R}} - \Theta\|_1}\right)$ **(right)**: $p = T^{0.85}$, $K = 2(\log(T))^{0.5}$.

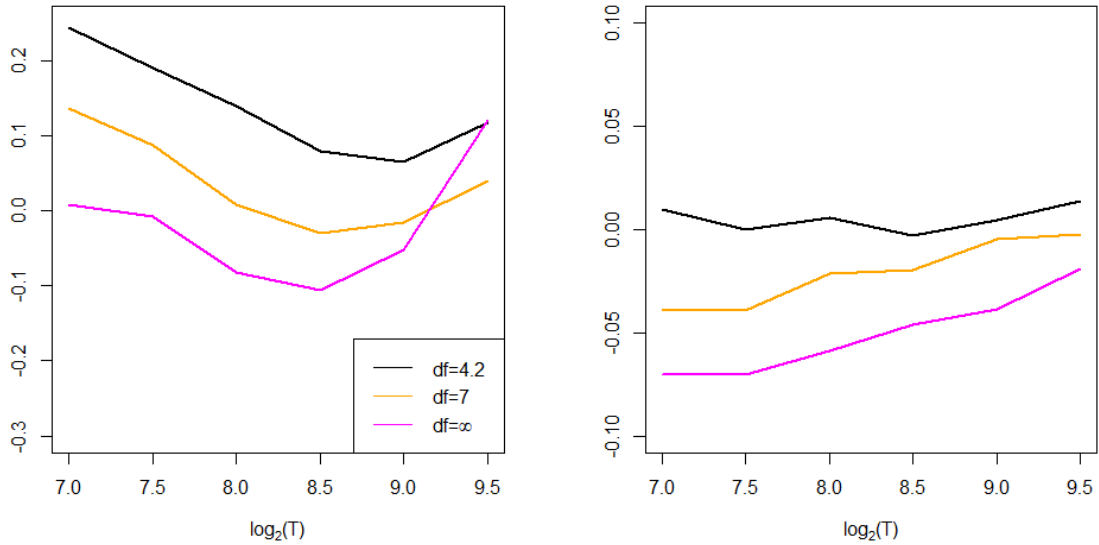|	Markowitz (risk-constrained)				Markowitz (weight-constrained)				Global Minimum-Variance			
	Return	Risk	SR	Turnover	Return	Risk	SR	Turnover	Return	Risk	SR	Turnover
Without TC												
EW	0.0081	0.0519	0.1553	-	0.0081	0.0519	0.1553	-	0.0081	0.0519	0.1553	-
Index	0.0063	0.0453	0.1389	-	0.0063	0.0453	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	-
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	-
POET	-	-	-	-	-0.1041	2.0105	-0.0518	-	0.5984	11.0064	0.0544	-
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
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
POET	-	-	-	-	-0.1876	1.7274	-0.1086	152.3298	1.0287	14.2676	0.0721	354.6043
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 1: Monthly portfolio returns, risk, Sharpe Ratio (SR) and turnover. Transaction costs are set to 50 basis points, targeted risk is set at $\sigma = 0.05$ (which is the standard deviation of the monthly excess returns on S&P 500 index from 1980 to 1995, the first training period), monthly targeted return is 0.7974% which is equivalent to 10% yearly return when compounded. In-sample: January 1, 1980 - December 31, 1995 (180 obs), Out-of-sample: January 1, 1995 - December 31, 2019 (300 obs).

	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	3.62E-04	0.0123		2.33E-04	3.62E-04	0.0123		2.33E-04	3.62E-04	0.0123	0.0292
Index	1.86E-04	1.36E-04	0.0159		1.86E-04	1.36E-04	0.0159		1.86E-04	1.36E-04	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	0.0633
CLIME	2.15E-03	8.46E-02	0.0254		2.02E-04	9.85E-03	0.0205		2.73E-04	1.07E-02	0.0255	0.0290
LW	4.34E-04	2.65E-02	0.0164		3.12E-04	9.96E-03	0.0313		3.10E-04	9.38E-03	0.0330	0.1194
POET	-	-	-		-7.06E-04	2.74E-01	-0.0026		1.07E-03	2.71E-01	0.0039	
FGL (FF1)	7.96E-04	7.82E-04	0.0285		3.73E-04	7.62E-05	0.0427		3.52E-04	7.42E-05	0.0408	0.0887
FGL (FF3)	6.51E-04	7.48E-04	0.0238		3.52E-04	8.04E-05	0.0393		3.39E-04	8.00E-05	0.0379	0.0911
FGL (FF5)	5.87E-04	7.31E-04	0.0217		3.47E-04	8.81E-05	0.0370		3.36E-04	8.62E-05	0.0362	0.0937
With TC												
EW	2.01E-04	3.62E-04	0.0106	0.0292	2.01E-04	3.62E-04	0.0106	0.0292	2.01E-04	3.62E-04	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
CLIME	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
LW	-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
POET	-	-	-	-	-2.28E-02	5.55E-01	-0.0411	113.3848	-2.81E-02	4.21E-01	-0.0666	132.8215
FGL (FF1)	3.86E-04	7.83E-04	0.0138	0.4068	2.82E-04	7.64E-05	0.0323	0.0903	2.63E-04	7.45E-05	0.0305	0.0887
FGL (FF3)	2.47E-04	7.50E-04	0.0090	0.4043	2.60E-04	8.06E-05	0.0290	0.0928	2.49E-04	8.02E-05	0.0278	0.0911
FGL (FF5)	1.83E-04	7.32E-04	0.0068	0.4032	2.53E-04	8.83E-05	0.0269	0.0952	2.43E-04	8.64E-05	0.0262	0.0937

Table 2: Daily portfolio returns, risk, Sharpe Ratio (SR) and turnover. Transaction costs are set to 50 basis points, targeted risk is set at $\sigma = 0.013$ (which is the standard deviation of the daily excess returns on S&P 500 index from 2000 to 2002, the first training period), daily targeted return is 0.0378% which is equivalent to 10% yearly return when compounded. In-sample: January 20, 2000 - January 24, 2002 (504 obs). Out-of-sample: January 17, 2002 - January 31, 2020 (4536 obs).

	Surge #1 Argentine Great Depression (2002)		Surge #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
CLIME	-0.0079	0.0348	-0.8912	0.1484	0.5331	0.0383	0.2346	0.0557
LW	0.0308	0.0231	0.2885	0.0315	0.3164	0.0118	0.5520	0.0287
Markowitz Weight-Constrained (MWC)								
FGL	-0.0138	0.0082	-0.1956	0.0135	0.1398	0.0044	0.3787	0.0072
CLIME	-0.1045	0.0124	-0.3974	0.0204	0.1309	0.0041	0.2595	0.0078
LW	-0.0158	0.0080	-0.2789	0.0126	0.1267	0.0037	0.3018	0.0085
POET	-0.2820	0.0324	-0.9989	0.1198	0.5720	0.0630	1.4756	0.0403
Global Minimum-Variance Portfolio (GMV)								
FGL	-0.0044	0.0081	-0.2113	0.0138	0.1384	0.0045	0.3703	0.0072
CLIME	-0.1061	0.0129	-0.4410	0.0241	0.1264	0.0041	0.2829	0.0081
LW	-0.0151	0.0080	-0.2926	0.0128	0.1323	0.0037	0.2994	0.0084
POET	-0.3190	0.0330	-0.9928	0.0931	-1.0000	0.2414	1.6301	0.0318

Table 3: Cumulative excess return (CER) and risk of portfolios using daily data. Transaction costs are set to 50 basis points, targeted risk is set at $\sigma = 0.013$ (which is the standard deviation of the daily excess returns on S&P 500 index from 2000 to 2002, the first training period), daily targeted return is 0.0378% which is equivalent to 10% yearly return when compounded. In-sample: January 20, 2000 - January 24, 2002 (504 obs), Out-of-sample: January 17, 2002 - January 31, 2020 (4536 obs).

35

Supplemental Appendix

A.1 Lemmas for Theorem 1

Lemma 1. *Under the assumptions of Theorem 1,*

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

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

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

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

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

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:

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

where $\zeta_{st} = \boldsymbol{\varepsilon}_s'\boldsymbol{\varepsilon}_t/p - \mathbb{E}[\boldsymbol{\varepsilon}_s'\boldsymbol{\varepsilon}_t]/p$, $\eta_{st} = \mathbf{f}_s'\sum_{i=1}^{p}\mathbf{b}_i\varepsilon_{it}/p$ 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) $(1/T)\sum_{t=1}^{T}\left[(1/T)\sum_{t=1}^{T}\hat{f}_{is}\mathbb{E}[\boldsymbol{\varepsilon}_s'\boldsymbol{\varepsilon}_t]/p\right]^2 = \mathcal{O}_p(T^{-1})$,

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

$$\frac{1}{T} \sum_{t=1}^T \left(\frac{1}{T} \sum_{s=1}^T \hat{f}_{is} \eta_{st}\right)^2 \leq \left\|\frac{1}{T} \sum_{s=1}^T \|\hat{f}_{is} \mathbf{f}_s'\|^2 \frac{1}{T} \sum_{t=1}^T \frac{1}{p} \|\sum_{j=1}^p \mathbf{b}_i \varepsilon_{jt}\|\right\|^2$$

$$\leq \frac{1}{Tp^2} \sum_{t=1}^T \left\|\sum_{j=1}^p \mathbf{b}_i \varepsilon_{jt}\right\|^2 \left(\frac{1}{T} \sum_{s=1}^T \hat{f}_{is}^2 \frac{1}{T} \sum_{s=1}^T \|\mathbf{f}_s\|^2\right)$$

$$= \mathcal{O}_p\left(\frac{K}{p} \cdot K\right) = \mathcal{O}_p\left(\frac{K^2}{p}\right).$$

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

$$\frac{1}{T} \sum_{t=1}^T \left(\frac{1}{T} \sum_{s=1}^T \hat{f}_{is} \xi_{st}\right)^2 = \frac{1}{T} \sum_{t=1}^T \left|\frac{1}{T} \sum_{s=1}^T \mathbf{f}_t' \sum_{j=1}^p \varepsilon_{js} \frac{1}{p} \hat{f}_{is}\right|^2 \leq \left(\frac{1}{T} \sum_{t=1}^T \|\mathbf{f}_t\|^2\right) \left\|\frac{1}{T} \sum_{s=1}^T \sum_{j=1}^p \mathbf{b}_j \varepsilon_{js} \frac{1}{p} \hat{f}_{is}\right\|^2$$

$$\leq \left(\frac{1}{T} \sum_{t=1}^T \|\mathbf{f}_t\|^2\right) \frac{1}{T} \sum_{s=1}^T \left\|\sum_{j=1}^p \mathbf{b}_j \varepsilon_{js} \frac{1}{p}\right\|^2 \left(\frac{1}{T} \sum_{s=1}^T \hat{f}_{is}^2\right)$$

$$= \mathcal{O}_p\left(K \cdot \frac{pK}{p^2} \cdot 1\right) = \mathcal{O}_p\left(\frac{K^2}{p}\right)$$

$\square$

Lemma 5.

(a) $\max_{t \leq T} \left\|(1/(Tp)) \sum_{s=1}^T \widehat{\mathbf{f}}_s' \mathbb{E}[\varepsilon_s' \varepsilon_t]\right\| = \mathcal{O}_p(K/\sqrt{T})$,

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

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

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

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

$$\max_{t \leq T} \left\|\frac{1}{Tp} \sum_{s=1}^T \widehat{\mathbf{f}}_s' \mathbb{E}[\varepsilon_s' \varepsilon_t]\right\| \leq \max_{t \leq T} \left[\frac{1}{T} \sum_{s=1}^T \|\widehat{\mathbf{f}}_s\| \frac{1}{T} \sum_{s=1}^T \left(\frac{\mathbb{E}[\varepsilon_s' \varepsilon_t]}{p}\right)^2\right]^{1/2} \leq \mathcal{O}_p(K) \max_{t \leq T} \left[\frac{1}{T} \sum_{s=1}^T \left(\frac{\mathbb{E}[\varepsilon_s' \varepsilon_t]}{p}\right)^2\right]^{1/2}$$

$$\leq \mathcal{O}_p(K) \max_{s,t} \sqrt{\left|\frac{\mathbb{E}[\varepsilon_s' \varepsilon_t]}{p}\right|} \max_{t \leq T} \left[\frac{1}{T} \sum_{s=1}^T \left|\frac{\mathbb{E}[\varepsilon_s' \varepsilon_t]}{p}\right|\right]^{1/2} = \mathcal{O}_p\left(K \cdot 1 \cdot \frac{1}{\sqrt{T}}\right) = \mathcal{O}_p\left(\frac{K}{\sqrt{T}}\right).$$

37

(b) Using the Cauchy-Schwarz inequality,

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

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 $\max_t (1/T) \sum_{s=1}^{T} \zeta_{st}^2 = \mathcal{O}_p \left(\sqrt{T}/p \right)$.

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

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

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

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

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

$\square$

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 (A.1), for some constant $C > 0$,

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

38

(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 A.1 and Lemma 5.

$\square$

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 $\widehat{\mathrm{cov}}(\mathbf{H}\mathbf{f}_t) = (1/T)\sum_{t=1}^{T}\mathbf{H}\mathbf{f}_t(\mathbf{H}\mathbf{f}_t)'$. Using the triangular inequality we get

$$\left\|\mathbf{H}\mathbf{H}' - \mathbf{I}_{\hat{K}}\right\|_F \le \left\|\mathbf{H}\mathbf{H}' - \widehat{\mathrm{cov}}(\mathbf{H}\mathbf{f}_t)\right\|_F + \left\|\widehat{\mathrm{cov}}(\mathbf{H}\mathbf{f}_t) - \mathbf{I}_{\hat{K}}\right\|_F. \tag{A.2}$$

To bound the first term in (A.2), we use Lemma 1: $\|\mathbf{H}\mathbf{H}' - \widehat{\mathrm{cov}}(\mathbf{H}\mathbf{f}_t)\|_F \le \|\mathbf{H}\|^2\|\mathbf{I}_K - \widehat{\mathrm{cov}}(\mathbf{H}\mathbf{f}_t)\|_F = \mathcal{O}_p(K^{5/2}/\sqrt{T})$.

To bound the second term in (A.2), we use the Cauchy-Schwarz inequality and Lemma 6:

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{H}\mathbf{f}_t(\mathbf{H}\mathbf{f}_t)' - \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t'\right\|_F \le \left\|\frac{1}{T}\sum_{t=1}^{T}(\mathbf{H}\mathbf{f}_t - \widehat{\mathbf{f}}_t)(\mathbf{H}\mathbf{f}_t)'\right\|_F + \left\|\frac{1}{T}\sum_t\widehat{\mathbf{f}}_t(\widehat{\mathbf{f}}_t' - (\mathbf{H}\mathbf{f}_t)')\right\|_F$$

$$\le \left(\frac{1}{T}\sum_{t=1}^{T}\left\|\mathbf{H}\mathbf{f}_t - \widehat{\mathbf{f}}_t\right\|^2\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{H}\mathbf{f}_t\|^2\right)^{1/2} + \left(\frac{1}{T}\sum_{t=1}^{T}\left\|\mathbf{H}\mathbf{f}_t - \widehat{\mathbf{f}}_t\right\|^2\frac{1}{T}\sum_{t=1}^{T}\left\|\widehat{\mathbf{f}}_t\right\|^2\right)^{1/2}$$

$$= \mathcal{O}_p\left(\left(\frac{K}{T} + \frac{K^3}{p}\cdot K\right)^{1/2} + \left(\frac{K}{T} + \frac{K^3}{p}\cdot K^2\right)^{1/2}\right) = \mathcal{O}_p\left(\frac{K^{3/2}}{\sqrt{T}} + \frac{K^{5/2}}{\sqrt{p}}\right).$$

(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.

$\square$

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

$$\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i = \frac{1}{T}\sum_{t=1}^{T}\mathbf{H}\mathbf{f}_t\varepsilon_{it} + \frac{1}{T}\sum_{t=1}^{T}r_{it}(\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t) + \mathbf{H}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t' - \mathbf{I}_K\right)\mathbf{b}_i. \tag{A.3}$$

Let us bound each term on the right-hand side of (A.3):

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

where we used Lemmas 1 and 7 together with Bonferroni's method.

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

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 $\mathcal{O}_p(K^2 T^{-1/2})$ since $\|(1/T) \sum_{t=1}^{T} \mathbf{f}_t \mathbf{f}_t' - \mathbf{I}_K\| = \mathcal{O}_p \left(K T^{-1/2} \right)$, $\|\mathbf{H}\| = \mathcal{O}_p(K)$
and $\max_i \|\mathbf{b}\|_i = \mathcal{O}(1)$ by Assumption (**B.1**).

A.3 Corollary 1

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

Corollary 1. *Under the assumptions of Theorem 1,*

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

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:

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

$\square$

A.4 Proof of Theorem 2

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

$$
\max_{i\leq p} \frac{1}{T}\sum_{t=1}^{T}(\varepsilon_{it} - \hat{\varepsilon}_{it})^2 \leq 4\max_i\|\mathbf{b}_i'\mathbf{H}'\|^2 \frac{1}{T}\sum_{t=1}^{T}\left\|\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\right\|^2 + 4\max_i\left\|\hat{\mathbf{b}}_i' - \mathbf{b}_i'\mathbf{H}'\right\|^2 \frac{1}{T}\sum_{t=1}^{T}\left\|\hat{\mathbf{f}}_t\right\|^2
$$

$$
+ 4\max_i\|\mathbf{b}_i'\| \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{f}_t\|^2\|\mathbf{H}'\mathbf{H} - \mathbf{I}_K\|_F^2
$$

$$
= \mathcal{O}\left(K^2\cdot\left(\frac{K}{T} + \frac{K^3}{p}\right)\right) + \mathcal{O}\left(\left(\frac{K^3\log p}{T} + \frac{K^2}{p}\right)\cdot K\right) + \mathcal{O}\left(K\cdot\left(\frac{K^5}{T} + \frac{K^5}{p}\right)\right)
$$

$$
= \mathcal{O}\left(\frac{K^4\log p}{T} + \frac{K^6}{p}\right).
$$

Let $\omega_{3T} \equiv K^2\sqrt{\log p/T} + K^3/\sqrt{p}$. Denote $\max_{i\leq p}(1/T)\sum_{t=1}^{T}(\varepsilon_{it} - \hat{\varepsilon}_{it})^2 = \mathcal{O}_p(\omega_{3T}^2)$. 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).

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) $\lambda_T^{-1}\max_{1\leq i\leq p}\left\|\hat{\mathbf{b}}_i - \mathbf{H}'\mathbf{b}_i\right\| = o_p(1/\sqrt{K})$ *and* $\max_{1\leq i\leq p}\left\|\hat{\mathbf{b}}_i\right\| = \mathcal{O}_p(\sqrt{K})$.

(c) $\lambda_T^{-1}\left\|\hat{\mathbf{B}} - \mathbf{B}\mathbf{H}'\right\| = o_p\left(\sqrt{p/K}\right)$ *and* $\left\|\hat{\mathbf{B}}\right\| = \mathcal{O}_p(\sqrt{p})$.

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

(a) Part (a) easily follows from **(B.1)**: $\text{tr}(\boldsymbol{\Sigma} - \mathbf{B}\mathbf{B}') = \text{tr}(\boldsymbol{\Sigma}) - \|\mathbf{B}\|^2 \geq 0$, since $\text{tr}(\boldsymbol{\Sigma}) = \mathcal{O}(p)$ 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 $\max_{i\leq p}\left\|\hat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\right\| = \mathcal{O}_p(\omega_{1T})$. Using the definition of λ_T from Theorem 2, it follows that $\lambda_T^{-1}\omega_{1T} = o_p(\omega_{1T}\omega_{3T}^{-1})$. Let $\tilde{z}_T \equiv \omega_{1T}\omega_{3T}^{-1}$. Consider $\lambda_T^{-1}\max_{1\leq i\leq p}\left\|\hat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\right\| = o_p(z_T)$. The latter holds for any $z_t \geq \tilde{z}_T$, with the tightest bound obtained when $z_T = \tilde{z}_T$. For the ease of representation, we use $z_T = 1/\sqrt{K}$ instead of $\tilde{z}_T$. The second result in Part (b) is obtained using the fact that $\max_{1\leq i\leq p}\left\|\hat{\mathbf{b}}_i\right\| \leq \sqrt{K}\|\mathbf{B}\|_{\max}$, where $\|\mathbf{B}\|_{\max} = \mathcal{O}(1)$ by **(B.1)**.

$\square$

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

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

(b) $\|\mathbf{\Pi}\|_2 = \mathcal{O}(1/p)$.

(c) $\lambda_T^{-1}\left\|\left\|\widehat{\mathbf{\Theta}}_f - \mathbf{\Theta}_f\right\|\right\|_2 = o_p\left(1/\sqrt{K}\right)$.

(d) $\lambda_T^{-1}\left\|\left\|\widehat{\mathbf{\Pi}} - \mathbf{\Pi}\right\|\right\|_2 = \mathcal{O}_p\left(s_T/p + 1/\left(p^2\sqrt{K}\right)\right)$ *and* $\left\|\left\|\widehat{\mathbf{\Pi}}\right\|\right\|_2 = \mathcal{O}_p(1/p)$.

Proof.

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

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

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

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

Let $\omega_{4T} = K^{3/2}/\sqrt{T} + K^{5/2}/\sqrt{p}$. Using the definition of λ_T from Theorem 2, it follows that $\lambda_T^{-1}\omega_{4T} = o_p(\omega_{4T}\omega_{3T}^{-1})$. Let $\widetilde{\gamma}_T \equiv \omega_{4T}\omega_{3T}^{-1}$. Consider $\lambda_T^{-1}\left\|\left\|\widehat{\mathbf{\Theta}}_f - \mathbf{\Theta}_f\right\|\right\|_2 = o_p(\gamma_T)$. 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

$$\left\|\left\|\widehat{\mathbf{B}}'\widehat{\mathbf{\Theta}}_\varepsilon\widehat{\mathbf{B}} - (\mathbf{BH'})'\mathbf{\Theta}_\varepsilon(\mathbf{BH'})\right\|\right\|_2 \leq \left\|\left\|\widehat{\mathbf{B}} - \mathbf{BH'}\right\|\right\|_2\left\|\left\|\widehat{\mathbf{\Theta}}_\varepsilon\right\|\right\|_2\left\|\left\|\widehat{\mathbf{B}}\right\|\right\|_2 + \left\|\left\|\mathbf{BH'}\right\|\right\|_2\left\|\left\|\widehat{\mathbf{\Theta}}_\varepsilon - \mathbf{\Theta}_\varepsilon\right\|\right\|_2\left\|\left\|\widehat{\mathbf{B}}\right\|\right\|_2$$

$$+ \left\|\left\|\mathbf{BH'}\right\|\right\|_2\|\mathbf{\Theta}_\varepsilon\|_2\left\|\left\|\widehat{\mathbf{B}} - \mathbf{BH'}\right\|\right\|_2 = \mathcal{O}_p\left(p \cdot s_T \cdot \lambda_T\right). \quad \text{(A.4)}$$

Now we combine (A.4) with the results from Parts (b)-(c):

$$\lambda_T^{-1}\left\|\!\left\|\boldsymbol{\Pi}\left(\widehat{\boldsymbol{\Pi}}^{-1}-\boldsymbol{\Pi}^{-1}\right)\right\|\!\right\|_2 = \mathcal{O}_p\Big(s_t + \frac{1}{p\sqrt{K}}\Big).$$

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

$$\lambda_T^{-1}\left\|\!\left\|\widehat{\boldsymbol{\Pi}}-\boldsymbol{\Pi}\right\|\!\right\|_2 \le \lambda_T^{-1}\frac{\left\|\!\left\|\boldsymbol{\Pi}\right\|\!\right\|_2\left\|\!\left\|\boldsymbol{\Pi}\left(\widehat{\boldsymbol{\Pi}}^{-1}-\boldsymbol{\Pi}^{-1}\right)\right\|\!\right\|_2}{1-\left\|\!\left\|\boldsymbol{\Pi}\left(\widehat{\boldsymbol{\Pi}}^{-1}-\boldsymbol{\Pi}^{-1}\right)\right\|\!\right\|_2} = \mathcal{O}_p\left(\frac{1}{p}\Big(s_t + \frac{1}{p\sqrt{K}}\Big)\right).$$

$\square$

A.6   Proof of Theorem 3

Using the Sherman-Morrison-Woodbury formula, we have

$$\left\|\!\left\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\right\|\!\right\|_l \le \left\|\!\left\|\widehat{\boldsymbol{\Theta}}_\varepsilon-\boldsymbol{\Theta}_\varepsilon\right\|\!\right\|_l + \left\|\!\left\|(\widehat{\boldsymbol{\Theta}}_\varepsilon-\boldsymbol{\Theta}_\varepsilon)\widehat{\mathbf{B}}\widehat{\boldsymbol{\Pi}}\widehat{\mathbf{B}}'\widehat{\boldsymbol{\Theta}}_\varepsilon\right\|\!\right\|_l + \left\|\!\left\|\boldsymbol{\Theta}_\varepsilon(\widehat{\mathbf{B}}-\mathbf{BH}')\widehat{\boldsymbol{\Pi}}\widehat{\mathbf{B}}'\widehat{\boldsymbol{\Theta}}_\varepsilon\right\|\!\right\|_l$$
$$+ \left\|\!\left\|\boldsymbol{\Theta}_\varepsilon\mathbf{BH}'(\widehat{\boldsymbol{\Pi}}-\boldsymbol{\Pi})\widehat{\mathbf{B}}'\widehat{\boldsymbol{\Theta}}_\varepsilon\right\|\!\right\|_l + \left\|\!\left\|\boldsymbol{\Theta}_\varepsilon\mathbf{BH}'\boldsymbol{\Pi}(\widehat{\mathbf{B}}-\mathbf{B})'\widehat{\boldsymbol{\Theta}}_\varepsilon\right\|\!\right\|_l + \left\|\!\left\|\boldsymbol{\Theta}_\varepsilon\mathbf{BH}'\boldsymbol{\Pi}(\mathbf{BH}')'(\widehat{\boldsymbol{\Theta}}_\varepsilon-\boldsymbol{\Theta}_\varepsilon)\right\|\!\right\|_l$$
$$= \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 + \Delta_5 + \Delta_6. \tag{A.5}$$

We now bound the terms in (A.5) for $l = 2$ and $l = \infty$. We start with $l = 2$. First, note that $\lambda_T^{-1}\Delta_1 = \mathcal{O}_p(s_T)$ by Theorem 2. Second, using Lemmas 8-9 together with Theorem 2, we have $\lambda_T^{-1}(\Delta_2 + \Delta_6) = \mathcal{O}_p(s_T \cdot \sqrt{p} \cdot (1/p) \cdot \sqrt{p} \cdot 1) = \mathcal{O}_p(s_T)$. Third, $\lambda_T^{-1}(\Delta_3 + \Delta_5)$ is negligible according to Lemma 8(c). Finally, $\lambda_T^{-1}\Delta_4 = \mathcal{O}_p\Big(1 \cdot \sqrt{p} \cdot \big((s_T/p) + (1/(p^2\sqrt{K}))\big) \cdot \sqrt{p} \cdot 1\Big) = \mathcal{O}_p(s_T + 1/(p\sqrt{K}))$ by Lemmas 8-9 and Theorem 2.

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

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) $\|\!|\boldsymbol{\Theta}|\!\|_1 = \mathcal{O}(d_T K^{3/2})$, *where* d_T *was defined in Assumption* **B.3**.

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:

$$\|\|\mathbf{\Theta}\|\|_1 \leq \|\|\mathbf{\Theta}_\varepsilon\|\|_1 + \|\|\mathbf{\Theta}_\varepsilon\mathbf{B}[\mathbf{\Theta}_f + \mathbf{B}'\mathbf{\Theta}_\varepsilon\mathbf{B}]^{-1}\mathbf{B}'\mathbf{\Theta}_\varepsilon\|\|_1$$

$$= \mathcal{O}(\sqrt{d_T}) + \mathcal{O}\left(\sqrt{d_T}\cdot p\cdot\frac{\sqrt{K}}{p}\cdot K\cdot\sqrt{d_T}\right) = \mathcal{O}(d_T K^{3/2}). \tag{A.6}$$

The last equality in (A.6) 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 $a = \boldsymbol{\iota}'_p\mathbf{\Theta}\boldsymbol{\iota}_p/p$, $b = \boldsymbol{\iota}'_p\mathbf{\Theta}\mathbf{m}/p$, $d = \mathbf{m}'\mathbf{\Theta}\mathbf{m}/p$, $g = \sqrt{\mathbf{m}'\mathbf{\Theta}\mathbf{m}}/p$ and $\hat{a} = \boldsymbol{\iota}'_p\widehat{\mathbf{\Theta}}\boldsymbol{\iota}_p/p$, $\hat{b} = \boldsymbol{\iota}'_p\widehat{\mathbf{\Theta}}\widehat{\mathbf{m}}/p$, $\hat{d} = \widehat{\mathbf{m}}'\widehat{\mathbf{\Theta}}\widehat{\mathbf{m}}/p$, $\hat{g} = \sqrt{\widehat{\mathbf{m}}'\widehat{\mathbf{\Theta}}\widehat{\mathbf{m}}}/p$. Under the assumptions of Theorem 4 and assuming $(ad - b^2) > 0$,*

(a) $a \geq C_0 > 0$, $b = \mathcal{O}(1)$, $d = \mathcal{O}(1)$.

(b) $|\hat{a} - a| = \mathcal{O}_p(\lambda_T d_T K^{3/2}(s_T + (1/p))) = o_p(1)$.

(c) $\left|\hat{b} - b\right| = \mathcal{O}_p(\lambda_T d_T K^{3/2}(s_T + (1/p))) = o_p(1)$

(d) $\left|\hat{d} - d\right| = \mathcal{O}_p(\lambda_T d_T K^{3/2}(s_T + (1/p))) = o_p(1)$.

(e) $|\hat{g} - g| = \mathcal{O}_p\left(\left[\lambda_T d_T K^{3/2}(s_T + (1/p))\right]^{1/2}\right) = o_p(1)$.

(f) $\left|(\hat{a}\hat{d} - \hat{b}^2) - (ad - b^2)\right| = \mathcal{O}_p\left(\lambda_T d_T K^{3/2}(s_T + (1/p))\right) = o_p(1)$.

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

Proof.

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

(b) Using the Hölders inequality, we have

$$|\hat{a} - a| = \left|\frac{\boldsymbol{\iota}'_p(\widehat{\mathbf{\Theta}} - \mathbf{\Theta})\boldsymbol{\iota}_p}{p}\right| \leq \frac{\left\|(\widehat{\mathbf{\Theta}} - \mathbf{\Theta})\boldsymbol{\iota}_p\right\|_1\|\boldsymbol{\iota}_p\|_{\max}}{p} \leq \|\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|\|_1$$

$$= \mathcal{O}_p\left(\lambda_T d_T K^{3/2}(s_T + (1/p))\right) = o_p(1),$$

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

(c) First, rewrite the expression of interest:

$$\hat{b} - b = [\boldsymbol{\iota}'_p(\widehat{\mathbf{\Theta}} - \mathbf{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})]/p + [\boldsymbol{\iota}'_p(\widehat{\mathbf{\Theta}} - \mathbf{\Theta})\mathbf{m}]/p + [\boldsymbol{\iota}'_p\mathbf{\Theta}(\widehat{\mathbf{m}} - \mathbf{m})]/p. \tag{A.7}$$

We now bound each of the terms in (A.7) using the expressions derived in Callot et al. (2019) (see their Proof of Lemma A.2) and the fact that $\log(p)/T = o(1)$.

$$\left|\boldsymbol{\iota}_p'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})\right|/p \leq \left\|\left|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right|\right\|_1 \|\widehat{\mathbf{m}} - \mathbf{m}\|_{\max} = \mathcal{O}_p\left(\lambda_T d_T K^{3/2}(s_T + (1/p)) \cdot \sqrt{\frac{\log(p)}{T}}\right).$$
(A.8)

$$\left|\boldsymbol{\iota}_p'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\mathbf{m}\right|/p \leq \left\|\left|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right|\right\|_1 = \mathcal{O}_p\left(\lambda_T d_T K^{3/2}(s_T + (1/p))\right).$$
(A.9)

$$\left|\boldsymbol{\iota}_p'\boldsymbol{\Theta}(\widehat{\mathbf{m}} - \mathbf{m})\right|/p \leq \|\|\boldsymbol{\Theta}\|\|_1 \|\widehat{\mathbf{m}} - \mathbf{m}\|_{\max} = \mathcal{O}_p\left(d_T K^{3/2} \cdot \sqrt{\frac{\log(p)}{T}}\right).$$
(A.10)

(d) First, rewrite the expression of interest:

$$\widehat{d} - d = [(\widehat{\mathbf{m}} - \mathbf{m})'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})]/p + [(\widehat{\mathbf{m}} - \mathbf{m})'\boldsymbol{\Theta}(\widehat{\mathbf{m}} - \mathbf{m})]/p$$
$$+ [2(\widehat{\mathbf{m}} - \mathbf{m})'\boldsymbol{\Theta}\mathbf{m}]/p + [2\mathbf{m}'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})]/p$$
$$+ [\mathbf{m}'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\mathbf{m}]/p.$$
(A.11)

We now bound each of the terms in (A.11) using the expressions derived in Callot et al. (2019) (see their Proof of Lemma A.3) and the fact that $\log(p)/T = o(1)$.

$$\left|(\widehat{\mathbf{m}} - \mathbf{m})'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})\right|/p \leq \|\widehat{\mathbf{m}} - \mathbf{m}\|_{\max}^2 \left\|\left|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right|\right\|_1$$
$$= \mathcal{O}_p\left(\frac{\log(p)}{T} \cdot \lambda_T d_T K^{3/2}(s_T + (1/p))\right)$$
(A.12)

$$\left|(\widehat{\mathbf{m}} - \mathbf{m})'\boldsymbol{\Theta}(\widehat{\mathbf{m}} - \mathbf{m})\right|/p \leq \|\widehat{\mathbf{m}} - \mathbf{m}\|_{\max}^2 \|\|\boldsymbol{\Theta}\|\|_1 = \mathcal{O}_p\left(\frac{\log(p)}{T} \cdot d_T K^{3/2}\right).$$
(A.13)

$$\left|(\widehat{\mathbf{m}} - \mathbf{m})'\boldsymbol{\Theta}\mathbf{m}\right|/p \leq \|\widehat{\mathbf{m}} - \mathbf{m}\|_{\max} \|\|\boldsymbol{\Theta}\|\|_1 = \mathcal{O}_p\left(\sqrt{\frac{\log(p)}{T}} \cdot d_T K^{3/2}\right).$$
(A.14)

$$\left|\mathbf{m}'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})\right|/p \leq \|\widehat{\mathbf{m}} - \mathbf{m}\|_{\max} \left\|\left|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right|\right\|_1$$
$$= \mathcal{O}_p\left(\sqrt{\frac{\log(p)}{T}} \cdot \lambda_T d_T K^{3/2}(s_T + (1/p))\right).$$
(A.15)

$$\left|\mathbf{m}'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\mathbf{m}\right|/p \leq \left\|\left|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right|\right\|_1 = \mathcal{O}_p\left(\lambda_T d_T K^{3/2}(s_T + (1/p))\right).$$
(A.16)

(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

$$\left|(\widehat{a}\widehat{d} - \widehat{b}^2) - (ad - b^2)\right| \leq \left[|\widehat{a} - a|\left|\widehat{d} - d\right| + |\widehat{a} - a|d + a\left|\widehat{d} - d\right| + (\widehat{b} - b)^2 + 2|b|\left|\widehat{b} - b\right|\right]$$
$$= \mathcal{O}_p\left(\lambda_T d_T K^{3/2}(s_T + (1/p))\right) = o_p(1).$$

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

$\square$

A.8 Proof of Theorem 4

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

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

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:

$$
\begin{aligned}
\|\widehat{\mathbf{w}}_{\text{MWC}} - \mathbf{w}_{\text{MWC}}\|_1 &\leq |(\widehat{\kappa}_1 - \kappa_1)|\left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\boldsymbol{\iota}_p\right\|_1/p + |(\widehat{\kappa}_1 - \kappa_1)|\|\boldsymbol{\Theta}\boldsymbol{\iota}_p\|_1/p + |\kappa_1|\left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\boldsymbol{\iota}_p\right\|_1/p \\
&\quad + |(\widehat{\kappa}_2 - \kappa_2)|\left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})\right\|_1/p + |(\widehat{\kappa}_2 - \kappa_2)|\|\boldsymbol{\Theta}(\widehat{\mathbf{m}} - \mathbf{m})\|_1/p \\
&\quad + |(\widehat{\kappa}_2 - \kappa_2)|\left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\mathbf{m}\right\|_1/p + |(\widehat{\kappa}_2 - \kappa_2)|\|\boldsymbol{\Theta}\mathbf{m}\|_1/p \\
&\quad + |\kappa_2|\left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})\right\|_1/p + |\kappa_2|\left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\mathbf{m}\right\|_1/p \quad \text{(A.17)}
\end{aligned}
$$

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

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

where the rate trivially follows from Lemma 11.
Similarly, we get

$$|(\widehat{\kappa}_2 - \kappa_2)| = \mathcal{O}_p\Big(\lambda_T d_T K^{3/2}(s_T + (1/p))\Big) = o_p(1).$$

Callot et al. (2019) showed that $|\kappa_1| = \mathcal{O}(1)$ and $|\kappa_2| = \mathcal{O}(1)$. Therefore, we can get the rate of (A.17):

$$\|\widehat{\mathbf{w}}_{\text{MWC}} - \mathbf{w}_{\text{MWC}}\|_1 = \mathcal{O}_p\Big(\lambda_T d_T^2 K^3(s_T + (1/p))\Big) = o_p(1).$$

We now proceed with the MRC weight formulation:

$$
\left\|\widehat{\mathbf{w}}_{\mathrm{MRC}} - \mathbf{w}_{\mathrm{MRC}}\right\|_1 \leq \frac{\frac{g}{p}\left[\left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})(\widehat{\mathbf{m}} - \mathbf{m})\right\|_1 + \left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\mathbf{m}\right\|_1 + \|\boldsymbol{\Theta}(\widehat{\mathbf{m}} - \mathbf{m})\|_1\right] + |\widehat{g} - g|\|\boldsymbol{\Theta}\mathbf{m}\|_1}{|\widehat{g}|g}
$$

$$
\leq \frac{\frac{g}{p}\left[p\left\|\!\left\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right\|\!\right\|_1\|(\widehat{\mathbf{m}} - \mathbf{m})\|_{\max} + p\left\|\!\left\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right\|\!\right\|_1\|\mathbf{m}\|_{\max} + p\|\!\|\boldsymbol{\Theta}\|\!\|_1\|(\widehat{\mathbf{m}} - \mathbf{m})\|_{\max}\right] + p|\widehat{g} - g|\|\!\|\boldsymbol{\Theta}\|\!\|_1\|\mathbf{m}\|_{\max}}{|\widehat{g}|g}
$$

$$
= \mathcal{O}_p\!\left(\lambda_T d_T K^{3/2}(s_T + (1/p)) \cdot \sqrt{\frac{\log(p)}{T}}\right) + \mathcal{O}_p\!\left(\lambda_T d_T K^{3/2}(s_T + (1/p))\right)
$$

$$
+ \mathcal{O}_p\!\left(d_T K^{3/2} \cdot \sqrt{\frac{\log(p)}{T}}\right) + \mathcal{O}_p\!\left([\lambda_T d_T K^{3/2}(s_T + (1/p))]^{1/2} \cdot d_T K^{3/2}\right) = o_p(1),
$$

where we used Lemmas 10-11.