Solving the Markowitz Optimization Problem for Large Portfolios [*][†]

Mengmeng Ao[‡] Yingying Li[§] Xinghua Zheng[¶]

First Draft: October 6, 2014
This Draft: December 22, 2017

Abstract

This paper studies the large dimensional Markowitz optimization problem. Given any risk constraint level, we introduce a new approach for estimating the optimal portfolio. The approach relies on a novel *unconstrained* regression representation of the mean-variance optimization problem, combined with high-dimensional sparse regression methods. Our estimated portfolio, under a mild sparsity assumption, asymptotically achieves mean-variance efficiency and meanwhile effectively controls the risk. To the best of our knowledge, this is the first approach that can achieve these two goals simultaneously for large portfolios. The superior properties of our approach are demonstrated via comprehensive simulation and empirical studies.

Keywords: Markowitz optimization; Large portfolio selection; Unconstrained regression, LASSO; Sharpe ratio

[*] Research partially supported by the RGC grants GRF16305315, GRF 16502014 and GRF 16518716 of the HKSAR, and The Fundamental Research Funds for the Central Universities 20720171073.

[†] We thank Yacine Ait-Sahalia, Torben G. Andersen, Federico Bandi, Tim Bollerslev, Robert Engle, Jianqing Fan, Eric Ghysels, Campbell Harvey, Joel Hasbrouck, Ravi Jagannathan, Raymond Kan, Steven Lalley, Jia Li, Per Mykland, Deniela Osterrieder, Andrew Patton, Matt Richardson, Jeff Russel, George Tauchen, Allan Timmermann, Viktor Tordorov, Ruey Tsay, Dacheng Xiu, and participants of various conferences and seminars for their valuable and helpful comments.

[‡] Wang Yanan Institute for Studies in Economics & Department of Finance, School of Economics, Xiamen University, China. mengmengao@xmu.edu.cn

[§] Hong Kong University of Science and Technology, HKSAR. yyli@ust.hk

[¶] Hong Kong University of Science and Technology, HKSAR. xhzheng@ust.hk

1 INTRODUCTION

1.1 Markowitz Optimization Enigma

The groundbreaking mean-variance portfolio theory proposed by Markowitz (1952) continues to play significant roles in research and practice. The mean-variance efficient portfolio has a simple explicit expression[1] that only depends on two population characteristics, the mean and covariance matrix of asset returns. Under the ideal situation when the underlying mean and covariance matrix are known, mean-variance investors can easily compute the optimal portfolio weights based on their preferred level of risk or targeted rate of return. In the real world, however, the true parameters are unknown. Sample mean and sample covariance matrix are used as proxies, and the resulting "plug-in" portfolio had been widely adopted. Such an approach is justified by the classical statistics theory because the plug-in portfolio is an MLE of the optimal portfolio. However, as documented in Michaud (1989) and others, the out-of-sample performance of the plug-in portfolio is poor. Moreover, the situation worsens as the number of assets increases. (For additional details, see Best and Grauer (1991), Green and Hollifield (1992), Chopra and Ziemba (1993), Britten-Jones (1999), Kan and Zhou (2007), and Basak et al. (2009) among others.) Termed the "Markowitz Optimization Enigma" by Michaud (1989), the issues of constructing the mean-variance efficient portfolio based on sample estimates limit the use of Markowitz's mean-variance framework.

1.2 Challenges for Large Portfolios

Modern portfolios often include a large number of assets. This makes the optimization problem high-dimensional in nature and induces serious challenges. Take the plug-in portfolio for example, as we will see in Figure 1, the risk of the plug-in portfolio can be substantially higher than the pre-specified risk level even when the portfolio weights are computed based on simulated i.i.d. returns. On the other hand, such a high risk is not well compensated by a high return, resulting in a significantly suboptimal Sharpe ratio. The key message is, **even in the ideal situation when all assumptions of Markowitz optimization are satisfied** (i.e., normally distributed returns, no time-varying parameters and regime switching, etc.), **there are intrinsic challenges towards the estimation of the mean-variance**

[1] See details in Section 2.1.

efficient portfolio.

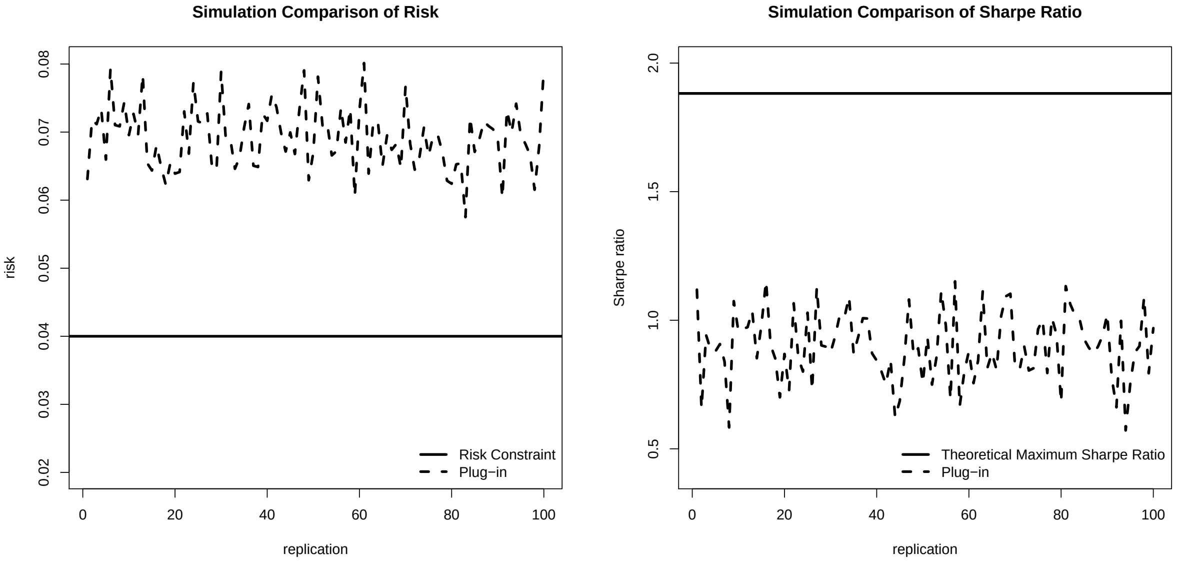

Figure 1. *Comparisons of risks and Sharpe ratios of the plug-in portfolio versus the risk constraint and the theoretical maximum Sharpe ratio, respectively. The portfolios are constructed based on data generated from i.i.d. multivariate normal distribution with parameters calibrated from real data (see Section 3.2 for details). The left panel plots the portfolio risks, and the right panel plots the Sharpe ratios. The asset pool includes 100 stocks and 3 factors, and the number of observations is 240. The comparison is replicated 100 times.*

Figure 1 shows the comparisons between the plug-in portfolio (black dashed lines) and the theoretical optimal portfolio (black solid lines). The asset pool consists of 100 stocks and 3 factors. The underlying mean and covariance matrix are calibrated from real data (see Section 3.2 for details). We generate 20 years of monthly returns from i.i.d. multivariate normal distribution, based on which we construct the plug-in portfolio and compare its risk and Sharpe ratio with the risk constraint and the theoretical maximum Sharpe ratio. Such simulated returns satisfy all the assumptions of Markowitz's mean-variance framework. However, as we observe from Figure 1, in all 100 replications, the plug-in portfolio carries a risk that is almost twice the specified level. Meanwhile, as shown in the right panel of Figure 1, the Sharpe ratio of the plug-in portfolio is only around 50% of the theoretical maximum Sharpe ratio.

The phenomenon above has been noted in Kan and Zhou (2007), and later investigated in Bai et al. (2009) and El Karoui (2010). These papers document that the deviation of the plug-in portfolio from the optimal portfolio is systematic, and the bias is due to the dimension (number of assets) being not small compared with the sample size. More precisely, one has that, under normality assumption on the returns and if the ratio between the number of assets, N, and the sample size, T, satisfies that

$$\frac{N}{T} \to \rho \in (0,1),$$

then the Sharpe ratio of the plug-in portfolio, $SR(\text{plug-in})$, is asymptotically related to the theoretical maximum Sharpe ratio, SR^*, as follows

$$\frac{SR(\text{plug-in})}{SR^*} \xrightarrow{P} \sqrt{\frac{1-\rho}{1+\rho/(SR^*)^2}} < \sqrt{1-\rho} < 1, \tag{1.1}$$

where "$\xrightarrow{P}$" stands for convergence in probability. The proof of (1.1) is given in Appendix A. For global minimum variance portfolio (GMV), Basak et al. (2009) derive a result in a similar spirit, which says that the plug-in GMV carries, on average, a risk that is a bigger-than-1 multiple of the true minimum risk, and the multiplier explicitly depends on the number of assets and sample size.

1.3 Existing Alternative Methods

The plug-in portfolio is obtained by replacing the population mean and covariance matrix in the formula for the optimal portfolio with their sample estimates. Alternatively, people seek to improve portfolio performance by plugging in better estimates of the underlying mean and covariance matrix. For estimation of covariance matrix, a widely used alternative estimator is the linear shrinkage estimator proposed in Ledoit and Wolf (2003, 2004), which estimates the covariance matrix by a suitable linear combination of the sample covariance matrix and a target matrix (e.g., identity or single-index matrix). More recently, Ledoit and Wolf (2017) propose a nonlinear shrinkage estimator of the covariance matrix and its factor-model-adjusted version that are suitable for portfolio optimization. For estimation of mean, among other works, Black and Litterman (1991) propose a quasi-Bayesian approach by combining investors' views with returns implied by CAPM. This quasi-Bayesian approach is extended to a fully Bayesian approach by Lai et al. (2011), who consider the mean-variance problem from a different angle and aim to maximize a certain utility function. Garlappi et al. (2007) propose to adjust estimates of expected returns by a multi-prior approach, also with an aim of maximizing a utility function. The aforementioned paper Bai et al. (2009), which analyze the systematic bias in the plug-in portfolio, propose a "bootstrap-corrected" method to estimate the optimal portfolio. However, as pointed out in a more recent working paper (Bai et al. (2013)), the bootstrap-corrected method fails to satisfy the risk constraint.

Another direction to improve portfolio performance is to modify the original framework by imposing various constraints on portfolio weights. Most research in this direction focuses on the GMV portfolio. Imposing constraints on weights has been empirically shown to be helpful; see, for example, Jagannathan and Ma (2003), DeMiguel et al. (2009), Brodie et al. (2009) and Fastrich et al. (2012). Fan et al. (2012b) give theoretical justifications to the empirical results in Jagannathan and Ma (2003), and also investigate the GMV portfolio with gross-exposure constraints where some short positions are allowed. Fan et al. (2012a) consider GMV portfolio with high-frequency data under gross-exposure constraints.

In addition to the approaches mentioned above, combinations of different portfolios have been studied; see, for example, Kan and Zhou (2007) and Tu and Zhou (2011).

The aforementioned methods lead to improved portfolio performance. However, they are still suboptimal. Take the latest development, the nonlinear shrinkage method in Ledoit and Wolf (2017) as an example, we see in Figure 2 that although its risk is substantially lower than that of the plug-in portfolio, the portfolio still violates the risk constraint, and is also significantly suboptimal in terms of Sharpe ratio.

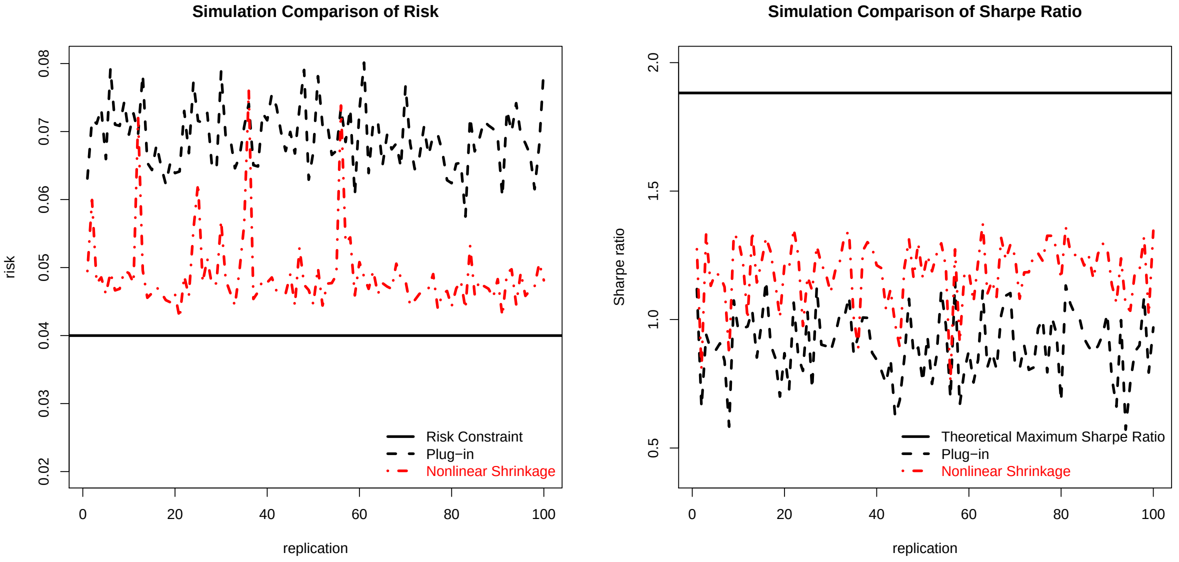

Figure 2. *Comparisons of the plug-in and nonlinear shrinkage portfolios. The portfolios are constructed based on the same data used for Figure 1. The left and right panels plot the portfolio risks and Sharpe ratios, respectively. There are 100 stocks and 3 factors in the asset pool, and the sample size is 240. The comparison is replicated 100 times.*

Comprehensive comparisons including various other strategies are made in Sections 3 and 4, based on both simulated and empirical data. The comparisons reveal similar conclusions.

1.4 Our Contributions

In this paper, a new methodology for estimating the mean-variance efficient portfolio is proposed, which we call the MAXimum-Sharpe-ratio Estimated & sparse Regression (MAXSER) method. MAXSER is a general approach that can be applied to various situations when the number of assets is not small compared with sample size. We show that, under mild assumptions, the MAXSER portfolio can asymptotically (1) achieve mean-variance efficiency and (2) satisfy the risk constraint. To the best of our knowledge, this is the first method that can achieve these two objectives simultaneously for large portfolio optimization.

Our first main contribution is establishing an equivalent unconstrained regression representation of the Markowitz optimization problem. This special regression representation is a novel finding made in this paper. There is existing literature that links regression to the mean-variance optimization problem. Two most closely related approaches are Britten-Jones (1999), who uses 1 as the response; and Brodie et al. (2009), who use the maximum expected return as the response. The issues with the two approaches are that, the regression in Britten-Jones (1999) calls for a challenging scaling to obtain the Markowitz portfolio; and Brodie et al. (2009) deal with a constrained regression, which is complicated and involves error and biases caused by replacing the constraint with the analogous sample version; see Remarks 1 and 2 for more detailed discussions. Our regression representation, in contrast, is on the one hand, *equivalent* to the original optimization problem, and on the other hand, *unconstrained*, so that it can be conveniently combined with high-dimensional regression techniques. See more details in Sections 2.1 and 2.2.

Our method further involves the following important aspects:

(i) Consistent estimation of the response in our regression representation, which directly links to consistent estimation of the maximum Sharpe ratio achieved by the tangency portfolio.

(ii) Proper sparse regression and rigorous analysis under our regression framework which possesses unique features.

(iii) The estimation in (i) and the sparse regression in (ii) constitute the core of MAXSER.

To implement MAXSER, we consider two scenarios, one without factor investing, the other with factor investing.

- Scenario I: When the asset pool consists only of individual assets. In this case, sparse regression with the consistently estimated response is directly applied to asset returns to estimate the optimal portfolio. See Section 2.3.
- Scenario II: When factors are also included in the asset pool. In this second case, we establish a framework that decomposes the optimal portfolio on all assets into the optimal portfolio on factors and that on idiosyncratic components, based on which we develop our estimator of the optimal full portfolio. See Section 2.4.

Under both scenarios, we theoretically prove the convergences of the expected return and risk of our estimated portfolio towards the theoretical maximum expected return and risk constraint, respectively. These properties guarantee that our portfolio can asymptotically achieve mean-variance efficiency as both the number of assets and sample size get large.

(iv) In practice, when solving for the sparse regression solution, the choice of the ℓ_1-norm constraint parameter is critical. We propose a new cross-validation procedure for choosing the tuning parameter. The method can help effectively control the risk, and constitutes another contribution of ours to the literature. The procedure can also be easily adapted to other portfolio optimization methods with norm constraint or short-sale constraint. See Section 2.5 for details.

The superior properties of our MAXSER portfolio are supported by simulation and empirical studies. We compare our strategy with various other methods including the plug-in portfolio, the equally weighted portfolio, the linear/nonlinear shrinkage portfolio of Ledoit and Wolf (2004) and Ledoit and Wolf (2017), and several other variations of mean-variance/GMV portfolios with constraints on portfolio weights. The complete simulation results are given in Section 3. Figure 3 shows the comparisons among the plug-in, nonlinear shrinkage and MAXSER portfolios. The added blue dashed lines are for our portfolio. We see that our portfolio **effectively controls the risk**. More importantly, the comparisons of Sharpe ratios show that our MAXSER portfolio ***nearly achieves the mean-variance efficiency*** and significantly outperforms others.

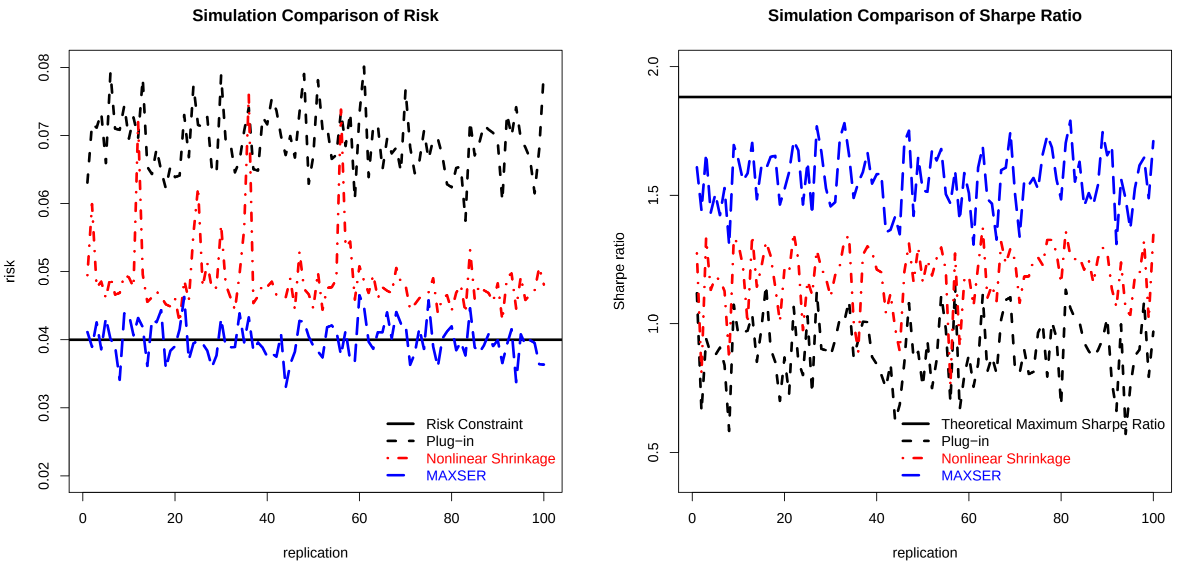

Figure 3. *Comparisons of our MAXSER portfolio with the plug-in and nonlinear shrinkage portfolios, based on the same simulated data as for Figures 1 and 2. The added blue dashed lines represent the MAXSER portfolio. The comparison is replicated 100 times.*

Comprehensive empirical studies are presented in Section 4. We have the following observations:

- MAXSER effectively controls the risk and yields high Sharpe ratios.
- Comparisons based on 50-year returns of DJIA 30 index constituents show that Scenario II with factor investing leads to substantially better performance in terms of Sharpe ratio than Scenario I without factor investing.
- Comparison of the Sharpe ratios of MAXSER (on both factors and stocks) and the Markowitz portfolio on factors shows that additionally investing in individual assets using our MAXSER strategy yields significant gain.
- Our cross-validation procedure for choosing the tuning parameter proves to be effective in controlling risk, not only for our own MAXSER, but also for other constrained portfolios. More importantly, the decisive advantage of MAXSER over various constrained portfolios confirms that the advantage of MAXSER is fundamentally due to its methodology rather than solely imposing constraints.

- We conduct statistical tests on Sharpe ratios. The test results based on 100 random stock pools drawn from S&P 500 index constituents, both without and with transaction costs taken into account, show that MAXSER has dominating advantage in terms of mean-variance efficiency.

2 The MAXSER Methodology

2.1 An Unconstrained Regression Representation

Suppose that we have a pool of N risky assets. Denote their (random excess) returns by $\boldsymbol{r} = (r_1, r_2, \ldots, r_N)'$, where for any vector $\boldsymbol{v}$, $\boldsymbol{v}'$ stands for the transpose of $\boldsymbol{v}$. Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ be the mean vector and covariance matrix of $\boldsymbol{r}$, respectively, and let $\boldsymbol{w}$ be a vector of portfolio weights on the assets. For a given level of risk constraint σ, the Markowitz optimization problem is

$$\arg\max_{\boldsymbol{w}} E(\boldsymbol{w}'\boldsymbol{r}) = \boldsymbol{w}'\boldsymbol{\mu} \quad \text{subject to} \quad \mathrm{Var}(\boldsymbol{w}'\boldsymbol{r}) = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} \leq \sigma^2. \tag{2.1}$$

If we denote by $\theta = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ the square of the maximum Sharpe ratio of the optimal portfolio, then the optimization problem (2.1) can be represented in its dual form with a return constraint $r^* = \sigma\sqrt{\theta}$:

$$\arg\min_{\boldsymbol{w}} \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} \quad \text{subject to} \quad \boldsymbol{w}'\boldsymbol{\mu} \geq r^*. \tag{2.2}$$

The optimal portfolio, $\boldsymbol{w}^*$, admits the following explicit expression:

$$\boldsymbol{w}^* = \frac{\sigma}{\sqrt{\theta}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}. \tag{2.3}$$

Because of the difficulties in estimating the mean and covariance matrix in high-dimensional situations, instead of working with the formula (2.3) and plugging in estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we propose a novel *equivalent* and *unconstrained* regression representation of Markowitz optimization.

Proposition 1. *The unconstrained regression*

$$\arg\min_{\boldsymbol{w}} E(r_c - \boldsymbol{w}'\boldsymbol{r})^2, \quad \textit{where} \quad r_c := \frac{1+\theta}{\theta} r^* \equiv \sigma\frac{1+\theta}{\sqrt{\theta}}, \tag{2.4}$$

is equivalent to the Markowitz optimizations (2.1) *and* (2.2).

Remark 1. *Brodie et al. (2009) use a constrained regression representation of Markowitz optimization* (2.2)*:*

$$\arg\min_{\boldsymbol{w}} E(r^* - \boldsymbol{w}'\boldsymbol{r})^2 \quad \textit{subject to} \quad \boldsymbol{w}'\boldsymbol{\mu} = r^*.$$

Such a constrained regression is theoretically difficult to analyze and numerically complicated to solve. To implement the regression, Brodie et al. (2009) replace the constraint $\boldsymbol{w}'\boldsymbol{\mu} = r^*$ *with its sample counterpart* $\boldsymbol{w}'\overline{\boldsymbol{r}} = r^*$*. Notice however that in the high-dimensional case,* $\overline{\boldsymbol{r}}$ *is* not *a consistent estimator of* $\boldsymbol{\mu}$ *(in the sense that* $||\overline{\boldsymbol{r}} - \boldsymbol{\mu}||_2 \nrightarrow 0$*).*

Remark 2. *Britten-Jones (1999) connects the estimation of the tangency portfolio with OLS regression with response 1. The regression yields a multiple of the sample (plug-in) tangency portfolio. Scaling the weights such that the weights add up to 1 recovers the plug-in tangency portfolio. Britten-Jones (1999) did not consider the Markowitz optimization problem* (2.1) *or* (2.2)*. To find the Markowitz portfolio for a given risk constraint* σ*, notice that, first of all, the plug-in tangency portfolio will be suboptimal in high dimensions (see equation* (1.1) *for the precise statement), and secondly, even if one knew the tangency portfolio, because it has a risk of* $\sqrt{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}/(\boldsymbol{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2}$*, a further scaling is needed and is challenging.*

We emphasize that our regression representation (2.4) is intrinsically different from existing regression representations in the literature for mean-variance portfolio estimation. Our representation is *unconstrained* and *equivalent* to the Markowitz problem. The elimination of constraint is particularly helpful for large portfolio selection, for which important techniques like sparse regression becomes directly applicable.

2.2 High-dimensional Issues & Sparse Regression

Proposition 1 translates the original Markowitz optimization problems (2.1) and (2.2) into an equivalent unconstrained regression problem. The next step is to make use of such a regression and estimate the optimal portfolio by replacing $\arg\min_{\boldsymbol{w}} E(r_c - \boldsymbol{w}'\boldsymbol{r})^2$ with its sample version

$$\arg\min_{\boldsymbol{w}} \frac{1}{T}\sum_{t=1}^{T} \left(r_c - \boldsymbol{w}'\boldsymbol{R_t}\right)^2, \tag{2.5}$$

where $\boldsymbol{R_t} = (R_{t1}, \ldots, R_{tN})'$, $t = 1, \ldots, T$, are T i.i.d. copies of the return vector $\boldsymbol{r}$. When the number of assets N is not small compared with T, (2.5) is a high-dimensional regression problem. It is now well known that in general it is *impossible* to consistently estimate the coefficients in a high-dimensional regression, and some sort of sparsity is necessary to turn the impossible into the possible. The most widely used sparsity assumption is boundedness on the ℓ_1-norm of the regression coefficients. In our case, this corresponds to assuming that $||\boldsymbol{w}^*||_1$ is bounded, where $||\boldsymbol{v}||_1 = \sum_{i=1}^{N} |v_i|$ for any $\boldsymbol{v} = (v_1, \ldots, v_N)'$. We adopt the sparse regression technique LASSO (Tibshirani (1996)), which leads to the following estimation of the optimal portfolio $\boldsymbol{w}^*$:

$$\boldsymbol{w}(r_c) := \arg\min_{\boldsymbol{w}} \frac{1}{T} \sum_{t=1}^{T} (r_c - \boldsymbol{w}'\boldsymbol{R_t})^2 \quad \text{subject to} \quad ||\boldsymbol{w}||_1 \le \lambda, \tag{2.6}$$

where λ is an ℓ_1-norm constraint tuning parameter. Note that the solution (2.6) is *infeasible* because the response r_c is unknown. Below we will develop a feasible solution with a consistent estimator of r_c. For clarity, let us suppose for the moment that r_c is known.

The solution (2.6) involves a tuning parameter λ. It can be easily seen that if λ is chosen to be larger than the ℓ_1-norm of the OLS solution (2.5), then one always gets the OLS solution. On the other hand, as λ varies from 0 to the ℓ_1-norm of the OLS solution, one gets a so-called solution path, each solution corresponding to an estimated portfolio. An algorithm called LARS (Efron et al. (2004)) has been developed to efficiently solve for the whole solution path.

Figure 4 shows the risks and Sharpe ratios of all estimated portfolios on the solution path based on simulated i.i.d. returns following multivariate normal distribution. The mean and covariance matrix are the parameters calibrated from idiosyncratic returns[2] from fitting Fama-French three factor model of 100 randomly chosen constituents of S&P 500 index; see Section 3.2 for details.

[2] In order to be consistent with our simulation and empirical studies, here we use simulated idiosyncratic returns to demonstrate the LASSO solution paths. See more details in Sections 2.4.3 and 2.5.3.

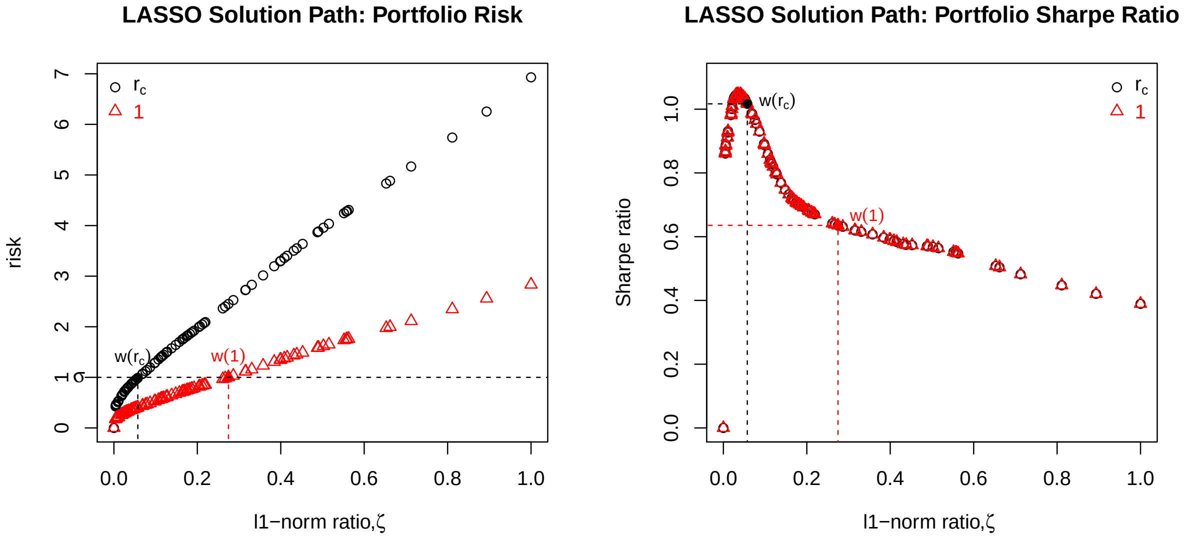

Figure 4. *Simulation comparisons of the risks and Sharpe ratios of all estimated portfolios on two LASSO solution paths, one with response r_c (setting $\sigma = 1$) and the other with response 1. The x-axis stands for the ratio between the ℓ_1-norms of a LASSO solution and the OLS solution. The risk and Sharpe ratio are true values based on the underlying mean and covariance matrix.*

In Figure 4, we show two LASSO solution paths, one with response r_c (setting $\sigma = 1$) and the other with response 1. The x-axis stands for the ratio of the ℓ_1-norm of a solution, $\boldsymbol{w}$, on the solution path relative to that of the OLS solution:

$$\zeta = \frac{||\boldsymbol{w}||_1}{||\boldsymbol{w}_{ols}||_1}. \tag{2.7}$$

For the whole range of ζ from 0 to 1, the left panel shows the risks of all estimated portfolios on the solution paths based on the two response values, and the right panel shows the corresponding Sharpe ratios. We observe that

- For each solution path, the risk is increasing with respect to the ℓ_1-norm ratio ζ. Take the solution path with the correct response r_c for example. When ζ is small, the risk is lower than the risk constraint level, and as ζ increases, the risk increases and eventually exceed the constraint level. Such a feature shows the importance of the choice of the ℓ_1-norm ratio ζ, or equivalently, that of the tuning parameter λ in practice.

- To identify the pursued portfolio from each solution path, we look for the portfolio with a risk closest to the risk constraint[3]. This leads to portfolios $\boldsymbol{w}(r_c)$ and $\boldsymbol{w}(1)$. The left panel shows that they correspond to very different ℓ_1-norm ratios. The ℓ_1-norm ratio of $\boldsymbol{w}(r_c)$ is lower than that of $\boldsymbol{w}(1)$, indicating that $\boldsymbol{w}(r_c)$ is sparser than $\boldsymbol{w}(1)$.

- Turning to the Sharpe ratio comparison, we see that the Sharpe ratio paths with different responses coincide with each other. However, because the portfolios $\boldsymbol{w}(r_c)$ and $\boldsymbol{w}(1)$ correspond to different values of ratio ζ, their Sharpe ratios differ significantly. The portfolio with the correct response r_c almost achieves the highest Sharpe ratio on the path, while the other portfolio with response 1 yields a much lower Sharpe ratio.

In summary, this illustration provides insight into the importance of (1) our corrected response r_c, and (2) the choice of tuning parameter. We will discuss the tuning parameter selection in Section 2.5.1 in more detail.

2.3 Scenario I: When Asset Pool Includes Individual Assets Only

Let us first consider the situation where the asset pool only includes individual assets (e.g., stocks).

2.3.1 Estimating the Maximum Sharpe Ratio and the Regression Response

In our regression representation, the response r_c is unknown[4] and needs to be estimated. The estimation of the response r_c is closely related to the estimation of the maximum Sharpe ratio, which has been considered in Kan and Zhou (2007). It is shown that the square of the plug-in Sharpe ratio follows a non-centralized F-distribution; see equation (49) therein. The result implies that the plug-in Sharpe ratio is heavily biased when the number of assets, N, is not negligible compared with the sample size T. Suppose that $\boldsymbol{R_t} = (R_{t1}, \ldots, R_{tN})'$, $t = 1, \ldots, T$, are T i.i.d. copies of the (excess) return. Let $\boldsymbol{R} = (\boldsymbol{R_1}, ..., \boldsymbol{R_T})'$ be the $T \times N$ observation matrix, and $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ be the corresponding sample mean and sample covariance

[3] Alternatively, one may consider using Sharpe ratio to identify the "optimal" ℓ_1-norm ratio ζ. Notice however that the estimation of Sharpe ratio is less accurate than that of risk. What's more, if the response is not r_c, then even if one could identify the portfolio with the highest Sharpe ratio on the solution path, to obtain the Markowitz portfolio with risk σ, an additional scaling is needed, which is even more challenging than what is discussed in Remark 2.

[4] In the dual form of Markowitz problem where return constraint r^* is given, the response is still unknown because it also depends on θ, the square of the maximum Sharpe ratio.

matrix, respectively. The following unbiased estimator is proposed in Kan and Zhou (2007):

$$\widehat{\theta} := \frac{(T-N-2)\widehat{\theta}_s - N}{T}, \tag{2.8}$$

where $\widehat{\theta}_s := \widehat{\boldsymbol{\mu}}'\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}$ is the sample estimate of $\theta = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Under the situation where $N/T \to \rho \in (0,1)$, we establish the following

Proposition 2. *Suppose that the maximum Sharpe ratio is bounded. Under normality assumption on returns and assuming that $N/T \to \rho \in (0,1)$, we have*

$$|\widehat{\theta} - \theta| \xrightarrow{P} 0. \tag{2.9}$$

Consequently,

$$\widehat{r_c} := \sigma \frac{1+\widehat{\theta}}{\sqrt{\widehat{\theta}}} \tag{2.10}$$

satisfies that

$$|\widehat{r_c} - r_c| \xrightarrow{P} 0. \tag{2.11}$$

We emphasize that our estimation of $\theta = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is *not* via consistently estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, which would be impossible without imposing strong structural assumptions on them. The challenge is due to error accumulation. Take the estimation of $\boldsymbol{\mu}$ for example. Each entry of $\boldsymbol{\mu}$ can be estimated with an error of order $1/\sqrt{T}$, however, because the dimension of $\boldsymbol{\mu}$ is N, the total estimation error under ℓ_2-norm is of order $\sqrt{N/T}$, which is $O(1)$ under our assumption that $N/T \to \rho \in (0,1)$. In contrast, we estimate θ directly. Our $\widehat{\theta}$ is consistent, and actually, as can be seen from the proof, the error $|\widehat{\theta} - \theta|$ is of order $1/\sqrt{T}$.

2.3.2 A LASSO-type Estimator

On the basis of our unconstrained regression representation (2.4), we aim to estimate the optimal portfolio $\boldsymbol{w}^*$ for a given risk constraint σ in the high-dimensional setting where N and T are both large and the optimal portfolio has a bounded ℓ_1-norm. With the consistent estimation of r_c in Proposition 2, following (2.6) we now construct our feasible LASSO-type estimator $\widehat{\boldsymbol{w}^*} = \left(\widehat{w_1^*}, ..., \widehat{w_N^*}\right)'$ as follows:

$$\widehat{\boldsymbol{w}^*} = \arg\min_{\boldsymbol{w}} \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{r_c} - \boldsymbol{w}'\boldsymbol{R_t}\right)^2 \quad \text{subject to} \quad ||\boldsymbol{w}||_1 \leq \lambda. \tag{2.12}$$

Before we give the theoretical properties of $\widehat{\boldsymbol{w}^*}$, we list the assumptions that will be needed.

Assumption:

A1 The (excess) return $\boldsymbol{r} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$;

A2 There exists $M < \infty$ such that $\max\left(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \max_{1\leq j\leq N} |\mu_j|\right) \leq M$;

A3 There exists $L < \infty$ such that $\max_{1\leq j\leq N} |\sigma_{jj}| \leq L$, where $\sigma_{jk} := \boldsymbol{\Sigma}(j,k)$;

A4 $||\boldsymbol{w}^*||_1 \leq \lambda$ for some constant λ;

A5 The number of assets N and the sample size T satisfy that $\rho_T := N/T \to \rho \in (0,1)$.

Remark 3. *About Assumption A1, in this paper we focus on the most fundamental form of the mean-variance problem, so we assume normal distribution of returns. Numerically, we find that our proposed method works well even when heavy-tailedness is present. Extensions incorporating heteroscedasticity and heavy-tailedness will be studied in subsequent papers.*

Remark 4. *Assumption A2 is equivalent to the common belief that the maximum Sharpe ratio and asset expected returns are bounded. We emphasize that we do not require the eigenvalues of* $\boldsymbol{\Sigma}$ *to be bounded, and consequently factor structure in returns is allowed. Nor do we require the eigenvalues of* $\boldsymbol{\Sigma}$ *to be bounded from below.*

Remark 5. *Assumption A4, the boundedness of* $||\boldsymbol{w}^*||_1$*, is our sparsity requirement on the optimal portfolio. Note that Assumption A4 does not require most weights to be zero. For example, it does not rule out value-weighted portfolios. In addition, we will see in Section 2.4 that when factor investing is allowed, the sparsity requirement will be significantly relaxed and imposed only upon the optimal portfolio on idiosyncratic components.*

Remark 6. *Assumption A5 says that we are in a high-dimensional setting where the number of assets* N *and the sample size* T *proportionally grow up to infinity. We require the sample size* T *to be larger than the dimension* N *only because we need to take inverse of the sample covariance matrix in estimating the maximum Sharpe ratio; see Proposition 2.*

2.3.3 Main Result I: MAXSER without Factor Investing

We now state our first main result, which establishes the asymptotic optimality of our MAXSER portfolio when the asset pool consists only of individual assets.

Theorem 1. *Under Assumptions A1* $\sim$ *A5, the MAXSER portfolio* $\widehat{\boldsymbol{w}^*}$ *defined in* (2.12) *with* $\widehat{r_c}$ *given by* (2.10) *satisfies that, as* $N \to \infty$,

$$|\boldsymbol{\mu}'\widehat{\boldsymbol{w}^*} - r^*| \xrightarrow{P} 0, \tag{2.13}$$

and

$$\left|\sqrt{\widehat{\boldsymbol{w}^*}'\boldsymbol{\Sigma}\widehat{\boldsymbol{w}^*}} - \sigma\right| \xrightarrow{P} 0. \tag{2.14}$$

Theorem 1 guarantees that our MAXSER portfolio $\widehat{\boldsymbol{w}^*}$ asymptotically (i) achieves the maximum expected return and (ii) satisfies the risk constraint. An immediate implication is that $\widehat{\boldsymbol{w}^*}$ approaches the mean-variance efficiency.

2.4 Scenario II: When Factor Investing Is Allowed

2.4.1 The Optimal Portfolio: A Factor-Idiosyncratic Component Separation

In addition to individual assets, oftentimes factors are also included in the investment asset pool. Motivated by such consideration, we now illustrate the implementation of MAXSER when factor investing is allowed. Consider the following model:

$$r_i = \alpha_i + \sum_{j=1}^{K} \beta_{ij} f_j + e_i := \sum_{j=1}^{K} \beta_{ij} f_j + u_i, \qquad i = 1, \ldots, N, \tag{2.15}$$

where f_j's are factor returns, β_{ij}'s represent individual stock sensitivities to the factors, and e_i's are the remaining errors in the model that are independent of the factor returns (f_j). Unlike the approximate factor model where the "idiosyncratic returns" $(u_i = \alpha_i + e_i)$ are assumed to have no factor structure, here in (2.15) we allow (u_i) to still admit factor structures, and in the below, we shall still refer to (u_i) as the idiosyncratic returns. As to the factors, they can be any well-recognized factors like Fama-French three factors or other factors identified in the large literature of asset pricing (see, e.g., Jegadeesh and Titman (2001) and Korajczyk and Sadka (2008)). They can also be statistical factors (identified based on a separate set of historical returns of a larger pool of assets).

Model (2.15) can be written in a compact form as

$$\boldsymbol{r} = \boldsymbol{\beta}\boldsymbol{f} + \boldsymbol{u}, \tag{2.16}$$

where $\boldsymbol{\beta} = (\beta_{ij})_{N\times K}$, $\boldsymbol{f} = (f_1, \ldots, f_K)'$, and $\boldsymbol{u} = (u_1, \ldots, u_N)'$. Let $\boldsymbol{\mu}_f$ be the mean of the factor returns, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$, the mean of idiosyncratic return $\boldsymbol{u}$. Let $\boldsymbol{\Sigma}_f$ and $\boldsymbol{\Sigma}_u$ be the covariance matrix of factor and idiosyncratic returns, respectively. Then the return vector $\boldsymbol{r}$ has the following mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\mu} = \boldsymbol{\beta}\boldsymbol{\mu}_f + \boldsymbol{\alpha}, \qquad \boldsymbol{\Sigma} = \boldsymbol{\beta}\boldsymbol{\Sigma}_f\boldsymbol{\beta}' + \boldsymbol{\Sigma}_u. \tag{2.17}$$

Denote the mean and covariance matrix of the returns on the full pool of factors and assets by $\boldsymbol{\mu}_{all}$ and $\boldsymbol{\Sigma}_{all}$, respectively:

$$\boldsymbol{\mu}_{all} = \begin{pmatrix} \boldsymbol{\mu}_f \\ \boldsymbol{\mu} \end{pmatrix}, \quad \boldsymbol{\Sigma}_{all} = \begin{pmatrix} \boldsymbol{\Sigma}_f & \boldsymbol{\beta}'\boldsymbol{\Sigma}_f \\ \boldsymbol{\Sigma}_f\boldsymbol{\beta} & \boldsymbol{\Sigma} \end{pmatrix}. \tag{2.18}$$

We make the following assumptions:

Assumption:

B1 The number of factors included, K, is bounded;

B2 There exists $M < \infty$ such that $\max\left(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\alpha}, \max_{1\leq j\leq N} |\alpha_j|\right) \leq M$;

B3 There exists $L < \infty$ such that $\max_{1\leq i\leq N} |\sigma_{ii}| \leq L$, where $\sigma_{ij} := \boldsymbol{\Sigma}_u(i,j)$.

Remark 7. *Similarly to Assumption A2, for $\boldsymbol{\Sigma}_u$, we do not require its eigenvalues to be bounded, in other words, we do not require $(f_1, \ldots, f_K)'$ to be the full set of factors, and $(u_1, \ldots, u_N)'$ can still have factor structure. Our recommendation is to only include a small number of strong factors. This also justifies why we still include $\boldsymbol{\alpha}$ in the model.*

With the K factors, our plan is to invest in not only the N assets but also the K factors. Such a plan is straightforward to implement when the factors are taken to be tradable risk factors like the Fama-French factors and many others (see, e.g., the supplementary file of Feng et al. (2017)), and is also feasible when the factors are statistical factors as mentioned above.

We aim to find an optimal allocation $(w_1^f, \ldots, w_K^f; w_1, \ldots, w_N) := (\boldsymbol{w}_f, \boldsymbol{w})$, where $\boldsymbol{w}_f$ and $\boldsymbol{w}$ represent the weight vectors on the K factors and N individual assets, respectively. The following result shows that finding such an optimal allocation can be decomposed into three steps:

(i) Find the optimal portfolio on the factors with 1 unit of risk (denoted by $\boldsymbol{w}_f^*$),

(ii) Find the optimal portfolio on the idiosyncratic components with 1 unit of risk (denoted by $\boldsymbol{w}_u^*$), and

(iii) Suitably combine these two portfolios.

Specifically, we have the following

Proposition 3. *For any given risk constraint level σ, the optimal portfolio $\boldsymbol{w}_{all} := (\boldsymbol{w}_f, \boldsymbol{w})$ is given by*

$$\sigma\left(\sqrt{\frac{\theta_f}{\theta_{all}}}\boldsymbol{w}_f^* - \sqrt{\frac{\theta_u}{\theta_{all}}}\boldsymbol{\beta}'\boldsymbol{w}_u^*, \quad \sqrt{\frac{\theta_u}{\theta_{all}}}\boldsymbol{w}_u^*\right),$$

where $\theta_f = \boldsymbol{\mu}_f'\boldsymbol{\Sigma}_f^{-1}\boldsymbol{\mu}_f$, $\theta_u = \boldsymbol{\alpha}'\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\alpha}$, and $\theta_{all} = \boldsymbol{\mu}_{all}'\boldsymbol{\Sigma}_{all}^{-1}\boldsymbol{\mu}_{all}$ are the squared maximum Sharpe ratios of portfolios on the factors, the idiosyncratic components, and the full set of factors and individual assets, respectively. Moreover, $\boldsymbol{w}_f^*$ and $\boldsymbol{w}_u^*$ admit the following explicit expressions:

$$\boldsymbol{w}_f^* = \frac{1}{\sqrt{\theta_f}}\boldsymbol{\Sigma}_f^{-1}\boldsymbol{\mu}_f, \quad \boldsymbol{w}_u^* = \frac{1}{\sqrt{\theta_u}}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\alpha}. \tag{2.19}$$

Remark 8. *In the case when $\alpha_i = 0$ for all $i = 1, \ldots, N$, we have $\theta_u = 0$ and the optimal allocation is given by $\left(\sigma\boldsymbol{w}_f^*, \boldsymbol{0}\right)$. In other words, the optimal portfolio will be fully invested in factors. However, this is unlikely in practice especially when one wants to only include a small number of strong factors. If, on the other hand, one intends to incorporate a large number of factors, then estimating the optimal portfolio on the factors will be of high-dimensional nature, in which case our strategy in Section 2.3 can be applied.*

According to Proposition 3, in order to estimate the optimal portfolio $(\boldsymbol{w}_f, \boldsymbol{w})$, we need to estimate θ_f, θ_u, $\boldsymbol{w}_f^*$ and $\boldsymbol{w}_u^*$. We will deal with them one by one, starting with the estimation of the maximum Sharpe ratios.

2.4.2 Estimating the Maximum Sharpe Ratios

Suppose that $\boldsymbol{R_t} = (R_{t1}, \ldots, R_{tN})'$ and $\boldsymbol{F_t} = (F_{t1}, \ldots, F_{tK})'$, $t = 1, \ldots, T$, are T i.i.d. copies of the excess return $\boldsymbol{r}$ and the factor excess return $\boldsymbol{f}$, respectively. Let $\boldsymbol{R} = (\boldsymbol{R_1}, ..., \boldsymbol{R_T})'$ and $\boldsymbol{F} = (\boldsymbol{F_1}, \ldots, \boldsymbol{F_T})'$. We estimate the coefficient $\boldsymbol{\beta}$ in (2.16) by regressing $\boldsymbol{R}$ on $\boldsymbol{F}$. Denote by $\widehat{\boldsymbol{\beta}}$ the estimated beta matrix. Correspondingly, let $\widehat{\boldsymbol{U}} = (\widehat{\boldsymbol{U}_1}, \ldots, \widehat{\boldsymbol{U}_T})' = \boldsymbol{R} - \boldsymbol{F}\widehat{\boldsymbol{\beta}}$ be the estimator of $\boldsymbol{U} = \boldsymbol{R} - \boldsymbol{F}\boldsymbol{\beta}$. Denote by $\widehat{\boldsymbol{\mu}}_f$ and $\widehat{\boldsymbol{\Sigma}}_f$ the sample mean and sample covariance matrix of the factor return $\boldsymbol{F}$, respectively.

There are three Sharpe ratios that need to be estimated under the current setting: $\sqrt{\theta_f}$, $\sqrt{\theta_u}$ and $\sqrt{\theta_{all}}$, which are the maximum Sharpe ratios on factors, idiosyncratic components and all assets, respectively. Because the number of factors included is bounded, $\sqrt{\theta_f}$ can be consistently estimated by its plug-in estimator:

$$\left|\sqrt{\widehat{\theta}_f} - \sqrt{\theta_f}\right| \xrightarrow{P} 0 \quad \text{as} \quad T \to \infty, \quad \text{where} \quad \widehat{\theta}_f = \widehat{\boldsymbol{\mu}}_f'\widehat{\boldsymbol{\Sigma}}_f^{-1}\widehat{\boldsymbol{\mu}}_f. \tag{2.20}$$

The remaining two Sharpe ratios, $\sqrt{\theta_u}$ and $\sqrt{\theta_{all}}$, both involve a large number of assets, and we need to adjust for the bias in the plug-in estimator. In parallel with Proposition 2, we have the following

Proposition 4. *Define the following estimator of* θ_{all}*:*

$$\widehat{\theta}_{all} := \frac{(T-N-K-2)\widehat{\theta}_{s,all} - N - K}{T}, \tag{2.21}$$

where $\widehat{\theta}_{s,all} := \widehat{\boldsymbol{\mu}}'_{all}\widehat{\boldsymbol{\Sigma}}_{all}^{-1}\widehat{\boldsymbol{\mu}}_{all}$ *is the sample estimate of* θ_{all}*. Suppose that* θ_{all} *is bounded. Under normality assumption on returns and assuming that* $N/T \to \rho \in (0,1)$*, we have*

$$\left|\sqrt{\widehat{\theta}_{all}} - \sqrt{\theta_{all}}\right| \xrightarrow{P} 0.$$

There is one last Sharpe ratio to be estimated, $\sqrt{\theta_u}$, the maximum Sharpe ratio on the idiosyncratic components. This quantity is a bit trickier to deal with because the idiosyncratic return $\boldsymbol{U}$ is not observable. A natural idea is to work with $\widehat{\boldsymbol{U}}$, the estimated idiosyncratic return. However, it can be shown that an estimator similar to (2.21) applied to $\widehat{\boldsymbol{U}}$ will be biased.

The solution to the aforementioned difficulty lies in the relationship among θ_f, θ_u and θ_{all}. Based on model (2.16), one can show that

$$\theta_{all} = \theta_f + \theta_u. \tag{2.22}$$

By (2.20) and Proposition 4, both θ_f and θ_u can be consistently estimated, so we get the following

Proposition 5. *Define*

$$\widehat{\theta}_u := \widehat{\theta}_{all} - \widehat{\theta}_f.$$

Under the assumptions of Proposition 4, we have

$$\left|\sqrt{\widehat{\theta}_u} - \sqrt{\theta_u}\right| \xrightarrow{P} 0.$$

Therefore for $r_c := (1+\theta_u)/\sqrt{\theta_u}$*,*

$$\widehat{r_c} := \frac{1+\widehat{\theta}_u}{\sqrt{\widehat{\theta}_u}} \tag{2.23}$$

satisfies that

$$|\widehat{r_c} - r_c| \xrightarrow{P} 0.$$

2.4.3 Estimating the Optimal Portfolio on Idiosyncratic Components

The optimal portfolio on the idiosyncratic components, $\boldsymbol{w}_u^*$, solves the following Markowitz optimization problem:

$$\arg\max_{\boldsymbol{w}} \boldsymbol{\alpha}'\boldsymbol{w} \quad \text{subject to} \quad \boldsymbol{w}'\boldsymbol{\Sigma}_u\boldsymbol{w} \leq 1. \tag{2.24}$$

The optimal portfolio yields an expected return of

$$r_u^* = \sqrt{\theta_u}, \tag{2.25}$$

which equals the maximum Sharpe ratio of portfolios on (u_i). Following our regression representation (2.4) in Section 2.1, we estimate $\boldsymbol{w}_u^*$ based on the following:

$$\arg\min_{\boldsymbol{w}} E\left(r_c - \boldsymbol{u}'\boldsymbol{w}\right)^2, \quad \text{where } r_c := \frac{1+\theta_u}{\theta_u} r_u^* = \frac{1+\theta_u}{\sqrt{\theta_u}}. \tag{2.26}$$

One major difference here is that the idiosyncratic returns are not observable, and we will have to rely on the estimated idiosyncratic return $\widehat{\boldsymbol{U}}$. More explicitly, based on both estimated idiosyncratic return $\widehat{\boldsymbol{U}}$ and estimated $\widehat{r_c}$ in (2.23), we define our estimator of $\boldsymbol{w}_u^*$ analogous to (2.12) as the following

$$\widehat{\boldsymbol{w}_u^*} = \arg\min_{\boldsymbol{w}} \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{r_c} - \boldsymbol{w}'\widehat{\boldsymbol{U}_t}\right)^2 \quad \text{subject to} \quad ||\boldsymbol{w}||_1 \leq \lambda, \tag{2.27}$$

where λ is the ℓ_1-norm tuning parameter. To give the theoretical properties of $\widehat{\boldsymbol{w}_u^*}$, the following assumptions will be needed.

Assumption:

C1 $\boldsymbol{f} \sim N(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$, $\boldsymbol{u} \sim N(\boldsymbol{\alpha}, \boldsymbol{\Sigma}_u)$;

C2 $||\boldsymbol{w}_u^*||_1 \leq \lambda$ for some constant λ;

C3 The number of assets N and the sample size T satisfy that $\rho_T := N/T \to \rho \in (0,1)$.

Remark 9. *Assumption C2 is our sparsity requirement in this setting when factor investing is allowed. We emphasize that the sparsity assumption is put on the optimal portfolio on the* ***idiosyncratic components****. This amounts to say, the mean-variance efficiency is achieved by an addition of a sparse portfolio of stocks to the optimal portfolio on factors.*

We are now ready to give the asymptotic properties of the portfolio $\widehat{\boldsymbol{w}_u^*}$. Note that because $\widehat{\boldsymbol{U}}$ contains estimation errors, Theorem 1 does not readily apply to this case.

Proposition 6. *Under Assumptions B1 $\sim$ B3 and C1 $\sim$ C3, we have as $N \to \infty$,*

$$|\boldsymbol{\alpha}'\widehat{\boldsymbol{w}_u^*} - \boldsymbol{\alpha}'\boldsymbol{w}_u^*| \xrightarrow{P} 0, \tag{2.28}$$

and

$$\left|\sqrt{\widehat{\boldsymbol{w}_u^*}'\boldsymbol{\Sigma}_u\widehat{\boldsymbol{w}_u^*}} - 1\right| \xrightarrow{P} 0. \tag{2.29}$$

Proposition 6 states that the portfolio $\widehat{\boldsymbol{w}_u^*}$ asymptotically attains the maximum expected return and carries a risk that is close to the given risk constraint, in this case, 1.

2.4.4 Main Result II: MAXSER with Factor Investing

By far we have achieved consistency in estimating $\boldsymbol{w}_u^*$, θ_u and θ_f. There is one more item to be estimated, $\boldsymbol{w}_f^*$, the optimal portfolio on factors (with risk equal to 1). This is easy because the number of factors included is bounded, and the simple "plug-in" estimator works. Specifically, $\widehat{\boldsymbol{w}_f^*} := \frac{1}{\sqrt{\widehat{\theta}_f}}\widehat{\boldsymbol{\Sigma}}_f^{-1}\widehat{\boldsymbol{\mu}}_f$ will be a consistent estimator of $\boldsymbol{w}_f^*$. Combining these results with Proposition 3, we obtain the following main result for our estimator of the optimal full portfolio $\widehat{\boldsymbol{w}_{all}}$.

Theorem 2. *Define our estimator of the optimal full portfolio $\boldsymbol{w}_{all}$ as*

$$\widehat{\boldsymbol{w}_{all}} := (\widehat{\boldsymbol{w}}_f, \widehat{\boldsymbol{w}}) = \sigma\left(\sqrt{\frac{\widehat{\theta}_f}{\widehat{\theta}_{all}}}\widehat{\boldsymbol{w}_f^*} - \sqrt{\frac{\widehat{\theta}_u}{\widehat{\theta}_{all}}}\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{w}_u^*},\ \sqrt{\frac{\widehat{\theta}_u}{\widehat{\theta}_{all}}}\widehat{\boldsymbol{w}_u^*}\right). \tag{2.30}$$

Under Assumptions B1 $\sim$ B3 and C1 $\sim$ C3, as $N \to \infty$, we have

$$|\widehat{\boldsymbol{w}_{all}}'\boldsymbol{\mu}_{all} - r^*| \xrightarrow{P} 0, \quad \textit{and} \quad |\widehat{\boldsymbol{w}_{all}}'\boldsymbol{\Sigma}_{all}\widehat{\boldsymbol{w}_{all}} - \sigma^2| \xrightarrow{P} 0. \tag{2.31}$$

Theorem 2 guarantees that our MAXSER portfolio with factor investing can again asymptotically yield the maximum expected return and meanwhile satisfy the risk constraint, and consequently, achieve mean-variance efficiency.

2.5 Practical Implementation of MAXSER

2.5.1 Choosing λ in (2.12) or (2.27)

To implement MAXSER, it is important to choose λ in (2.12) or (2.27). Because one of our goals is to meet the risk constraint, also in light of Figure 4, we choose λ such that the estimated portfolio possesses a risk that is close to the given risk constraint.

In practice, we do not know the underlying covariance matrix $\mathbf{\Sigma}$ or $\mathbf{\Sigma}_{all}$. To circumvent this difficulty, we use a cross-validation method. Specifically, for a 10-fold cross-validation procedure, we randomly split the sample into 10 groups to form 10 validation sets. For each validation set, the training set is taken to be the rest of the observations. Next, for each such training set i, let the ℓ_1-norm ratio ζ (defined in (2.7)) vary from 0 to 1 to obtain the whole solution path $(\widehat{\boldsymbol{w}_\zeta^*})_{0\leq\zeta\leq 1}$ $((\widehat{\boldsymbol{w}_{all,\zeta}^*})_{0\leq\zeta\leq 1}$ under Scenario II), and find the value $\zeta(i)$ such that the estimated portfolio minimizes the difference between the risk computed using the validation set and the given risk constraint. Let $\lambda(i) = ||\widehat{\boldsymbol{w}_{\zeta(i)}^*}||_1$. The ultimate $\widehat{\lambda}$ is then taken to be the average of $(\lambda(i),\ i = 1, \ldots, 10)$.

To our knowledge, the above cross-validation procedure for determining the constraint parameter is new and constitutes another contribution of ours to the literature. By taking the parameter selection criterion to be the risk, our method of choosing the tuning parameter effectively helps control out-of-sample risk. The method can be easily adapted to other portfolio optimization methods with norm constraint or short-sale constraint. In our numerical studies below, we apply such a method to other constrained portfolios to help them control the risk and compare their performance with ours.

2.5.2 Adjustment of $\widehat{\theta}$, $\widehat{\theta}_u$ and $\widehat{\theta}_{all}$

Kan and Zhou (2007) notice that $\widehat{\theta}$ in (2.8), the unbiased estimator of the square of maximum Sharpe ratio, often takes negative values, and they propose the adjusted estimator that improves over the unbiased one:

$$\widehat{\theta}_{adj} = \frac{(T-N-2)\widehat{\theta}_s - N}{T} + \frac{2(\widehat{\theta}_s)^{N/2}(1+\widehat{\theta}_s)^{-(T-2)/2}}{TB_{\widehat{\theta}_s/(1+\widehat{\theta}_s)}(N/2,(T-N)/2)}, \tag{2.32}$$

where, recall that, $\widehat{\theta_s}$ is the plug-in estimators of θ, and

$$B_x(a,b) = \int_0^x y^{a-1}(1-y)^{b-1}dy.$$

Under the second setting that allows for factor investing, we adopt the following adjustment of $\widehat{\theta}_{all}$:

$$\widehat{\theta}_{all,adj} = \frac{(T-N-K-2)\widehat{\theta}_{s,all} - N - K}{T} + \frac{2(\widehat{\theta}_{s,all})^{(N+K)/2}(1+\widehat{\theta}_{s,all})^{-(T-2)/2}}{TB_{\widehat{\theta}_{s,all}/(1+\widehat{\theta}_{s,all})}((N+K)/2,(T-N-K)/2)}, \tag{2.33}$$

where, recall that, $\widehat{\theta}_{s,all}$ is the plug-in estimator of θ_{all}. The adjusted $\widehat{\theta}_u$ is $\widehat{\theta}_{u,adj} := \widehat{\theta}_{all,adj} - \widehat{\theta}_f$.

2.5.3 Implementation Steps

Scenario I: When the asset pool consists only of individual assets

In this case, our method consists of the following steps:

Step 1 Estimate the square of the maximum Sharpe ratio by $\widehat{\theta}_{adj}$ in (2.32), and compute the response $\widehat{r_c}$ in (2.10);

Step 2 Choose λ by cross-validation according to the procedure described in Section 2.5.1. Denote the chosen value by $\widehat{\lambda}$;

Step 3 Set λ in (2.12) to be $\widehat{\lambda}$ and solve for $\widehat{\boldsymbol{w}^*}$, the MAXimum-Sharpe-ratio Estimated sparse Regression (MAXSER) portfolio.

Scenario II: When factors are included in the asset pool

Incorporating factor investing, MAXSER is implemented as follows:

Step 1 Perform OLS regressions of asset returns $\boldsymbol{R}$ on factor returns $\boldsymbol{F}$ to obtain $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{U}}$;

Step 2 Compute the estimates of the square of the maximum Sharpe ratios $\widehat{\theta}_f$, $\widehat{\theta}_{all,adj}$ and $\widehat{\theta}_{u,adj}$, and compute the response $\widehat{r_c}$ in (2.23);

Step 3 Choose λ by cross-validation according to the procedure described in Section 2.5.1. Denote the chosen value by $\widehat{\lambda}$;

Step 4 Set λ in (2.27) to be $\widehat{\lambda}$ and solve for $\widehat{\boldsymbol{w}_u^*}$;

Step 5 Compute $\widehat{\boldsymbol{w}_f^*}$ and plug in the estimates from the previous steps into (2.30) to obtain the MAXSER portfolio $\widehat{\boldsymbol{w}_{all}}$.

3 SIMULATION STUDIES

In this section, we examine and compare the performance of MAXSER with various competing strategies. As the empirical studies on DJIA 30 index constituents in Sections 4.2.3 and 4.2.4 show, it is beneficial to invest in both stocks and (Fama-French three) factors. For this reason, in the following simulation studies, the simulated asset pool includes both stocks and factors. Parameters for generating returns are calibrated from S&P 500 index constituents and Fama-French three factors; see Section 3.2 for details.

3.1 Methods to be Compared with

In addition to the plug-in and nonlinear shrinkage portfolios that we discussed in the Introduction, we include various other strategies in our comparisons. The complete list is given in Table 1.

Table 1

List of competing portfolios and their abbreviations. "MV" stands for mean-variance portfolio, and "GMV" stands for global minimum variance portfolio.

Portfolio	Abbreviation
Plug-in MV on factors	Factor
Three-fund portfolio by Kan and Zhou (2007)	KZ
MV/GMV with different covariance matrix estimates	
MV with sample cov	MV-P
MV with linear shrinkage cov	MV-LS
MV with nonlinear shrinkage cov	MV-NLS
MV with nonlinear shrinkage cov adjusted for factor models	MV-NLSF
GMV with linear shrinkage cov	GMV-LS
GMV with nonlinear shrinkage cov	GMV-NLS
MV with no-short-sale constraint	
MV with sample cov & no-short-sale constraint	MV-P-NSS
MV with linear shrinkage cov & no-short-sale constraint	MV-LS-NSS
MV with nonlinear shrinkage cov & no-short-sale constraint	MV-NLS-NSS
MV with short-sale constraint & cross-validation	
MV with sample cov & short-sale-CV	MV-P-SSCV
MV with linear shrinkage cov & short-sale-CV	MV-LS-SSCV
MV with nonlinear shrinkage cov & short-sale-CV	MV-NLS-SSCV

Among the portfolios under comparison, a special one is the portfolio Factor, which is the Markowitz portfolio on factors. Specifically, recall that $\widehat{\boldsymbol{\mu}}_f$ and $\widehat{\boldsymbol{\Sigma}}_f$ are the sample mean and sample covariance matrix of the factor returns, respectively. The Factor portfolio is defined as follows

$$\widehat{\boldsymbol{w}}_{Fac} = \frac{\sigma}{\sqrt{\widehat{\boldsymbol{\mu}}_f' \widehat{\boldsymbol{\Sigma}}_f^{-1} \widehat{\boldsymbol{\mu}}_f}} \widehat{\boldsymbol{\Sigma}}_f^{-1} \widehat{\boldsymbol{\mu}}_f \left(= \sigma \widehat{\boldsymbol{w}_f^*} \right). \tag{3.1}$$

This portfolio is special in the sense that it only involves a small number of assets (three in our case). Consequently, the plug-in formula (3.1) indeed gives a nearly optimal portfolio. Including such a portfolio in the comparison would reveal whether there is benefit to invest in additional individual assets.

As to the other portfolios, the MV/GMV portfolios are constructed by replacing covariance matrix with the sample/linear shrinkage (Ledoit and Wolf (2004))/nonlinear shrinkage/nonlinear shrinkage adjusted for factor model (Ledoit and Wolf (2017)) estimators in the formulas of MV/GMV portfolio weights, respectively. Details about the portfolio "KZ"[5] can be found in Kan and Zhou (2007).

In addition, we construct portfolios with either no-short-sale or short-sale constraint on portfolio weights. The "MV-P-NSS", "MV-LS-NSS" and "MV-NLS-NSS" portfolios are with *no-short-sale constraint*, and are using the sample/linear shrinkage/nonlinear shrinkage covariance matrix, respectively. More generally, the MV portfolios with *short-sale constraint*[6], "MV-P-SSCV", "MV-LS-SSCV" and "MV-NLS-SSCV", are having short position threshold determined by the cross-validation procedure that we proposed in Section 2.5.1. In such a way, these portfolios enjoy the same benefit in terms of risk control as our MAXSER portfolio. We include these portfolios to demonstrate the effectiveness of our cross-validation procedure in controlling risk, and more importantly, to show that the advantage of MAXSER is not

[5] Following Kan and Zhou (2007), the risk aversion is set to be 3. Note that such a risk aversion is not designed to meet a given risk constraint, hence in the following we will not comment on the risk of portfolio KZ.

[6] Specifically, suppose that $\widehat{\boldsymbol{\mu}}_{all}$ is the sample mean of returns on stocks and factors, and $\widehat{\boldsymbol{\Sigma}}_{all}$ is an estimate of the covariance matrix, which can be the sample/linear shrinkage/nonlinear shrinkage covariance matrix. The MV portfolio with short-sale constraint is solved by

$$\arg\max_{\boldsymbol{w}} \boldsymbol{w}' \widehat{\boldsymbol{\mu}}_{all} \quad \text{subject to} \quad \boldsymbol{w}' \widehat{\boldsymbol{\Sigma}}_{all} \boldsymbol{w} \le \sigma^2 \text{ and } w_i > -\lambda_{SS} \text{ for all } i = 1, \ldots, N. \tag{3.2}$$

Here $\lambda_{SS} > 0$ is the short position threshold determined via a 10-fold cross-validation as follows: Split the sample into 10 groups of validation sets. For each validation set, its corresponding training set consists of the rest of the observations. Next, for each training set, solve the optimization (3.2) for a sequence of λ_{SS} to get a solution path, and find the value of λ_{SS} such that the difference between the risk on validation set and the given constraint is minimized. Set $\widehat{\lambda_{SS}}$ to be the average value over the 10 groups. Finally, plug $\widehat{\lambda_{SS}}$ into (3.2) and use the full dataset to solve for the ultimate portfolio.

only due to the ℓ_1-norm constraint[7], but rather, more fundamentally, due to its methodology.

3.2 Parameter Setting

We simulate monthly returns from model (2.15) with parameters calibrated from real data. Specifically,

- For the parameters of factor returns, the mean $\boldsymbol{\mu}_f$ and covariance matrix $\boldsymbol{\Sigma}_f$ are taken to be the sample mean and sample covariance matrix of the Fama-French three factor (FF3) monthly returns from 2007 to 2016.

- As to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, out of the stocks that stayed in S&P 500 index during the period of 2007 – 2016, we randomly pick 100 of them. We then regress their monthly excess returns over FF3 returns, and set the resulting slopes to be the β_i's; the α_i's are obtained by hard thresholding the estimated intercepts with a threshold of two standard errors.

- Finally, the covariance matrix of idiosyncratic returns, $\boldsymbol{\Sigma}_u$, is obtained by applying the soft-thresholding method proposed in Rothman (2012)[8] to the sample covariance matrix of the residuals in the regression above.

Notice that the idiosyncratic returns generated from the parameters chosen as above may still possess factor structure.

3.3 Simulation Comparisons

3.3.1 When Returns Are Normally Distributed

We first show results for data generated under multivariate normal distribution. Returns of 100 stocks and 3 factors are generated using the parameters described in Section 3.2. The level of risk constraint is fixed to be $\sigma = 0.04$.

[7] A more straightforward comparison with other ℓ_1-norm constrained portfolios is made in Section 4.2, where we examine empirical performance based on DJIA 30 index constituents. There, it is seen clearly that adding ℓ_1-norm constraint to existing methods has similar effect to adding either no-short-sale or short-sale constraints. On the other hand, these ℓ_1-norm constrained portfolios incur prohibitively high computational costs, which is why we only include them in the empirical study based on DJIA index constituents with about 30 stocks.

[8] The soft-thresholding method can be implemented in R by the package "PDSCE". In our setting the penalty parameter "lam" is set to be 0.5.

We run 1,000 replications to evaluate the portfolio performance in terms of both risk and (annualized) Sharpe ratio. The comparison results for sample sizes $T = 120$ and 240 are summarized in Table 2.

Table 2

Simulation comparison of risks and Sharpe ratios of the portfolios under comparison based on 1,000 replications where returns of sample size $T = 120$ or 240 are generated from multivariate normal distribution. The risk constraint is set to be 0.04. The theoretical maximum Sharpe ratio is 1.882. Average value and standard deviation (in brackets) of each performance measure are reported.

Normal Distribution	$\sigma = 0.04$	$T = 120$
Portfolio	Risk	Sharpe Ratio
Factor	0.041 (0.003)	0.401 (0.169)
KZ	0.052 (0.040)	0.329 (0.184)
MAXSER	0.042 (0.005)	**1.194** (0.260)
MV/GMV with different covariance matrix estimates		
MV-P	0.296 (0.072)	0.367 (0.168)
MV-LS	0.082 (0.006)	0.697 (0.160)
MV-NLS	0.054 (0.017)	0.945 (0.183)
MV-NLSF	0.044 (0.002)	0.837 (0.139)
GMV-LS	0.013 (0.001)	0.438 (0.132)
GMV-NLS	0.015 (0.003)	0.553 (0.148)
MV with no-short-sale constraint		
MV-P-NSS	0.044 (0.003)	0.399 (0.040)
MV-LS-NSS	0.044 (0.003)	0.409 (0.036)
MV-NLS-NSS	0.043 (0.003)	0.416 (0.035)
MV with short-sale constraint & cross-validation		
MV-P-SSCV	0.044 (0.003)	0.399 (0.040)
MV-LS-SSCV	0.044 (0.003)	0.409 (0.036)
MV-NLS-SSCV	0.044 (0.004)	0.501 (0.169)

Normal Distribution	$\sigma = 0.04$	$T = 240$
Portfolio	Risk	Sharpe Ratio
Factor	0.041 (0.002)	0.467 (0.108)
KZ	0.091 (0.031)	0.909 (0.130)
MAXSER	0.041 (0.003)	**1.506** (0.140)
MV/GMV with different covariance matrix estimates		
MV-P	0.070 (0.005)	0.911 (0.123)
MV-LS	0.061 (0.004)	0.943 (0.117)
MV-NLS	0.049 (0.004)	1.199 (0.117)
MV-NLSF	0.042 (0.001)	1.068 (0.104)
GMV-LS	0.009 (0.000)	0.450 (0.102)
GMV-NLS	0.009 (0.001)	0.539 (0.167)
MV with no-short-sale constraint		
MV-P-NSS	0.042 (0.002)	0.415 (0.032)
MV-LS-NSS	0.042 (0.002)	0.420 (0.031)
MV-NLS-NSS	0.041 (0.002)	0.427 (0.030)
MV with short-sale constraint & cross-validation		
MV-P-SSCV	0.042 (0.002)	0.415 (0.032)
MV-LS-SSCV	0.042 (0.002)	0.420 (0.031)
MV-NLS-SSCV	0.042 (0.003)	0.468 (0.137)

From Table 2, we observe that

- In terms of *risk control*,
  - The risk of our MAXSER portfolio is close to the given constraint.
  - The Factor portfolio also has a risk close to the constraint. The reason is that as we pointed out earlier, for such a low-dimensional portfolio, simple plug-in estimator would be consistent.
  - MV portfolios with covariance matrix estimated by sample, linear shrinkage and nonlinear shrinkage estimators severely violate the risk constraint. When $T = 120$, their risks exceed the constraint by about 640%, 105% and 35%, respectively.

  - In sharp contrast, when either no-short-sale or short-sale constraint is imposed, the corresponding MV strategies lead to portfolios with risk close to the risk constraint level. This shows that our cross-validation procedure of choosing the short-sale threshold works for these strategies just as well as for our own MAXSER portfolio in terms of risk control.

- In terms of *Sharpe ratio*,

  - As the sample size T increases from 120 to 240, all portfolios improve.
  - MAXSER achieves the highest Sharpe ratio among all portfolios under comparison. In the $T = 240$ case, on average, MAXSER attains about 80% of the theoretical maximum Sharpe ratio, whereas the Sharpe ratio of the MV-NLS portfolio, the second highest among all portfolios, is about 64%.
  - In the $T = 240$ case, the 95% confidence interval of the mean Sharpe ratio of MAXSER is $[1.497, 1.515]$. For comparison, the 95% confidence interval for the difference between the mean Sharpe ratios of MAXSER and MV-NLS is $[0.301, 0.317]$. Such a range confirms that the improvement in Sharpe ratio achieved by MAXSER is not only statistically significant but also rather substantial.
  - Compared with MV-P-SSCV, MV-LS-SSCV and MV-NLS-SSCV portfolios, the Sharpe ratio of MAXSER is significantly higher, indicating that the outstanding performance of MAXSER is fundamentally due to our methodology, which is much more than solely imposing constraint.

- In summary, our MAXSER portfolio effectively controls risk, and is significantly more mean-variance efficient than the other portfolios.

3.3.2 When Returns Are Heavy-Tailed

Given the empirical evidence that financial returns tend to be heavy-tailed, in the following we conduct a simulation study for data with heavy-tails. More specifically, we shall let the factor and idiosyncratic returns be all *Student-t* distributed with 6 degrees of freedom. The mean and covariance matrix are taken to be the same as in Section 3.2.

Table 3

Simulation comparison of risks and Sharpe ratios of the portfolios under comparison based on 1,000 replications where returns of sample size $T = 120$ or 240 are generated from t-distribution with 6 degrees of freedom. The underlying mean and covariance matrix are the same as those set in Section 3.3.1. The risk constraint is set to be 0.04. The theoretical maximum Sharpe ratio is 1.882. Average value and standard deviation (in brackets) of each performance measure are reported.

$t(6)$ Distribution	$\sigma = 0.04$	$T = 120$
Portfolio	Risk	Sharpe Ratio
Factor	0.034 (0.003)	0.350 (0.202)
KZ	0.039 (0.031)	0.288 (0.191)
MAXSER	0.034 (0.004)	**1.037** (0.274)
MV/GMV with different covariance matrix estimates		
MV-P	0.246 (0.060)	0.321 (0.174)
MV-LS	0.062 (0.005)	0.635 (0.169)
MV-NLS	0.042 (0.009)	0.845 (0.179)
MV-NLSF	0.036 (0.002)	0.716 (0.150)
GMV-LS	0.013 (0.001)	0.459 (0.130)
GMV-NLS	0.014 (0.003)	0.572 (0.125)
MV with no-short-sale constraint		
MV-P-NSS	0.036 (0.003)	0.394 (0.039)
MV-LS-NSS	0.036 (0.003)	0.406 (0.035)
MV-NLS-NSS	0.035 (0.003)	0.411 (0.036)
MV with short-sale constraint & cross-validation		
MV-P-SSCV	0.036 (0.003)	0.394 (0.039)
MV-LS-SSCV	0.036 (0.003)	0.406 (0.035)
MV-NLS-SSCV	0.037 (0.004)	0.531 (0.196)

$t(6)$ Distribution	$\sigma = 0.04$	$T = 240$
Portfolio	Risk	Sharpe Ratio
Factor	0.033 (0.002)	0.427 (0.141)
KZ	0.059 (0.023)	0.802 (0.154)
MAXSER	0.033 (0.003)	**1.377** (0.185)
MV/GMV with different covariance matrix estimates		
MV-P	0.058 (0.004)	0.807 (0.140)
MV-LS	0.048 (0.003)	0.847 (0.133)
MV-NLS	0.040 (0.004)	1.071 (0.138)
MV-NLSF	0.034 (0.001)	0.931 (0.117)
GMV-LS	0.010 (0.000)	0.469 (0.107)
GMV-NLS	0.010 (0.001)	0.538 (0.182)
MV with no-short-sale constraint		
MV-P-NSS	0.034 (0.002)	0.406 (0.035)
MV-LS-NSS	0.034 (0.002)	0.412 (0.033)
MV-NLS-NSS	0.034 (0.002)	0.418 (0.032)
MV with short-sale constraint & cross-validation		
MV-P-SSCV	0.034 (0.002)	0.406 (0.035)
MV-LS-SSCV	0.034 (0.002)	0.414 (0.046)
MV-NLS-SSCV	0.035 (0.003)	0.529 (0.221)

Table 3 shows that MAXSER continues to clearly outperform other portfolios. Another observation is that, if we compare Table 3 with Table 2 for the normal case, we see that heavy-tailedness does to some extent hurt all the strategies in terms of Sharpe ratio.

4 EMPIRICAL STUDIES

We investigate the performance of our strategy through two types of empirical studies:

- *Practical* evaluation:

  The asset pool containing the constituents of the DJIA 30 index is considered. Under a rolling-window scheme to be specified below, at each rebalancing time point, only the constituents at that time are included. Portfolio performances are compared on the basis of both raw returns and returns net of transaction costs; see Section 4.2 for details.

- *General statistical* evaluation:

  We compare the performance of the strategies under consideration using 100 random datasets, in which the stocks are randomly picked historical constituents of the S&P 500 index. See Section 4.3 for details.

4.1 Portfolios Under Comparison in Empirical Study

In addition to the strategies that we compared in simulation studies, we include five more portfolios in our empirical study: the index, the equally weighted portfolio ("1/N" rule), and three ℓ_1-norm constrained mean-variance portfolios. Specifically, based on covariance matrices estimated by the sample covariance matrix, the linear shrinkage estimator (Ledoit and Wolf (2004)) and the nonlinear shrinkage estimator (Ledoit and Wolf (2017)), we construct the portfolios "MV-P-L1CV", "MV-LS-L1CV" and "MV-NLS-L1CV" by imposing ℓ_1-norm constraint[9] for which the constraint is determined by the cross-validation method that we proposed in Section 2.5.1. We include these ℓ_1-norm constrained portfolios to examine the effect of imposing ℓ_1-norm constraint, and, more importantly, to demonstrate that the advantage of MAXSER is fundamentally due to its methodology rather than solely due to the ℓ_1-norm constraint.

4.2 Practical Evaluation

4.2.1 Data & Investment Rolling–Window Scheme

We first evaluate our proposed portfolio, MAXSER, based on the stock universe of DJIA 30 index constituents. We obtain the lists of DJIA 30 index constituents from COMPUSTAT and CRSP. In addition to the asset pool that consists only of stocks, we also consider the larger pool in which Fama-French three factors (FF3) are included. Correspondingly, we perform

[9] Due to the prohibitively high time cost of solving the original mean-variance optimization with ℓ_1-norm constraint, the three ℓ_1-norm constrained methods are only applied to the DJIA data with around 30 stocks. In contrast, the computational cost of MAXSER is very low thanks to the fast algorithm LARS (Efron et al. (2004)).

two studies: Study I for the stock only case, and Study II for the case that includes both stocks and Fama-French three factors. The evaluation is conducted based on a rolling-window scheme. More specifically, at the beginning of each month, one asset pool is formed by including the current constituents of DJIA 30 index (and the Fama-French three factors for Study II). Different portfolios are constructed using the monthly excess returns[10] during the past T months, where T is the sample size to be specified. If a stock has missing data in the T-month training period, it is excluded from the asset pool. As a result, the number of stocks would vary over time and can be slightly smaller than the total number of constituents. The risk constraint is fixed to be the standard deviation of DJIA 30 index returns during the first training period. The portfolios are held for one month, and the corresponding returns are recorded. We then evaluate the performance of the portfolios under comparison based on their out-of-sample monthly returns.

4.2.2 Performance Evaluation

We evaluate the performance of MAXSER and other competing portfolios in terms of risk and annualized Sharpe ratio. We also investigate the effect of transaction costs and report the comparisons based on returns net of transaction costs.

For the DJIA data set with around 30 stocks in each asset pool, we use a sample size of $T = 60$, in other words, each training set contains the monthly returns in the past five years. The testing period is February 1967 – December 2016, which results in, for each strategy, 599 out-of-sample monthly returns.

In addition to comparing out-of-sample risks and Sharpe ratios, to verify the statistical significance of the advantage of our portfolio MAXSER, we conduct hypothesis tests about the Sharpe ratio. More specifically, we test

$$H_0 : SR_{MAXSER} \leqslant SR_0 \quad vs \quad H_a : SR_{MAXSER} > SR_0, \tag{4.1}$$

where SR_{MAXSER} denotes the Sharpe ratio of MAXSER portfolio, and SR_0 denotes the Sharpe ratio of the portfolio under comparison. The test is conducted using Memmel (2003)'s corrected version of Jobson and Korkie (1981)'s test.

[10] For computing excess returns, we obtain the risk-free rate r_f from Fama/French Data Library.

4.2.3 Study I: DJIA Constituents Only

We first consider the asset pool that consists only of DJIA constituents. The comparisons among our MAXSER and the competing portfolios based on returns either without or with transaction costs are summarized below.

Without transaction costs

The summary without considering transaction costs is reported in Table 4, which shows the risk, Sharpe ratio, and the p-value for testing (4.1) for each portfolio.

Table 4

Summary of risk, Sharpe ratio and p-value of Sharpe ratio test (4.1) *of the portfolios under comparison based on DJIA 30 index constituents. The testing period is February 1967 – December 2016. The risk constraint* $\sigma = 0.037$ *is the standard deviation of the monthly excess returns on DJIA 30 index during February 1962 – January 1967.*

DJIA 30 Constituents	σ=0.037	$T = 60$	
Portfolio	Risk	Sharpe Ratio	p-value
Index	0.044	0.102	0.000
Equally weighted	0.046	0.026	0.000
KZ	0.063	0.069	0.000
MAXSER	0.042	**0.250**	–
MV/GMV with different covariance matrix estimates			
MV-P	0.078	0.020	0.000
MV-LS	0.057	0.049	0.000
MV-NLS	0.054	0.061	0.000
MV-NLSF	0.054	0.088	0.000
GMV-LS	0.039	0.063	0.000
GMV-NLS	0.038	0.097	0.003
MV with no-short-sale constraint			
MV-P-NSS	0.040	0.127	0.004
MV-LS-NSS	0.040	0.135	0.006
MV-NLS-NSS	0.040	0.136	0.007
MV with short-sale constraint & cross-validation			
MV-P-SSCV	0.042	0.017	0.000
MV-LS-SSCV	0.049	0.107	0.000
MV-NLS-SSCV	0.047	0.183	0.027
MV with ℓ_1-norm constraint & cross-validation			
MV-P-L1CV	0.041	0.103	0.000
MV-LS-L1CV	0.042	0.153	0.005
MV-NLS-L1CV	0.043	0.137	0.003

We observe from Table 4 that

- In terms of *risk control*,
  - The index, MAXSER, GMV, short-sale and ℓ_1-norm constrained portfolios carry similar risks, all of which are slightly higher than the given constraint.
  - The MV portfolios with different covariance matrix estimates, MV-P, MV-LS, MV-NLS and MV-NLSF, carry higher risks. Most notably, the plug-in portfolio MV-P violates the risk constraint by about 110%.
- In terms of *Sharpe ratio*,
  - MAXSER dominates the other competing portfolios with a Sharpe ratio about 37% higher than that of MV-NLS-SSCV portfolio, the second best portfolio in terms of Sharpe ratio.
  - The three portfolios, MV-P-L1CV, MV-LS-L1CV and MV-NLS-L1CV, are also constructed with ℓ_1-norm constraint determined by our cross-validation procedure, however, they yield significantly lower Sharpe ratios than MAXSER. Such a comparison clearly shows that the outstanding performance of MAXSER is fundamentally due to its methodology rather than solely imposing ℓ_1-norm constraint.
  - The p-values of Sharpe ratio tests show that the advantage of MAXSER is statistically significant.

In summary, our MAXSER effectively controls the risk and significantly outperforms the competing portfolios in terms of Sharpe ratio.

With transaction costs

Next, we take transaction costs into account and compute the returns net of transaction costs. Here we adopt a widely-used way of computing transaction costs (see, for example, Engle et al. (2012)), which is closely related to the portfolio turnover. The turnover at any rebalancing time t is defined as

$$\text{Turnover}(t) := \sum_{j=1}^{N} |w_j(t+1) - w_j(t+)|, \tag{4.2}$$

where $w_j(t+1)$ is the weight on asset j at the beginning of period $t+1$, and $w_j(t+)$ is the weight of the same asset at the end of period t. The transaction cost of the portfolio at time t

is proportional to Turnover(t) and a cost level c_0, which measures transaction cost per dollar traded. It can be derived that the portfolio return net of transaction cost in period t, $r_{net}(t)$, has the following relation with the portfolio return without transaction cost $r(t)$:

$$r_{net}(t) = (1 - c_0\text{Turnover}(t))\,(1 + r(t)) - 1. \tag{4.3}$$

In Engle et al. (2012), it is found that the average cost level for NYSE stocks is around 0.088%. In the following analysis we adopt $c_0 = 0.1\%$. For the index, we set its transaction cost to be zero. Table 5 shows the risks and Sharpe ratios of different portfolios after deducting transaction costs.

Table 5

Summary of risk, Sharpe ratio and p-value of Sharpe ratio test (4.1) *based on returns net of transaction costs of different portfolios based on DJIA 30 constituents. The testing period is February 1967 – December 2016.*

DJIA 30 Constituents	$c_0 = 0.1\%$	σ=0.037	
Portfolio	Risk	Sharpe Ratio	p-value
Index	0.044	0.102	0.000
Equally weighted	0.046	0.020	0.001
KZ	0.063	0.011	0.000
MAXSER	0.042	**0.209**	–
MV/GMV with different covariance matrix estimates			
MV-P	0.078	−0.043	0.000
MV-LS	0.057	0.001	0.000
MV-NLS	0.054	0.017	0.000
MV-NLSF	0.054	0.044	0.000
GMV-LS	0.039	0.038	0.001
GMV-NLS	0.038	0.074	0.008
MV with no-short-sale constraint			
MV-P-NSS	0.040	0.108	0.015
MV-LS-NSS	0.040	0.118	0.024
MV-NLS-NSS	0.040	0.119	0.026
MV with short-sale constraint & cross-validation			
MV-P-SSCV	0.042	−0.048	0.000
MV-LS-SSCV	0.049	0.023	0.000
MV-NLS-SSCV	0.047	0.091	0.000
MV with ℓ_1-norm constraint & cross-validation			
MV-P-L1CV	0.041	0.029	0.000
MV-LS-L1CV	0.042	0.073	0.000
MV-NLS-L1CV	0.043	0.047	0.000

Table 5 shows that when transaction cost is considered, MAXSER again performs significantly better than the competing portfolios.

4.2.4 Study II: DJIA Constituents & Fama-French Three Factors

In this section, we include the Fama-French three factors in the asset pool, and compare MAXSER with the competing portfolios which are also applied to both stocks and the three factors. Again we summarize the performances without and with transaction costs.

Without transaction costs

The summary without considering transaction costs is reported in Table 6, which shows the risk, Sharpe ratio, and the p-value for testing (4.1) for each portfolio.

Table 6

Summary of risk, Sharpe ratio and p-value of Sharpe ratio test (4.1) *of the portfolios under comparison based on DJIA 30 index constituents and Fama-French three factors. The testing period is February 1967 – December 2016. The risk constraint* $\sigma = 0.037$ *is the standard deviation of the monthly excess returns on DJIA 30 index during February 1962 – January 1967.*

DJIA 30 Constituents & FF3	σ=0.037	$T = 60$	
Portfolio	Risk	Sharpe Ratio	p-value
Index	0.044	0.102	0.000
Equally weighted	0.043	0.064	0.000
Factor	0.040	0.425	0.000
KZ	0.124	0.535	0.000
MAXSER	0.045	**0.706**	–
MV/GMV with different covariance matrix estimates			
MV-P	0.089	0.595	0.008
MV-LS	0.053	0.303	0.000
MV-NLS	0.053	0.457	0.000
MV-NLSF	0.051	0.524	0.000
GMV-LS	0.017	0.397	0.000
GMV-NLS	0.017	0.337	0.000
MV with no-short-sale constraint			
MV-P-NSS	0.040	0.463	0.000
MV-LS-NSS	0.035	0.431	0.000
MV-NLS-NSS	0.035	0.399	0.000
MV with short-sale constraint & cross-validation			
MV-P-SSCV	0.042	0.543	0.000
MV-LS-SSCV	0.047	0.371	0.000
MV-NLS-SSCV	0.041	0.294	0.000
MV with ℓ_1-norm constraint & cross-validation			
MV-P-L1CV	0.038	0.357	0.000
MV-LS-L1CV	0.045	0.440	0.000
MV-NLS-L1CV	0.046	0.386	0.000

From Table 6, we observe the following:

- In terms of *risk control*,
    - The risk of MAXSER is close to those of the index, the equally weighted, the Factor, the short-sale constrained and ℓ_1-norm constrained portfolios.
    - The risk of the plug-in ("MV-P") portfolio is more than twice of the risk constraint level and hardly bearable for investors.
- In terms of *Sharpe ratio*,
    - Our MAXSER portfolio yields the highest Sharpe ratio.
    - Compared with portfolios with similar risks, the Sharpe ratio of MAXSER is about 30% higher than that of MV-P-SSCV portfolio, which performs better than other portfolios under comparison with similar risks.
    - The ℓ_1-norm constrained portfolios, MV-P-L1CV, MV-LS-L1CV and MV-NLS-L1CV, yield much lower Sharpe ratios than MAXSER. This again confirms that the advantage of MAXSER is fundamental rather than solely due to imposing the constraint.
    - The small p-values of Sharpe ratio tests of MAXSER against all the other portfolios under comparison confirm the statistical significance of the advantage of MAXSER.
    - A new observation from this study is that, the comparison between the Sharpe ratios of MAXSER and Factor portfolios indicates that investing in individual stocks in addition to factors using our strategy MAXSER can yield substantial gain.

In summary, MAXSER effectively controls out-of-sample risk and dominates the competing portfolios in terms of mean-variance efficiency.

With transaction costs

Next, we again take transaction costs into account and compute the returns of each portfolio net of transaction costs as described in (4.3). Table 7 shows the risks and Sharpe ratios after deducting transaction costs.

Table 7

Summary of risk, Sharpe ratio and p-value of Sharpe ratio test (4.1) *based on returns net of transaction costs of different portfolios based on DJIA 30 constituents and Fama-French three factors. The testing period is February 1967 – December 2016.*

DJIA 30 Constituents & FF3	$c_0 = 0.1\%$	σ=0.037	
Portfolio	Risk	Sharpe Ratio	p-value
Index	0.044	0.102	0.000
Equally weighted	0.043	0.059	0.000
Factor	0.040	0.403	0.000
KZ	0.123	0.440	0.000
MAXSER	0.045	**0.634**	–
MV/GMV with different covariance matrix estimates			
MV-P	0.089	0.516	0.004
MV-LS	0.053	0.251	0.000
MV-NLS	0.053	0.379	0.000
MV-NLSF	0.051	0.466	0.000
GMV-LS	0.017	0.368	0.000
GMV-NLS	0.017	0.266	0.000
MV with no-short-sale constraint			
MV-P-NSS	0.040	0.437	0.000
MV-LS-NSS	0.035	0.409	0.000
MV-NLS-NSS	0.035	0.373	0.000
MV with short-sale constraint & cross-validation			
MV-P-SSCV	0.042	0.449	0.000
MV-LS-SSCV	0.047	0.282	0.000
MV-NLS-SSCV	0.041	0.150	0.000
MV with ℓ_1-norm constraint & cross-validation			
MV-P-L1CV	0.038	0.269	0.000
MV-LS-L1CV	0.045	0.339	0.000
MV-NLS-L1CV	0.045	0.273	0.000

Table 7 again shows the clear advantage of our MAXSER portfolio. MAXSER still yields a Sharpe ratio that is significantly higher than the other strategies.

4.2.5 Discussions on Study I & II

In the previous two Sections 4.2.3 and 4.2.4, we considered two scenarios, one investing only in stocks (the constituents of DJIA 30 index, to be precise), the other investing in both stocks and the Fama-French three factors. Comparing the two scenarios shows clearly that for all strategies, it is beneficial to invest in both stocks and factors such as the Fama-French three factors. It is for this reason that in the simulation studies (Section 3) we focused on the setting with factor investing. In the second type of empirical evaluation below, we will also focus on the scenario with investments in both stocks and Fama-French three factors.

Another important observation is that, if we compare MAXSER with Factor portfolio in either Table 6 or 7, we see that MAXSER substantially enhances the performance. In other words, the value added by applying MAXSER to invest in both stocks and factors is far beyond marginal. The conclusion is also supported by the general statistical evaluation results in Section 4.3.

4.3 General Statistical Evaluation

The comparisons in Section 4.2 are from a practical viewpoint, where for each period the stock pool is updated to include all index constituents. In this section, we evaluate the portfolio performances from a more statistical point of view, based on 100 random stock pools formed by historical constituents of S&P 500 index. Specifically, each stock pool consists of 100 stocks randomly chosen from the stock universe, which contains 369 S&P 500 index historical constituents that have complete price data during January 1992 – December 2016, the whole study period. Each time when we randomly select 100 stocks, they are taken to form the stock pool throughout the study period. In such a way, we eliminate the effect of inclusion/exclusion of stocks. We make overall evaluations based on 100 randomizations.

4.3.1 Comparison Summary

The following results are based on the same rolling-window scheme as described in Section 4.2.1. The sample size is $T = 120$, and we again include the Fama-French three factors in our asset pools. The means and standard deviations of portfolio risks and Sharpe ratios based on the 100 randomizations are reported in Table 8.

Table 8

Summary of risks and Sharpe ratios of the portfolios under comparison for 100 random asset pools, each containing 100 stocks randomly selected from S&P 500 index historical constituents and Fama-French three factors. The testing period is January 2002 – December 2016. The risk constraint is taken to be the standard deviation of the index excess returns during January 1992 – December 2001. Average value and standard deviation (in brackets) of each performance measure are reported.

S&P 500 Constituents & FF3	$\sigma = 0.041$	$T = 120$
Portfolio	Risk	Sharpe Ratio
Index	0.042	0.223
Factor	0.041	0.320
Equally weighted	0.050 (0.002)	0.261 (0.041)
KZ	0.072 (0.019)	0.311 (0.234)
MAXSER	0.044 (0.003)	**0.527** (0.184)
MV/GMV with different covariance matrix estimates		
MV-P	0.334 (0.031)	0.325 (0.230)
MV-LS	0.065 (0.004)	0.194 (0.180)
MV-NLS	0.061 (0.004)	0.188 (0.179)
MV-NLSF	0.057 (0.003)	0.353 (0.161)
GMV-LS	0.025 (0.001)	0.420 (0.127)
GMV-NLS	0.025 (0.001)	0.414 (0.123)
MV with no-short-sale constraint		
MV-P-NSS	0.047 (0.002)	0.345 (0.120)
MV-LS-NSS	0.043 (0.002)	0.288 (0.138)
MV-NLS-NSS	0.042 (0.002)	0.285 (0.146)
MV with short-sale constraint & cross-validation		
MV-P-SSCV	0.047 (0.002)	0.345 (0.120)
MV-LS-SSCV	0.043 (0.002)	0.286 (0.139)
MV-NLS-SSCV	0.043 (0.002)	0.286 (0.153)

Table 8 shows that

- MAXSER portfolio carries a risk close to the risk constraint.
- MAXSER achieves the highest average Sharpe ratio, which is 27% higher than the

average Sharpe ratio of GMV-LS portfolio, the second best among the portfolios under comparison.

Next, we take transaction costs into account. The returns net of transaction costs are computed by formula (4.3). The transaction cost level is again taken to be 0.1%. The comparisons are summarized in Table 9.

Table 9

Comparison of risks and Sharpe ratios based on portfolio returns net of transaction costs, for 100 random asset pools formed by 100 S&P 500 constituents and Fama-French three factors. Average value and standard deviation (in brackets) of each performance measure are reported. The testing period is January 2002 – December 2016.

S&P 500 Constituents & FF3	$c_0 = 0.1\%$	$\sigma = 0.0247$
Portfolio	Risk	Sharpe Ratio
Index	0.042	0.223
Factor	0.041	0.303
Equally weighted	0.050 (0.002)	0.256 (0.041)
KZ	0.072 (0.018)	0.140 (0.234)
MAXSER	0.044 (0.003)	**0.463** (0.183)
MV/GMV with different covariance matrix estimates		
MV-P	0.328 (0.029)	0.101 (0.231)
MV-LS	0.065 (0.004)	0.139 (0.182)
MV-NLS	0.061 (0.004)	0.123 (0.181)
MV-NLSF	0.057 (0.003)	0.300 (0.159)
GMV-LS	0.025 (0.001)	0.388 (0.126)
GMV-NLS	0.025 (0.001)	0.365 (0.122)
MV with no-short-sale constraint		
MV-P-NSS	0.047 (0.002)	0.326 (0.121)
MV-LS-NSS	0.043 (0.002)	0.272 (0.139)
MV-NLS-NSS	0.042 (0.002)	0.269 (0.147)
MV with short-sale constraint & cross-validation		
MV-P-SSCV	0.047 (0.002)	0.326 (0.121)
MV-LS-SSCV	0.043 (0.002)	0.269 (0.140)
MV-NLS-SSCV	0.043 (0.002)	0.265 (0.154)

Table 9 reveals that MAXSER portfolio maintains its advantage over other portfolios. Its Sharpe ratio is more than 20% higher than that of GMV-LS, the second best portfolio in terms of Sharpe ratio among the portfolios under comparison.

In summary, both without and with transaction costs, our MAXSER portfolio outperforms.

Moreover, as we will see in the next section, the advantage of MAXSER over other portfolios in terms of Sharpe ratio is dominating.

4.3.2 Statistical Tests About Sharpe Ratio

In this section, we conduct the Sharpe ratio test (4.1) based on both raw returns and returns net of transaction costs for the 100 random asset pools. The histograms of the p-values based on the 100 random asset pools are given in Figures 5 $\sim$ 6.

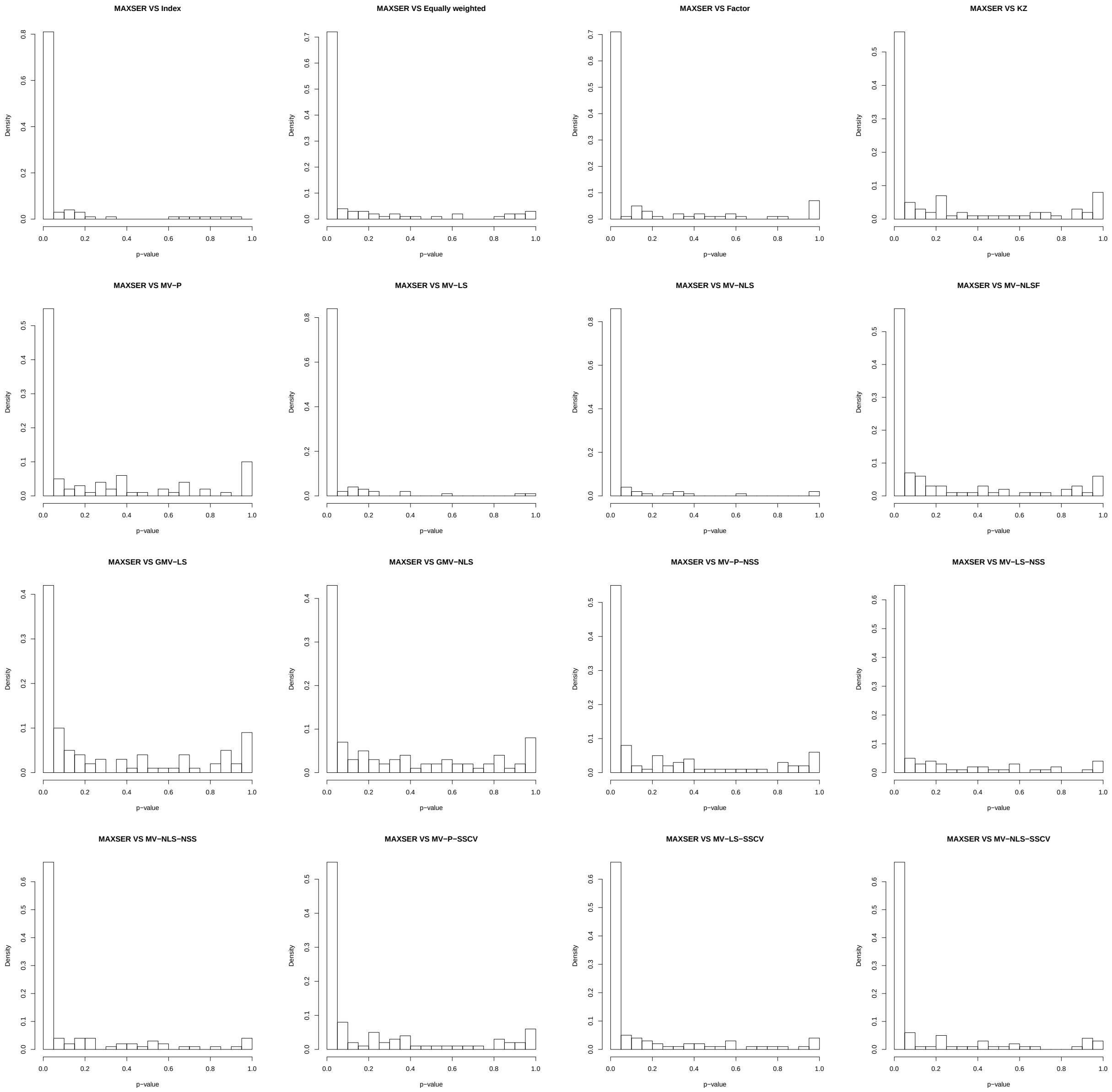

Figure 5. *Histograms of p-values for the Sharpe ratio test* (4.1) *of MAXSER against the portfolios under comparison, based on 100 random asset pools. In this figure transaction costs are not taken into account. The consistent feature of concentration of p-values around zero shows that MAXSER has dominating advantage over the competing portfolios in terms of Sharpe ratio.*

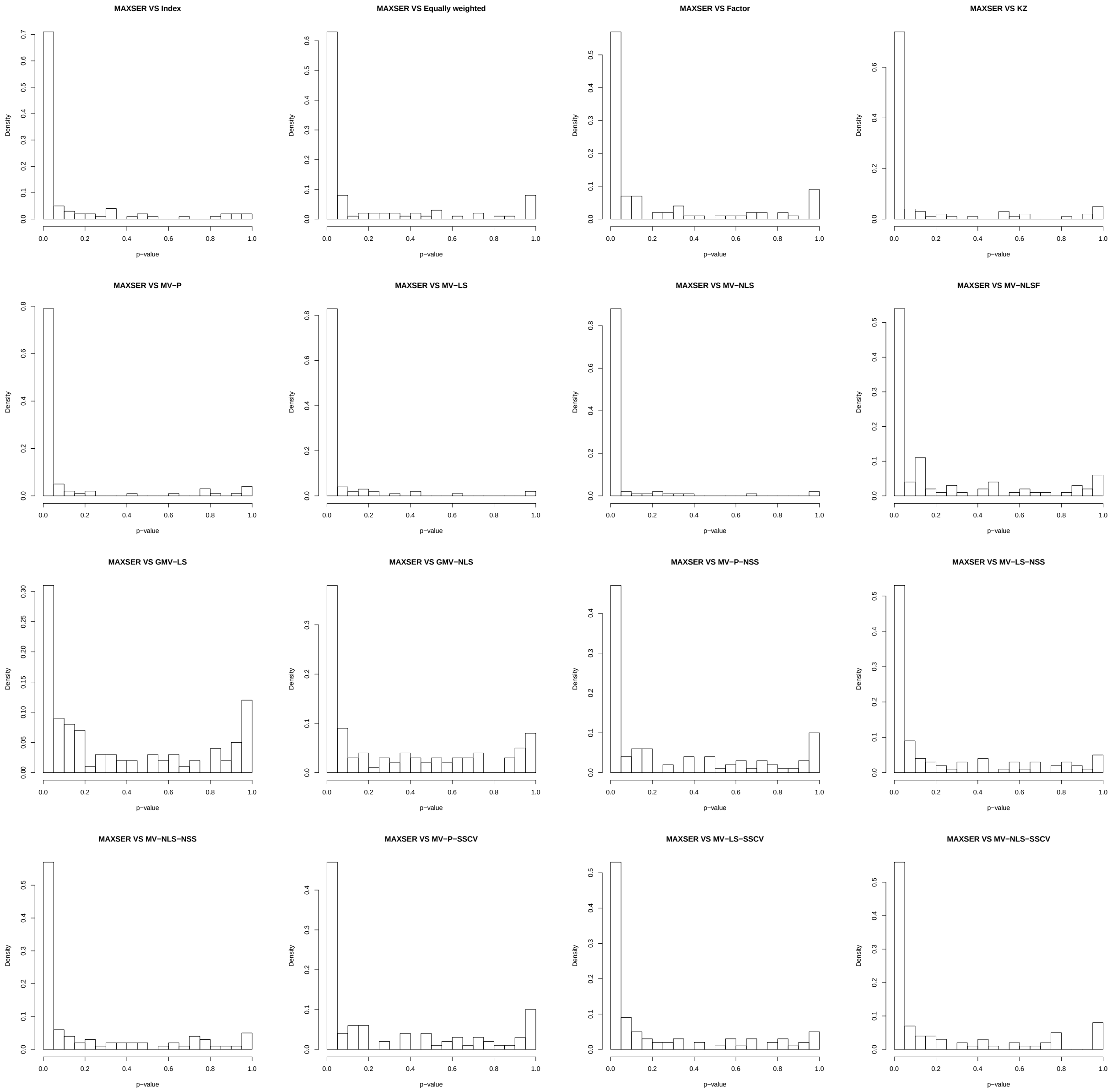

Figure 6. *Histograms of p-values for the Sharpe ratio test* (4.1) *of MAXSER against the portfolios under comparison, based on 100 random asset pools. In this figure transaction costs are deducted from portfolio returns. Similarly to Figure 5, MAXSER is seen to have dominating advantage over the competing portfolios in terms of Sharpe ratio.*

Notice that if the null hypothesis H_0 in (4.1) holds, then the p-values would be roughly uniformly distributed. This is obviously not the case here. We observe from Figures 5 and 6 that, for all competing portfolios and both without and with transaction costs taken into account, there is a consistent feature of p-values concentrating around zero. Such a feature shows that MAXSER has dominating advantage over the competing portfolios in terms of Sharpe ratio.

5 CONCLUSION

In this paper, we propose a novel approach to estimate the mean-variance efficient portfolio when the number of assets in the investment pool is not small compared with sample size. We consider two scenarios, one without factor investing, the other with factor investing. We prove that, under both scenarios, our strategy, MAXSER, asymptotically achieves the mean-variance efficiency and meanwhile effectively controls the risk. To the best of our knowledge, this is the first method that can simultaneously achieve these two goals for large portfolios.

The sound theoretical properties of MAXSER are supported by comprehensive numerical studies. In both simulation and empirical analyses, MAXSER can not only effectively control the risk, but more importantly, yield high Sharpe ratios.

Appendix A Proof of Convergence (1.1)

Proof. Let $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ be the sample mean and sample covariance matrix of i.i.d. returns following multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For given risk constraint σ, the plug-in portfolio is

$$\widehat{\boldsymbol{w}}_p = \frac{\sigma}{\sqrt{\widehat{\boldsymbol{\mu}}'\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}}}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}. \tag{A.1}$$

It follows that the ratio between the Sharpe ratio of $\widehat{\boldsymbol{w}}_p$ and $SR^* = \sqrt{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}$ equals

$$\frac{SR(\text{plug-in})}{SR^*} = \frac{\frac{\boldsymbol{\mu}'\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}}{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}}{\sqrt{\frac{\widehat{\boldsymbol{\mu}}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}}{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}}}. \tag{A.2}$$

By Theorem 4.6 in El Karoui (2010), for the numerator in (A.2) we have

$$\frac{\boldsymbol{\mu}'\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}}{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} \xrightarrow{P} \frac{1}{1-\rho}. \tag{A.3}$$

Furthermore, by Theorems 3.1 and 3.3 in El Karoui (2013), for the denominator in (A.2) we have

$$\frac{\widehat{\boldsymbol{\mu}}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}}{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} = \frac{1}{(1-\rho)^3}\left(1 + \frac{\rho}{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}\right) + o_p(1). \tag{A.4}$$

Combining (A.3) and (A.4) we obtain (1.1). □

Appendix B Proof of Proposition 1

Proof. Observe that

$$E(r_c - \boldsymbol{w}'\boldsymbol{r})^2 = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} + (\boldsymbol{w}'\boldsymbol{\mu})^2 - 2r_c\boldsymbol{w}'\boldsymbol{\mu}. \tag{A.5}$$

It follows that the minimizer of $E(r_c - \boldsymbol{w}'\boldsymbol{r})^2$ satisfies the first order condition below:

$$\boldsymbol{\Sigma}\boldsymbol{w} + (\boldsymbol{w}'\boldsymbol{\mu})\boldsymbol{\mu} - r_c\boldsymbol{\mu} = 0. \tag{A.6}$$

Left multiplying both sides by $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}$ yields

$$\boldsymbol{w}'\boldsymbol{\mu} + (\boldsymbol{w}'\boldsymbol{\mu}) \cdot \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - r_c \cdot \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = 0.$$

Recall that $\theta = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. We therefore have

$$\boldsymbol{w}'\boldsymbol{\mu} = \frac{\theta}{1+\theta}r_c = r^*. \tag{A.7}$$

Combining (A.6) and (A.7), we obtain

$$\boldsymbol{w} = \frac{r^*}{\theta}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \boldsymbol{w}^*.$$

□

Appendix C Proof of Proposition 2

Proof. Kan and Zhou (2007) noticed that the sample estimate $\widehat{\theta}_s$ is in fact Hotelling's T^2 , which follows non-centralized F-distribution:

$$\widehat{\theta}_s \sim \left(\frac{N}{T-N}\right) F_{N,T-N}(T\theta). \tag{A.8}$$

It immediately follows that $\widehat{\theta}$ is an unbiased estimator of θ. Moreover, the variance of $\widehat{\theta}_s$ is

$$\text{Var}\left(\widehat{\theta}_s\right) = \frac{2T^2\theta^2 + 2(T-2)(N+2T\theta)}{(T-N-2)^2(T-N-4)}. \tag{A.9}$$

Under the assumption that $N/T \to \rho \in (0,1)$ and that θ is bounded, $\text{Var}(\widehat{\theta})$ converges to 0 at rate $1/T$. The conclusion follows. □

Appendix D Proof of Theorem 1

In order to prove Theorem 1, we first give the following Proposition 7 about an infeasible estimator of $\boldsymbol{w}^*$ defined as

$$\widetilde{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} ||\boldsymbol{r_c} - \boldsymbol{R}\boldsymbol{w}||_2^2 \quad \text{subject to} \quad ||\boldsymbol{w}||_1 \leq \lambda, \tag{A.10}$$

where $\boldsymbol{r_c} = (r_c, \ldots, r_c)'$ with $r_c = (1+\theta)/\theta \cdot r^*$.

Proposition 7. *Under Assumptions A1 $\sim$ A5, as $N \to \infty$,*

$$E|\boldsymbol{\mu}'\boldsymbol{w}^* - \boldsymbol{\mu}'\widetilde{\boldsymbol{w}}| \to 0, \tag{A.11}$$

and

$$E\left|\sqrt{\widetilde{\boldsymbol{w}}'\boldsymbol{\Sigma}\widetilde{\boldsymbol{w}}} - \sqrt{\boldsymbol{w}^{*\prime}\boldsymbol{\Sigma}\boldsymbol{w}^*}\right| \to 0. \tag{A.12}$$

We first prove Proposition 7 in Section D.1. The proof of Theorem 1 is given in Section D.2.

D.1 Proof of Proposition 7

1. A Preliminary Theorem

We first establish a theorem that lays down the basis for Proposition 7. The following notation will be used.

Notation. The notation $\overset{\mathcal{D}}{=}$ means equal in distribution. The notation $Y_T = o_p(f(T))$ means that $Y_T/f(T) \xrightarrow{P} 0$, and $Y_T = O_p(f(T))$ means that the sequence $(|Y_T/f(T)|)$ is tight, i.e., for any $\varepsilon > 0$, there exists a finite $M > 0$ such that $P(|Y_T/f(T)| > M) < \varepsilon$ for all T.

Theorem 3. *Define the $\boldsymbol{\Sigma}$-norm estimation error of $\widetilde{\boldsymbol{w}}$ as $E||\boldsymbol{w}^* - \widetilde{\boldsymbol{w}}||^2_{\boldsymbol{\Sigma}}$, where $||\cdot||_{\boldsymbol{\Sigma}}$ is the norm induced by $\boldsymbol{\Sigma}$: $||\boldsymbol{x}||^2_{\boldsymbol{\Sigma}} := \boldsymbol{x}'\boldsymbol{\Sigma}\boldsymbol{x}$ for any $\boldsymbol{x} \in \mathbb{R}^N$. Further let $\boldsymbol{R}^* = \boldsymbol{R}\boldsymbol{w}^*$ and $\widetilde{\boldsymbol{R}} = \boldsymbol{R}\widetilde{\boldsymbol{w}}$, where $\widetilde{\boldsymbol{w}}$ is the infeasible estimator defined in* (A.10)*. Under Assumptions A1 and A3, we have*

$$E\left(\frac{1}{T}||\boldsymbol{R}^* - \widetilde{\boldsymbol{R}}||^2_2\right) \leq \frac{G_1}{\sqrt{T}} + \lambda\sigma\left(2\sqrt{2L} + 2M + \sqrt{\frac{L}{\theta}}\right)\sqrt{\frac{2\log(2N)}{T}}, \tag{A.13}$$

and

$$E||\boldsymbol{w}^* - \widetilde{\boldsymbol{w}}||^2_{\boldsymbol{\Sigma}} \leq G_1\frac{1}{\sqrt{T}} + G_2\sqrt{\frac{2\log(2N)}{T}} + G_3\sqrt{\frac{2\log(2N^2)}{T}}, \tag{A.14}$$

where $G_1 = 2\lambda\sigma\sqrt{2L}+\sigma^2/\sqrt{\theta}$, $G_2 = \lambda\left(\sigma\left(2\sqrt{2L} + 2M + \sqrt{\frac{L}{\theta}}\right) + 8\lambda L\right)$, and $G_3 = 4\lambda^2(\sqrt{2}L + 2M\sqrt{L})$.

Remark 10. *Theorem 3 is seemingly similar to but actually fundamentally different from Theorem 1 in Chatterjee (2013). The difference is due to that in our setting, the "noise" $\boldsymbol{\varepsilon} = \boldsymbol{r_c} - \boldsymbol{R}\boldsymbol{w}$ is not independent of the "covariate" $\boldsymbol{R}$.*

Before we prove Theorem 3, we give a few lemmas that will be used.

Lemma 1 (Lemma 3 in Chatterjee (2013))**.** *Suppose that $\zeta_i \sim N(0, \sigma_i^2)$ for $i = 1, \dots, m$, which need not be independent. Then*

$$E\left(\max_{1\leq i\leq m}|\zeta_i|\right) \leq \max_{1\leq i\leq m}\sigma_i \cdot \sqrt{2\log(2m)}.$$

Lemma 2. *Suppose that $\xi_i \sim \chi^2(T)$ for $i = 1, \dots, m$, which need not be independent. If $\sqrt{\log(2m)/2T} \leq 1/4$, then*

$$E\left(\max_{1\leq i\leq m}|\xi_i - T|\right) \leq 2\sqrt{2T\log(2m)}.$$

Proof. Using the moment generating function of $\chi^2(T)$, we have for $\eta < \frac{1}{2}$,

$$E\left(e^{\eta(\xi_i - T)}\right) = e^{-\frac{T}{2}\log(1-2\eta) - \eta T}.$$

Elementary calculus shows that

$$\log(1+x) \geq x - x^2, \text{ for all } |x| \leq \frac{1}{2}. \tag{A.15}$$

Therefore for $|\eta| \leq \frac{1}{4}$,

$$E\left(e^{\eta(\xi_i - T)}\right) \leq e^{2T\eta^2}.$$

It follows that for any $\eta \in (0, \frac{1}{4}]$,

$$\begin{aligned} E\left(\max_{1\leq i\leq m} |\xi_i - T|\right) &= \frac{1}{\eta}\ E\left(\log e^{\max\limits_{1\leq i\leq m} \eta|\xi_i - T|}\right) \\ &\leq \frac{1}{\eta}\log\left(\sum_{i=1}^{m} E\left(e^{\eta(\xi_i - T)} + e^{-\eta(\xi_i - T)}\right)\right) \\ &\leq \frac{1}{\eta}\log\left(\sum_{i=1}^{m} 2e^{2T\eta^2}\right) \\ &= \frac{\log(2m)}{\eta} + 2T\eta. \end{aligned}$$

Choosing $\eta = \sqrt{\log(2m)/2T}$ yields the desired conclusion. □

Lemma 3. *Suppose that $X_j \sim N(0,1)$ for $j = 1, ..., p$, and $Y_k \sim N(0,1)$ for $k = 1, ..., q$. The two sequences $(X_j)_{j=1}^p$ and $(Y_k)_{k=1}^q$ are independent, but X_j's do not need to be independent, neither do Y_k's. Let $\left(X_j^t\right)_{t=1}^T$ and $(Y_k^t)_{t=1}^T$ be i.i.d copies of $(X_j)_{j=1}^p$ and $(Y_k)_{k=1}^q$, respectively. Further assume that* $\log(2pq)/T \leq 1/2$. *Then*

$$E\left(\max_{j,k}\left|\sum_{t=1}^{T} X_j^t Y_k^t\right|\right) \leq 2\sqrt{T\log(2pq)}.$$

Proof. For two independent standard normal random variables Z_1 and Z_2, one has $E(e^{\psi Z_1 Z_2}) = (1-\psi^2)^{-\frac{1}{2}}$ for $|\psi| < 1$. Since $X_j^t Y_k^t$'s are independent for each pair (j,k), letting $\eta_{j,k} = \sum\limits_{t=1}^{T} X_j^t Y_k^t$, we have

$$E(e^{\psi\eta_{j,k}}) = E\left(\prod_{t=1}^{T} e^{\psi X_j^t Y_k^t}\right) = \prod_{t=1}^{T} E\left(e^{\psi X_j^t Y_k^t}\right) = e^{-\frac{T}{2}\log(1-\psi^2)}.$$

By inequality (A.15), for $|\psi| \leq \frac{\sqrt{2}}{2}$, one has $\log(1-\psi^2) \geq -\psi^2(1+\psi^2) \geq -2\psi^2$, hence

$$E(e^{\psi\eta_{j,k}}) \leq e^{T\psi^2}.$$

Thus for any $\psi \in \left[0, \frac{\sqrt{2}}{2}\right]$,

$$\begin{aligned}
E\left(\max_{j,k} |\eta_{j,k}|\right) &= \frac{1}{\psi} E\left(\log e^{\max\limits_{j,k} \psi |\eta_{j,k}|}\right) \\
&\leq \frac{1}{\psi} \log \left(\sum_{j=1}^{p} \sum_{k=1}^{q} E(e^{\psi \eta_{j,k}} + e^{-\psi \eta_{j,k}}) \right) \\
&\leq \frac{1}{\psi} \log \left(\sum_{j=1}^{p} \sum_{k=1}^{q} 2e^{T\psi^2} \right) \\
&= \frac{\log(2pq)}{\psi} + T\psi.
\end{aligned}$$

Choosing $\psi = \sqrt{\log(2pq)/T}$ yields the desired inequality. □

Lemma 4. *Suppose that $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,N})'$, $t = 1, \ldots, T$, are i.i.d random vectors from multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)'$ and $\boldsymbol{\Sigma} = (\sigma_{jk})_{1 \leq j,k \leq N}$. For $j, k = 1, \ldots, N$, define $Y_{j,k} := E(X_{1,j} X_{1,k}) - \frac{1}{T} \sum\limits_{t=1}^{T} X_{t,j} X_{t,k}$. If $\max\limits_{1 \leq j \leq N} |\mu_j| \leq M$, and $\max\limits_{1 \leq j \leq N} |\sigma_{jj}| \leq L$, then*

$$E\left(\max_{1 \leq j,k \leq N} |Y_{j,k}|\right) \leq 2L\sqrt{\frac{2\log(2N)}{T}} + \left(2L + 2M\sqrt{2L}\right)\sqrt{\frac{\log(2N^2)}{T}}. \tag{A.16}$$

Proof. Denote the standard deviation of $X_{1,j}$ by σ_j $(= \sqrt{\sigma_{jj}})$. Note that

$$\sum_{t=1}^{T} X_{t,j} X_{t,k} = \sum_{t=1}^{T} (X_{t,j} - \mu_j)(X_{t,k} - \mu_k) + \sum_{t=1}^{T} (X_{t,j} - \mu_j)\mu_k + \sum_{t=1}^{T} (X_{t,k} - \mu_k)\mu_j + T\mu_j \mu_k,$$

and

$$Y_{j,k} = \sigma_{jk} - \frac{1}{T} \sum_{t=1}^{T} (X_{t,j} - \mu_j)(X_{t,k} - \mu_k) - \frac{1}{T} \sum_{t=1}^{T} (X_{t,j} - \mu_j)\mu_k - \frac{1}{T} \sum_{t=1}^{T} (X_{t,k} - \mu_k)\mu_j.$$

Let $A_{j,k} = \sum\limits_{t=1}^{T} (X_{t,j} - \mu_j)(X_{t,k} - \mu_k)$, $B_{j,k} = \sum\limits_{t=1}^{T} (X_{t,j} - \mu_j)\mu_k$, $C_{j,k} = \sum\limits_{t=1}^{T} (X_{t,k} - \mu_k)\mu_j$, and $\rho_{jk} = \operatorname{corr}(X_{1,j}, X_{1,k})$. Let $(N_{t,j})_{t=1}^{T}, (M_{t,j})_{t=1}^{T}$ consist of i.i.d standard normal random variables. Then

$$\begin{aligned}
A_{j,k} &\overset{\mathcal{D}}{=} \sigma_j \sigma_k \sum_{t=1}^{T} N_{t,j} \left(\rho_{jk} N_{t,j} + \sqrt{1 - \rho_{jk}^2} M_{t,k}\right) \\
&= \sigma_{jk} \sum_{t=1}^{T} (N_{t,j}^2 - 1) + T\sigma_{jk} + \sigma_j \sigma_k \sqrt{1 - \rho_{jk}^2} \sum_{t=1}^{T} N_{t,j} M_{t,k},
\end{aligned}$$

so that

$$\sigma_{jk} - \frac{1}{T}A_{j,k} \overset{\mathcal{D}}{=} -\left(\frac{1}{T}\sigma_{jk}\sum_{t=1}^{T}(N_{t,j}^2 - 1) + \frac{1}{T}\sigma_j\sigma_k\sqrt{1-\rho_{jk}^2}\sum_{t=1}^{T}N_{t,j}M_{t,k}\right). \tag{A.17}$$

By Lemmas 2 and 3, we have

$$E\left(\max_{1\le j\le N}\left|\sum_{t=1}^{T}(N_{t,j}^2 - 1)\right|\right) \le 2\sqrt{2T\log(2N)}, \tag{A.18}$$

and

$$E\left(\max_{1\le j,k\le N}\left|\sum_{t=1}^{T}N_{t,j}M_{t,k}\right|\right) \le 2\sqrt{T\log(2N^2)}. \tag{A.19}$$

Moreover, since $B_{j,k} \sim N(0, T\mu_k^2\sigma_{jj})$ and $C_{j,k} \sim N(0, T\mu_j^2\sigma_{kk})$, by Lemma 1,

$$\max\left(E\left(\max_{1\le j,k\le N}|B_{j,k}|\right), E\left(\max_{1\le j,k\le N}|C_{j,k}|\right)\right) \le M\sqrt{2LT\log(2N^2)}. \tag{A.20}$$

Combining (A.17)–(A.20), we get the desired bound. □

We are now ready to prove Theorem 3.

Proof of Theorem 3. We start with inequality (A.13). Given $\boldsymbol{R}$, define

$$\boldsymbol{S_N} = \{\boldsymbol{Rw} : |w_1| + \cdots + |w_N| \le \lambda\}. \tag{A.21}$$

It is easy to see that $\boldsymbol{S_N}$ is a compact convex subset of $\mathbb{R}^T$. Note that $\widetilde{\boldsymbol{R}} = \boldsymbol{R}\widetilde{\boldsymbol{w}}$ is the projection of $\boldsymbol{r_c}$ onto $\boldsymbol{S_N}$, and $\boldsymbol{R}^* = \boldsymbol{Rw}^*$ is also in $\boldsymbol{S_N}$, therefore $(\boldsymbol{R}^* - \widetilde{\boldsymbol{R}})$ is at an obtuse angle to $(\boldsymbol{r_c} - \widetilde{\boldsymbol{R}})$, that is,

$$\left(\boldsymbol{R}^* - \widetilde{\boldsymbol{R}}\right)'\left(\boldsymbol{r_c} - \widetilde{\boldsymbol{R}}\right) \le 0.$$

Consequently, recalling the definition of r_c in (2.10), we have

$$
\begin{aligned}
||\boldsymbol{R}^* - \widetilde{\boldsymbol{R}}||_2^2 &\leq \left(\boldsymbol{R}^* - \widetilde{\boldsymbol{R}}\right)' (\boldsymbol{R}^* - \boldsymbol{r_c}) \\
&= \left(\boldsymbol{R}^* - \widetilde{\boldsymbol{R}}\right)' (\boldsymbol{R}^* - \boldsymbol{r}^*) - \frac{r^*}{\theta} \left(\boldsymbol{R}^* - \widetilde{\boldsymbol{R}}\right)' \boldsymbol{1} \\
&= \sum_{j=1}^{N} \left(w_j^* - \widetilde{w}_j\right) \sum_{t=1}^{T} R_{tj} \sum_{k=1}^{N} w_k^* (R_{tk} - \mu_k) - \frac{r^*}{\theta} \sum_{j=1}^{N} \left(w_j^* - \widetilde{w}_j\right) \sum_{t=1}^{T} R_{tj} \\
&= \sum_{j=1}^{N} \left(w_j^* - \widetilde{w}_j\right) \sum_{t=1}^{T} \sum_{k=1}^{N} w_k^* \left((R_{tk} - \mu_k)(R_{tj} - \mu_j) - \sigma_{kj}\right) \\
&\quad + \sum_{j=1}^{N} \left(w_j^* - \widetilde{w}_j\right) \sum_{t=1}^{T} \sum_{k=1}^{N} w_k^* (R_{tk} - \mu_k) \mu_j \\
&\quad + \sum_{j=1}^{N} \left(w_j^* - \widetilde{w}_j\right) \sum_{t=1}^{T} \sum_{k=1}^{N} w_k^* \sigma_{kj} - \frac{r^*}{\theta} \sum_{j=1}^{N} \left(w_j^* - \widetilde{w}_j\right) \sum_{t=1}^{T} R_{tj} \\
&:= \sum_{j=1}^{N} \left(w_j^* - \widetilde{w}_j\right) D_{j,1} + \sum_{j=1}^{N} \left(w_j^* - \widetilde{w}_j\right) D_{j,2} + D_3 := D,
\end{aligned}
\tag{A.22}
$$

where

$$
\begin{aligned}
D_{j,1} &= \sum_{t=1}^{T} \sum_{k=1}^{N} w_k^* \left((R_{tk} - \mu_k)(R_{tj} - \mu_j) - \sigma_{kj}\right), \\
D_{j,2} &= \sum_{t=1}^{T} \sum_{k=1}^{N} w_k^* (R_{tk} - \mu_k) \mu_j, \\
D_3 &= \sum_{j=1}^{N} \left(w_j^* - \widetilde{w}_j\right) \sum_{t=1}^{T} \sum_{k=1}^{N} w_k^* \sigma_{kj} - \frac{r^*}{\theta} \sum_{j=1}^{N} \left(w_j^* - \widetilde{w}_j\right) \sum_{t=1}^{T} R_{tj}.
\end{aligned}
\tag{A.23}
$$

For $D_{j,1}$, we shall prove that

$$
E\left(\max_{1 \leq j \leq N} |D_{j,1}|\right) \leq \sigma \sqrt{2LT} \left(1 + \sqrt{2 \log(2N)}\right). \tag{A.24}
$$

To see this, denote

$$
\rho_j = \operatorname{corr}\left(\sum_{k=1}^{N} w_k^* (R_{tk} - \mu_k), R_{tj} - \mu_j\right) = \frac{\sum\limits_{k=1}^{p} w_k^* \sigma_{kj}}{\sigma \sigma_j},
$$

where $\sigma_j = \operatorname{sd}(R_{tj})$ for $j = 1, ..., N$. Suppose that $(Z_t)_{t=1}^{T}, (W_{t,j})_{t=1}^{T},\ j = 1, ..., N$, consist of i.i.d standard normal random variables. Then

$$\begin{aligned}
D_{j,1} &\overset{\mathcal{D}}{=} \sum_{t=1}^{T} \sigma\sigma_j \left[Z_t(\rho_j Z_t + \sqrt{1-\rho_j^2}\; W_{t,j}) - \rho_j \right] \\
&= \sigma\sigma_j\rho_j \sum_{t=1}^{T} (Z_t^2 - 1) + \sigma\sigma_j \sqrt{1-\rho_j^2} \sum_{t=1}^{T} Z_t W_{t,j} \\
&:= V_{j,1} + V_{j,2}.
\end{aligned} \tag{A.25}$$

For $V_{j,1}$, by Assumption A3, we have

$$\begin{aligned}
E\left(\max_{1\le j\le N} |V_{j,1}| \right) &= E\left(\max_{1\le j\le N} \left| \sigma\sigma_j\rho_j \sum_{t=1}^{T} (Z_t^2 - 1) \right| \right) \\
&\le \sigma\sqrt{L} \sqrt{E\left(\sum_{t=1}^{T} (Z_t^2 - 1) \right)^2} \\
&= \sigma\sqrt{2LT}.
\end{aligned} \tag{A.26}$$

Moreover, by Lemma 3,

$$E\left(\max_{1\le j\le N} |V_{j,2}| \right) \le \sigma\sqrt{L}\; E\left(\max_{1\le j\le N} \left| \sum_{t=1}^{T} Z_t W_{t,j} \right| \right) \le 2\sigma\sqrt{LT\log(2N)}. \tag{A.27}$$

Combining inequalities (A.25), (A.26) and (A.27) we get (A.24).

As to $D_{j,2}$, because (w_k^*) is the optimal portfolio, we have $D_{j,2} \sim N(0, T\mu_j^2\sigma^2)$. By Lemma 1 and Assumption A2, we then get

$$E\left(\max_{1\le j\le N} |D_{j,2}| \right) \le \sigma M \sqrt{2T\log(2N)}. \tag{A.28}$$

For D_3, we shall prove that

$$E\,|D_3| \le \frac{\lambda\sigma\sqrt{L}}{\sqrt{\theta}} \sqrt{2T\log(2N)} + \frac{\sigma^2}{\sqrt{\theta}}\sqrt{T}. \tag{A.29}$$

To see this, using the fact that (w_k^*) is the optimal portfolio again, we can rewrite D_3 as

$$\begin{aligned}
D_3 &= \sum_{j=1}^{N} \widetilde{w}_j \left(\sum_{t=1}^{T} \left(\frac{r^*}{\theta} R_{tj} - \sum_{k=1}^{N} w_k^* \sigma_{kj} \right) \right) + \left(T\sigma^2 - \frac{r^*}{\theta} \sum_{t=1}^{T} \sum_{j=1}^{N} R_{tj} w_j^* \right) \\
&:= D_{3a} + D_{3b}.
\end{aligned}$$

Noting that $\boldsymbol{w^*} = \sigma/\sqrt{\theta} \cdot \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and $r^* = \sigma\sqrt{\theta}$, we have

$$D_{3a} = \sum_{j=1}^{N} \widetilde{w}_j \frac{\sigma}{\sqrt{\theta}} \sum_{t=1}^{T} \left(R_{tj} - \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}(,j) \right) = \sum_{j=1}^{N} \widetilde{w}_j \frac{\sigma}{\sqrt{\theta}} \sum_{t=1}^{T} \left(R_{tj} - \mu_j \right),$$

where $\mathbf{\Sigma}(,j)$ is the j-th column of $\mathbf{\Sigma}$. Because $\sum_{j=1}^{N} |\widetilde{w}_j| \leq \lambda$, we have

$$E\,|D_{3a}| \leq \frac{\lambda\sigma}{\sqrt{\theta}} E\left(\max_{1\leq j\leq N}\left|\sum_{t=1}^{T}(R_{tj}-\mu_j)\right|\right) \leq \frac{\lambda\sigma\sqrt{L}}{\sqrt{\theta}}\sqrt{2T\log(2N)}, \tag{A.30}$$

where in the last inequality we again used Lemma 1. Regarding D_{3b}, let $Z_t \overset{i.i.d}{\sim} N(0,1)$ for $t = 1,\ldots,T$, then $\sum_{j=1}^{N} R_{tj}w_j^* \overset{\mathcal{D}}{=} r^* + \sigma Z_t$, thus again noting that $r^* = \sigma\sqrt{\theta}$, we have

$$D_{3b} \overset{\mathcal{D}}{=} T\sigma^2 - \frac{r^*}{\theta}\sum_{t=1}^{T}(r^* + \sigma Z_t) = -\frac{\sigma^2}{\sqrt{\theta}}\sum_{t=1}^{T} Z_t.$$

It follows that

$$E\,|D_{3b}| \leq \frac{\sigma^2}{\sqrt{\theta}}\sqrt{T}. \tag{A.31}$$

Putting (A.30) and (A.31) together we get (A.29).

Combining (A.22), (A.24), (A.28) and (A.29), we obtain

$$\begin{aligned} & E\left(||\boldsymbol{R}^* - \widetilde{\boldsymbol{R}}||_2^2\right) \\ \leq & 2\lambda E\left(\max_{1\leq j\leq N}|D_{j,1}|\right) + 2\lambda E\left(\max_{1\leq j\leq N}|D_{j,2}|\right) + E|D_3| \\ \leq & 2\lambda\sigma\sqrt{2LT} + 2\lambda\sigma\left(\sqrt{2L} + M + \frac{\sqrt{L}}{2\sqrt{\theta}}\right)\sqrt{2T\log(2N)} + \frac{\sigma^2}{\sqrt{\theta}}\sqrt{T}. \end{aligned} \tag{A.32}$$

The desired bound (A.13) follows.

Next we prove inequality (A.14). Let $\mathcal{F}$ be the σ-algebra generated by $(R_{tj})_{1\leq t\leq T, 1\leq j\leq N}$. Let $\widetilde{Q} = \sum_{j=1}^{N}\widetilde{w}_j r_j$, and $Q^* = \sum_{j=1}^{N} w_j^* r_j$, where $(r_1,\ldots,r_N)'$ denotes a future return. Since $\widetilde{\boldsymbol{w}}$ is estimated from observed data, it is independent of $(r_1,\ldots,r_N)'$. We have

$$\begin{aligned} & E\left((Q^* - \widetilde{Q})^2 \mid \mathcal{F}\right) \\ = & \sum_{j,k=1}^{N}(w_j^* - \widetilde{w}_j)(w_k^* - \widetilde{w}_k)E(r_j r_k) \\ = & \sum_{j,k=1}^{N}(w_j^* - \widetilde{w}_j)(w_k^* - \widetilde{w}_k)\left(E\left[(r_j-\mu_j)(r_k-\mu_k)\right] + \mu_j\mu_k\right) \\ = & (\boldsymbol{w}^* - \widetilde{\boldsymbol{w}})'\mathbf{\Sigma}(\boldsymbol{w}^* - \widetilde{\boldsymbol{w}}) + \left(\sum_{j=1}^{N}(w_j^* - \widetilde{w}_j)\mu_j\right)^2 \\ \geqslant & ||\boldsymbol{w}^* - \widetilde{\boldsymbol{w}}||_{\mathbf{\Sigma}}^2. \end{aligned} \tag{A.33}$$

Also note that

$$\frac{1}{T}||\boldsymbol{R^*} - \widetilde{\boldsymbol{R}}||_2^2 = \frac{1}{T}\sum_{t=1}^{T}\sum_{j,k=1}^{N}(w_j^* - \widetilde{w}_j)(w_k^* - \widetilde{w}_k)R_{tj}R_{tk}.$$

Define $Y_{j,k} = E(r_j r_k) - \frac{1}{T}\sum\limits_{t=1}^{T} R_{tj}R_{tk}$. By Assumption A4 and that $\sum\limits_{k=1}^{N}|\widetilde{w}_k| \leq \lambda$, we have

$$\begin{aligned} E\left((Q^* - \widetilde{Q})^2 \mid \mathcal{F}\right) - \frac{1}{T}||\boldsymbol{R^*} - \widetilde{\boldsymbol{R}}||_2^2 &= \sum_{j,k=1}^{N}(w_j^* - \widetilde{w}_j)(w_k^* - \widetilde{w}_k)Y_{j,k} \\ &\leq 4\lambda^2 \max_{1\leq j,k\leq N}|Y_{j,k}|. \end{aligned} \tag{A.34}$$

Therefore by inequality (A.33),

$$\begin{aligned} E||\boldsymbol{w}^* - \widetilde{\boldsymbol{w}}||_{\boldsymbol{\Sigma}}^2 &\leq E\left(Q^* - \widetilde{Q}\right)^2 \\ &\leq \frac{1}{T}E(||\boldsymbol{R}^* - \widetilde{\boldsymbol{R}}||_2^2) + 4\lambda^2 E(\max_{1\leq j,k\leq N}|Y_{j,k}|), \end{aligned}$$

which, combined with (A.13) and Lemma 4, yields the desired inequality (A.14). □

2. Proof of Proposition 7

Proof. Note that

$$|\boldsymbol{\mu}'\boldsymbol{w}^* - \boldsymbol{\mu}'\widetilde{\boldsymbol{w}}| = \left|\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-\frac{1}{2}} \cdot \boldsymbol{\Sigma}^{\frac{1}{2}}\left(\boldsymbol{w}^* - \widetilde{\boldsymbol{w}}\right)\right| \leq \sqrt{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\cdot\left(\boldsymbol{w}^* - \widetilde{\boldsymbol{w}}\right)'\boldsymbol{\Sigma}\left(\boldsymbol{w}^* - \widetilde{\boldsymbol{w}}\right)}.$$

The convergence (A.11) then follows from Assumptions A2, A5 and the bound (A.14) in Theorem 3.

Next we prove (A.12). By the triangular inequality for the norm $||\cdot||_{\boldsymbol{\Sigma}}$, we have

$$|||\boldsymbol{w}^*||_{\boldsymbol{\Sigma}} - ||\widetilde{\boldsymbol{w}}||_{\boldsymbol{\Sigma}}| \leq ||\boldsymbol{w}^* - \widetilde{\boldsymbol{w}}||_{\boldsymbol{\Sigma}}.$$

The convergence (A.12) then again follows from Assumption A5 and the bound A.14.

□

D.2 Proof of Theorem 1

Proof. Consider again the convex set (A.21). Using $\widehat{r_c}$ as response implies that $\widehat{\boldsymbol{R^*}} := \boldsymbol{R}\widehat{\boldsymbol{w^*}}$ is the projection of $\widehat{\boldsymbol{r_c}} = (\widehat{r_c}, \ldots, \widehat{r_c})' \in \mathbb{R}^T$ onto the set $\boldsymbol{S_N}$. By the same reasoning as the proof for (A.22), we have

$$\begin{aligned}||\boldsymbol{R^*} - \widehat{\boldsymbol{R^*}}||_2^2 &\leq (\boldsymbol{R^*} - \widehat{\boldsymbol{R^*}})'(\boldsymbol{R^*} - \boldsymbol{r_c} + \boldsymbol{r_c} - \widehat{\boldsymbol{r_c}}) \\ &= \widehat{D} + (r_c - \widehat{r_c}) \sum_{j=1}^{N} (w_j^* - \widehat{w_j^*}) \left(\sum_{t=1}^{T} R_{tj} \right),\end{aligned}$$

where $\widehat{D}$ is defined in the same way as D in (A.22) by replacing $\widetilde{w}_j$ with $\widehat{w_j^*}$. It follows that

$$\frac{1}{T}||\boldsymbol{R^*} - \widehat{\boldsymbol{R^*}}||_2^2 \leq \frac{1}{T}\widehat{D} + 2\lambda|r_c - \widehat{r_c}| \max_{1\leq j\leq N} \left| \frac{1}{T}\sum_{t=1}^{T} R_{tj} \right|. \tag{A.35}$$

For the first term on the RHS of inequality (A.35), by the same reasoning as for (A.32) and Assumption A4, we have

$$E\left(\frac{1}{T}|\widehat{D}|\right) \leq O(\sqrt{\log(N)/T}). \tag{A.36}$$

For the second term, we rewrite $\frac{1}{T}\sum\limits_{t=1}^{T} R_{tj} = \mu_j + \frac{\sigma_j}{\sqrt{T}} Z_j$, where $Z_j, j = 1, \ldots, N$, are (correlated) standard normal random variables. Then we have

$$\begin{aligned}\max_{1\leq j\leq N} \left| \frac{1}{T}\sum_{t=1}^{T} R_{tj} \right| &= \max_{1\leq j\leq N} \left| \mu_j + \frac{\sigma_j}{\sqrt{T}} Z_j \right| \\ &\leq \max_{1\leq j\leq N} |\mu_j| + \max_{1\leq j\leq N} |\sigma_j| \frac{\max\limits_{1\leq j\leq N} |Z_j|}{\sqrt{T}} \\ &\leq M + \sqrt{L}\ \frac{\max\limits_{1\leq j\leq N} |Z_j|}{\sqrt{T}}.\end{aligned} \tag{A.37}$$

By Lemma 1 we have

$$E\left(\max_{1\leq j\leq N} |Z_j|\right) \leq \sqrt{2\log(2N)}. \tag{A.38}$$

Combining (A.35)∼(A.38) and noting that $\log N/T \to 0$, we get that

$$\frac{1}{T}||\boldsymbol{R^*} - \widehat{\boldsymbol{R^*}}||_2^2 \leq o_p(1) + 2\lambda M\ |r_c - \widehat{r_c}|. \tag{A.39}$$

Now similar to what we did above (A.33), define $\widehat{Q^*} = \sum\limits_{j=1}^{N} \widehat{w_j^*} r_j$, and recall that $Q^* = \sum\limits_{j=1}^{N} w_j^* r_j$, where $(r_1, \ldots, r_N)'$ denotes a future return. By the same arguments as in (A.33) and (A.34) and using (A.37), we have

$$||\boldsymbol{w}^* - \widehat{\boldsymbol{w}^*}||_{\boldsymbol{\Sigma}}^2 \leq E\left((Q^* - \widehat{Q^*})^2 \mid \mathcal{F}\right) \leq \frac{1}{T}||\boldsymbol{R^*} - \widehat{\boldsymbol{R^*}}||_2^2 + 4\lambda^2 \max_{1\leq j,k\leq N} |Y_{j,k}|. \tag{A.40}$$

For the second term, by Lemma 4 we have

$$4\lambda^2 \max_{1\le j,k\le N} |Y_{j,k}| = O_p(\sqrt{\log(N)/T}). \tag{A.41}$$

Combining (A.39), (A.40), (A.41) and noting that $\log N/T \to 0$, we get

$$||\boldsymbol{w}^* - \widehat{\boldsymbol{w}^*}||^2_{\boldsymbol{\Sigma}} \le o_p(1) + 2\lambda M \; |r_c - \widehat{r_c}|. \tag{A.42}$$

Therefore, under the assumptions of Theorem 1, using $|r_c - \widehat{r_c}| \xrightarrow{P} 0$ and Assumption A2, we obtain that

$$|\boldsymbol{\mu}'\boldsymbol{w}^* - \boldsymbol{\mu}'\widehat{\boldsymbol{w}^*}| = \left|\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-\frac{1}{2}} \cdot \boldsymbol{\Sigma}^{\frac{1}{2}}\left(\boldsymbol{w}^* - \widehat{\boldsymbol{w}^*}\right)\right| \le \sqrt{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} \cdot ||\boldsymbol{w}^* - \widehat{\boldsymbol{w}^*}||_{\boldsymbol{\Sigma}} \xrightarrow{P} 0,$$

and

$$\left|\sqrt{\widehat{\boldsymbol{w}^*}'\boldsymbol{\Sigma}\widehat{\boldsymbol{w}^*}} - \sqrt{\boldsymbol{w}^{*\prime}\boldsymbol{\Sigma}\boldsymbol{w}^*}\right| \le ||\boldsymbol{w}^* - \widehat{\boldsymbol{w}^*}||_{\boldsymbol{\Sigma}} \xrightarrow{P} 0.$$

□

Appendix E Proof of Proposition 3

Proof. By (2.16), given asset returns $\boldsymbol{R}$ and factor returns $\boldsymbol{f} := (f_1, \dots, f_K)'$, the return of the portfolio $(\boldsymbol{w}_f, \boldsymbol{w})$ is

$$R_{all} = \boldsymbol{f}'\boldsymbol{w}_f + \boldsymbol{R}'\boldsymbol{w} = \boldsymbol{f}'(\boldsymbol{w}_f + \boldsymbol{\beta}'\boldsymbol{w}) + \boldsymbol{u}'\boldsymbol{w}.$$

Note that the returns on factors and idiosyncratic components are uncorrelated, so the above expression decomposes the full portfolio return into two uncorrelated parts: the return on factors and that on idiosyncratic components. The mean and variance of R_{all} are

$$r_{all} = \boldsymbol{\mu}_f'(\boldsymbol{w}_f + \boldsymbol{\beta}'\boldsymbol{w}) + \boldsymbol{\alpha}'\boldsymbol{w},$$

and

$$\sigma^2 = (\boldsymbol{w}_f + \boldsymbol{\beta}'\boldsymbol{w})'\boldsymbol{\Sigma}_f(\boldsymbol{w}_f + \boldsymbol{\beta}'\boldsymbol{w}) + \boldsymbol{w}'\boldsymbol{\Sigma}_u\boldsymbol{w}.$$

In order to find the optimal weights, let x be the proportion of risk (in terms of variance) allocated to the idiosyncratic components. The maximum expected return on the idiosyncratic components at such a risk level, $\boldsymbol{\alpha}'\boldsymbol{w}$, must be $\sigma\sqrt{x} \cdot \sqrt{\theta_u}$ according to the definition of θ_u. On the other hand, the risk allocated to factors equals $(1-x)\sigma^2$, so the corresponding maximum expected return, $\boldsymbol{\mu}_f'(\boldsymbol{w}_f + \boldsymbol{\beta}'\boldsymbol{w})$, must be $\sigma\sqrt{1-x}\cdot\sqrt{\theta_f}$. Therefore, the optimal proportion of risk on idiosyncratic components, x, must be

$$\arg\max_x(\sigma\sqrt{(1-x)\theta_f} + \sigma\sqrt{x\theta_u}),$$

which can be easily shown to be $x = \theta_u/\theta_{all}$. It then follows that

$$\boldsymbol{w} = \sigma\sqrt{\frac{\theta_u}{\theta_{all}}}\cdot\frac{\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\alpha}}{\sqrt{\theta_u}} = \sigma\sqrt{\frac{\theta_u}{\theta_{all}}}\boldsymbol{w}_u^*,$$

and

$$\boldsymbol{w}_f + \boldsymbol{\beta}'\boldsymbol{w} = \sigma\sqrt{\frac{\theta_f}{\theta_{all}}}\cdot\frac{\boldsymbol{\Sigma}_f^{-1}\boldsymbol{\mu}_f}{\sqrt{\theta_f}} = \sigma\sqrt{\frac{\theta_f}{\theta_{all}}}\boldsymbol{w}_f^*,$$

which are equivalent to the expression of $\boldsymbol{w}_{all}$ in Proposition 3. □

Appendix F Proofs of Propositions 4 and 5

Proposition 4 can be proved exactly in the same way as Proposition 2.

Proof of Proposition 5. The convergence $|\widehat{\theta}_u - \theta_u| \xrightarrow{P} 0$ is a straightforward consequence of the relationship (2.22), Proposition 4 and (2.20). The convergence $|\widehat{r_c} - r_c| \xrightarrow{P} 0$ follows directly.

□

Appendix G Proof of Proposition 6

Similarly to the proof of Theorem 1, we first give the asymptotic properties of an infeasible estimator in the following proposition.

Proposition 8. *Define the infeasible estimator of $\boldsymbol{w}_u^*$ as*

$$\widetilde{\boldsymbol{w}_u} = \arg\min_{\boldsymbol{w}} ||\boldsymbol{r_c} - \widehat{\boldsymbol{U}}\boldsymbol{w}||_2^2 \quad \textit{subject to} \quad ||\boldsymbol{w}||_1 \le \lambda, \tag{A.43}$$

where $\boldsymbol{r_c} = (r_c, ..., r_c)' \in \mathbb{R}^T$. Under Assumptions B1∼B3 and C1∼C3, as $N \to \infty$,

$$|\boldsymbol{\alpha}'\boldsymbol{w}_u^* - \boldsymbol{\alpha}'\widetilde{\boldsymbol{w}_u}| \xrightarrow{P} 0, \tag{A.44}$$

and

$$\left|\sqrt{\widetilde{\boldsymbol{w}_u}'\boldsymbol{\Sigma}_u\widetilde{\boldsymbol{w}_u}} - 1\right| \xrightarrow{P} 0. \tag{A.45}$$

We prove Proposition 8 in Section G.1. Proposition 6 is proved in Section G.2.

G.1 Proof of Proposition 8

1. A Preliminary Theorem

Theorem 4. *Under Assumptions B3, C1 and C2, for the infeasible estimator $\widetilde{\boldsymbol{w}_u}$ defined in* (A.43)*, we have, as $N \to \infty$,*

$$||\boldsymbol{w}_u^* - \widetilde{\boldsymbol{w}_u}||_{\boldsymbol{\Sigma}_u}^2 = O_p\left(\sqrt{\frac{\log N}{T}}\right). \tag{A.46}$$

The following Lemma 5 about $\boldsymbol{\alpha}$ will be used.

Lemma 5. *Suppose $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$ is the vector of alpha's in model* (2.15)*, and $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_N)'$ is the OLS estimate of $\boldsymbol{\alpha}$. Then we have*

$$\max_{1 \le i \le N} |\widehat{\alpha}_i - \alpha_i| = O_p\left(\sqrt{\frac{\log N}{T}}\right). \tag{A.47}$$

Proof. For any fixed $1 \le i \le N$,

$$\widehat{\alpha}_i - \alpha_i = (1, 0, \ldots, 0)\left(\widetilde{\boldsymbol{F}}'\widetilde{\boldsymbol{F}}\right)^{-1}\widetilde{\boldsymbol{F}}' \cdot \boldsymbol{e}_i := \boldsymbol{\Gamma} \cdot \boldsymbol{e}_i,$$

where $\widetilde{\boldsymbol{F}} = (\boldsymbol{1}, \boldsymbol{F})$, $\boldsymbol{1} = (1, \ldots, 1)' \in \mathbb{R}^T$, and in the vector $(1, 0, \ldots, 0)'$ there are K zero's. It follows that, by the multivariate normality of $(e_{1t}, \ldots, e_{Nt})'$, conditional on $\widetilde{\boldsymbol{F}}$,

$$(\widehat{\alpha}_1 - \alpha_1, \ldots, \widehat{\alpha}_N - \alpha_N)' \sim N\left(\boldsymbol{0}, ||\boldsymbol{\Gamma}||_2^2 \cdot \boldsymbol{\Sigma}_e\right).$$

Therefore, by Lemma 1 and Assumption B3,

$$E\left(\max_{1 \le j \le N} |\widehat{\alpha}_i - \alpha_i| \;\middle|\; \widetilde{\boldsymbol{F}}\right) = O\left(\sqrt{||\boldsymbol{\Gamma}||_2^2 \cdot \log N}\right)$$

To prove the lemma, it then suffices to show that $||\boldsymbol{\Gamma}||_2^2 = O_p(1/T)$. Note that

$$T \cdot ||\boldsymbol{\Gamma}||_2^2 = (1, 0, \ldots, 0)\left(\frac{\widetilde{\boldsymbol{F}}'\widetilde{\boldsymbol{F}}}{T}\right)^{-1}(1, 0, \ldots, 0)' \le \left|\left|\left(\frac{\widetilde{\boldsymbol{F}}'\widetilde{\boldsymbol{F}}}{T}\right)^{-1}\right|\right|.$$

By law of large numbers,

$$\frac{\widetilde{\boldsymbol{F}}'\widetilde{\boldsymbol{F}}}{T} \xrightarrow{P} \begin{pmatrix} 1 & \boldsymbol{\mu}_f' \\ \boldsymbol{\mu}_f & \boldsymbol{\mu}_f\boldsymbol{\mu}_f' + \boldsymbol{\Sigma}_f \end{pmatrix}_{(K+1)\times(K+1)} := \boldsymbol{A}.$$

Note that

$$\det(\boldsymbol{A}) = \det\left(\boldsymbol{\mu}_f\boldsymbol{\mu}_f' + \boldsymbol{\Sigma}_f - \boldsymbol{\mu}_f\boldsymbol{\mu}_f'\right) = \det\left(\boldsymbol{\Sigma}_f\right) > 0.$$

Hence $\left|\left|\left(\frac{\widetilde{\boldsymbol{F}}'\widetilde{\boldsymbol{F}}}{T}\right)^{-1}\right|\right| = O_p(1)$ and the proof is completed. □

We are now ready to prove Theorem 4.

Proof of Theorem 4. For clearer presentation, we introduce the following notations and collect some of their relationships:

$$\begin{aligned}
&\boldsymbol{H} := \widetilde{\boldsymbol{F}}\left(\widetilde{\boldsymbol{F}}'\widetilde{\boldsymbol{F}}\right)^{-1}\widetilde{\boldsymbol{F}}',\\
&\widehat{\boldsymbol{e}} := (\widehat{\boldsymbol{e}}_1, \ldots, \widehat{\boldsymbol{e}}_N) = \boldsymbol{e} - \boldsymbol{H}\boldsymbol{e},\\
&\boldsymbol{U} = \mathbf{1}\boldsymbol{\alpha}' + \boldsymbol{e}, \qquad \text{and}\\
&\widehat{\boldsymbol{U}} = \mathbf{1}\widehat{\boldsymbol{\alpha}}' + \widehat{\boldsymbol{e}} = \boldsymbol{U} + \mathbf{1}\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right)' - \boldsymbol{H}\boldsymbol{e},
\end{aligned} \tag{A.48}$$

where $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_N)'$ contains the OLS estimates of intercepts in model (2.16).

We first prove the following result:

$$\frac{1}{T}||\boldsymbol{U}\widetilde{\boldsymbol{w}_u} - \boldsymbol{U}\boldsymbol{w}_u^*||_2^2 = O_p\left(\sqrt{\frac{\log N}{T}}\right). \tag{A.49}$$

In fact, given the residual matrix $\widehat{\boldsymbol{U}}$, define

$$\boldsymbol{S_N} = \left\{\widehat{\boldsymbol{U}}\boldsymbol{w} : |w_1| + \cdots + |w_N| \le \lambda\right\}. \tag{A.50}$$

By the same argument in the proof of Theorem 3, we have

$$\left(\boldsymbol{R^*} - \widetilde{\boldsymbol{R}}\right)'\left(\boldsymbol{r_c} - \widetilde{\boldsymbol{R}}\right) \le 0. \tag{A.51}$$

Plugging in $\widetilde{\boldsymbol{R}} = \widehat{\boldsymbol{U}}\widetilde{\boldsymbol{w}_u}$ and $\boldsymbol{R^*} = \widehat{\boldsymbol{U}}\boldsymbol{w}_u^*$ yields

$$\left((\widehat{\boldsymbol{U}} - \boldsymbol{U})(\boldsymbol{w}_u^* - \widetilde{\boldsymbol{w}_u}) + \boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{U}\widetilde{\boldsymbol{w}_u}\right)'\left(\boldsymbol{r_c} - \widehat{\boldsymbol{U}}\widetilde{\boldsymbol{w}_u}\right) \le 0,$$

or equivalently,

$$\begin{aligned}
&(\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{U}\widetilde{\boldsymbol{w}_u})'\left(\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{U}\widetilde{\boldsymbol{w}_u} + \boldsymbol{r_c} - \boldsymbol{U}\boldsymbol{w}_u^* + \left(\boldsymbol{U} - \widehat{\boldsymbol{U}}\right)\widetilde{\boldsymbol{w}_u}\right)\\
&\le (\boldsymbol{w}_u^* - \widetilde{\boldsymbol{w}_u})'\left(\widehat{\boldsymbol{U}} - \boldsymbol{U}\right)'\left(\widehat{\boldsymbol{U}}\widetilde{\boldsymbol{w}_u} - \boldsymbol{r_c}\right).
\end{aligned}$$

It follows that

$$\begin{aligned}
||\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{U}\widetilde{\boldsymbol{w}_u}||_2^2 \le{}& (\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{U}\widetilde{\boldsymbol{w}_u})'(\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{r_c}) + (\boldsymbol{w}_u^* - \widetilde{\boldsymbol{w}_u})'\boldsymbol{U}'\left(\widehat{\boldsymbol{U}} - \boldsymbol{U}\right)\widetilde{\boldsymbol{w}_u}\\
&+ (\boldsymbol{w}_u^* - \widetilde{\boldsymbol{w}_u})'\left(\widehat{\boldsymbol{U}} - \boldsymbol{U}\right)'\left(\widehat{\boldsymbol{U}}\widetilde{\boldsymbol{w}_u} - \boldsymbol{r_c}\right).
\end{aligned} \tag{A.52}$$

For the first term on the RHS of (A.52), by (A.22), (A.24), (A.28) and (A.29), we have

$$E\left|(\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{U}\widetilde{\boldsymbol{w}_u})'(\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{r_c})\right| \le c\sqrt{T} + \lambda\left(2\sqrt{2L} + 2M + \sqrt{\frac{L}{\theta_u}}\right)\sqrt{2T\log(2N)}, \tag{A.53}$$

Next we prove the following result for the second term on the RHS of (A.52):

$$(\boldsymbol{w}_u^* - \widetilde{\boldsymbol{w}_u})' \boldsymbol{U}' \left(\widehat{\boldsymbol{U}} - \boldsymbol{U}\right) \widetilde{\boldsymbol{w}_u} = O_p\left(\sqrt{T \log N}\right). \tag{A.54}$$

Let $\boldsymbol{V} = \boldsymbol{U}'(\widehat{\boldsymbol{U}} - \boldsymbol{U})\widetilde{\boldsymbol{w}_u}$. By (A.48) we have

$$\begin{aligned} \boldsymbol{V} &= \boldsymbol{U}'\mathbf{1}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})'\widetilde{\boldsymbol{w}_u} - \boldsymbol{\alpha}\mathbf{1}'\boldsymbol{H}\boldsymbol{e}\widetilde{\boldsymbol{w}_u} - \boldsymbol{e}'\boldsymbol{H}\boldsymbol{e}\widetilde{\boldsymbol{w}_u} \\ &:= \boldsymbol{V}_1 - \boldsymbol{V}_2 - \boldsymbol{V}_3. \end{aligned} \tag{A.55}$$

For $\boldsymbol{V}_1$, we have

$$|V_{1,j}| = T\left|\sum_{l=1}^{N}(\widehat{\alpha}_l - \alpha_l)\widetilde{w_{u,l}} \cdot (\alpha_j + \overline{e_j})\right| \leq T\lambda|\alpha_j + \overline{e_j}| \cdot \max_{1\leq l\leq N}|\widehat{\alpha}_l - \alpha_l|,$$

where $\overline{e_j} = \frac{1}{T}\mathbf{1}'\boldsymbol{e}_j$. It follows that

$$\max_{1\leq j\leq N}|V_{1,j}| \leq T\lambda \max_{1\leq j\leq N}|\alpha_j| \cdot \max_{1\leq l\leq N}|\widehat{\alpha}_l - \alpha_l| + T\lambda \max_{1\leq j\leq N}|\overline{e_j}| \cdot \max_{1\leq l\leq N}|\widehat{\alpha}_l - \alpha_l|. \tag{A.56}$$

Noting that $\overline{e_j} \sim N(0, \frac{1}{T}\sigma_j^2)$, by Lemmas 1 and 5 and Assumption B2, we obtain

$$\max_{1\leq j\leq N}|V_{1,j}| = O_p\left(\sqrt{T\log N}\right). \tag{A.57}$$

For $\boldsymbol{V}_2$, we have

$$\max_{1\leq j\leq N}|V_{2,j}| = \max_{1\leq j\leq N}|\alpha_j| \cdot |\mathbf{1}'\boldsymbol{H}\boldsymbol{e}\widetilde{\boldsymbol{w}_u}| \leq \lambda \max_{1\leq j\leq N}|\alpha_j| \cdot \max_{1\leq j\leq N}|q_j|, \tag{A.58}$$

where $\boldsymbol{q} = (q_1, \ldots, q_N)' := \boldsymbol{e}'\boldsymbol{H}\mathbf{1}$. Because $\boldsymbol{H}$ is the projection matrix, we have $\boldsymbol{q} = \boldsymbol{e}'\mathbf{1} = T \cdot (\overline{e_1}, \ldots, \overline{e_N})'$. Again by Lemma 1, we have

$$\max_{1\leq j\leq N}|q_j| = O_p(\sqrt{T\log N}), \tag{A.59}$$

which, combined with Assumption B2 leads to

$$\max_{1\leq j\leq N}|V_{2,j}| = O_p(\sqrt{T\log N}). \tag{A.60}$$

For $\boldsymbol{V}_3$, write the eigendecomposition of $\boldsymbol{H}$ as $\boldsymbol{H} = \boldsymbol{V_H}\boldsymbol{D}\boldsymbol{V_H'}$, where $\boldsymbol{V_H}$ is an orthogonal matrix, and $\boldsymbol{D} = \begin{pmatrix} \boldsymbol{I}_{K+1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$. Let $\widetilde{\boldsymbol{e}}_j = \boldsymbol{V_H}\boldsymbol{e}_j \sim N(\mathbf{0}, \sigma_j^2 \cdot \boldsymbol{I})$. Then $\boldsymbol{V}_3 = \widetilde{\boldsymbol{e}}'\boldsymbol{D}\widetilde{\boldsymbol{e}}\widetilde{\boldsymbol{w}_u}$ and

$$\max_{1\leq j\leq N}|V_{3,j}| \leq \lambda \max_{1\leq j,l\leq N}\left|\widetilde{\boldsymbol{e}}_j'\boldsymbol{D}\widetilde{\boldsymbol{e}}_l\right| = \lambda \max_{1\leq j,l\leq N}\left|\sum_{t=1}^{K+1}\widetilde{e}_{t,j}\widetilde{e}_{t,l}\right|. \tag{A.61}$$

By Lemma 4 and Assumptions B3 and B1, we obtain

$$\max_{1\le j\le N}|V_{3,j}| = O_p\left(\sqrt{2(K+1)\log(2N)}+\sqrt{(K+1)\log(2N^2)}\right) = O_p\left(\sqrt{\log N}\right). \tag{A.62}$$

Combining (A.55), (A.57), (A.60) and (A.62), we get

$$\left|(\boldsymbol{w}_u^* - \widetilde{\boldsymbol{w}_u})' \boldsymbol{U}' \left(\widehat{\boldsymbol{U}} - \boldsymbol{U}\right)\widetilde{\boldsymbol{w}_u}\right| \le 2\lambda \max_{1\le j\le N}|V_j| = O_p\left(\sqrt{T\log N}\right).$$

Next, for the third term on the RHS of (A.52), we shall prove

$$(\boldsymbol{w}_u^* - \widetilde{\boldsymbol{w}_u})' \left(\widehat{\boldsymbol{U}} - \boldsymbol{U}\right)' \left(\widehat{\boldsymbol{U}}\widetilde{\boldsymbol{w}_u} - \boldsymbol{r_c}\right) = O_p\left(\sqrt{T\log N}\right). \tag{A.63}$$

Let $\boldsymbol{g} = \left(\widehat{\boldsymbol{U}} - \boldsymbol{U}\right)' \left(\widehat{\boldsymbol{U}}\widetilde{\boldsymbol{w}_u} - \boldsymbol{r_c}\right)$. By (A.48) we have

$$\begin{aligned}\boldsymbol{g} &= (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})' \mathbf{1}'\boldsymbol{U}\widetilde{\boldsymbol{w}_u} + T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})'\widetilde{\boldsymbol{w}_u} - (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})'\mathbf{1}'\boldsymbol{H}\boldsymbol{e}\widetilde{\boldsymbol{w}_u} - \boldsymbol{e}'\boldsymbol{H}\mathbf{1}\widehat{\boldsymbol{\alpha}}'\widetilde{\boldsymbol{w}_u} \\ &\quad - r_c T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + r_c\boldsymbol{e}'\boldsymbol{H}\mathbf{1} \\ &:= \boldsymbol{g}_1 + \boldsymbol{g}_2 - \boldsymbol{g}_3 - \boldsymbol{g}_4 - \boldsymbol{g}_5 + \boldsymbol{g}_6.\end{aligned} \tag{A.64}$$

For $\boldsymbol{g}_1$, similarly to (A.56), (A.57) and by Lemmas 1 and 5 and Assumption B2, we have

$$\begin{aligned}\max_{1\le j\le N}|g_{1,j}| &= \max_{1\le j\le N}|\widehat{\alpha}_j - \alpha_j| \cdot |\mathbf{1}'\boldsymbol{U}\widetilde{\boldsymbol{w}_u}| \\ &\le T\lambda \max_{1\le j\le N}|\widehat{\alpha}_j - \alpha_j| \max_{1\le l\le N}|\alpha_l + \overline{e_l}| \\ &\le T\lambda \max_{1\le l\le N}|\alpha_l| \cdot \max_{1\le j\le N}|\widehat{\alpha}_j - \alpha_j| + T\lambda\max_{1\le l\le N}|\overline{e_l}| \cdot \max_{1\le j\le N}|\widehat{\alpha}_j - \alpha_j| \\ &= O_p\left(\sqrt{T\log N}\right).\end{aligned} \tag{A.65}$$

For $\boldsymbol{g}_2$, by Lemma 5, we have

$$\begin{aligned}\max_{1\le j\le N}|g_{2,j}| &= T\max_{1\le j\le N}|\widehat{\alpha}_j - \alpha_j| \cdot |(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})'\widetilde{\boldsymbol{w}_u}| \\ &\le T\lambda\max_{1\le j\le N}(\widehat{\alpha}_j - \alpha_j)^2 \\ &= O_p(\log N).\end{aligned} \tag{A.66}$$

For $\boldsymbol{g}_3$, by (A.59) and Lemma 5, we have

$$\begin{aligned}\max_{1\le j\le N}|g_{3,j}| &= \max_{1\le j\le N}|\widehat{\alpha}_j - \alpha_j| \cdot |\mathbf{1}'\boldsymbol{H}\boldsymbol{e}\widetilde{\boldsymbol{w}_u}| \\ &\le \lambda\max_{1\le j\le N}|\widehat{\alpha}_j - \alpha_j| \cdot \max_{1\le j\le N}|q_j| \\ &= O_p(\log N).\end{aligned} \tag{A.67}$$

For $\boldsymbol{g}_4$, by (A.59), Assumption B2 and Lemma 5, we have

$$\begin{aligned}
\max_{1\leq j\leq N}|g_{4,j}| &= \max_{1\leq j\leq N}|q_j|\cdot|\widehat{\boldsymbol{\alpha}}'\widetilde{\boldsymbol{w}_u}| \\
&\leq \lambda\max_{1\leq j\leq N}|q_j|\left(\max_{1\leq j\leq N}|\alpha_j|+\max_{1\leq j\leq N}|\widehat{\alpha}_j-\alpha_j|\right) \\
&= O_p\left(\sqrt{T\log N}\left(1+\sqrt{\frac{\log N}{T}}\right)\right) \\
&= O_p\left(\sqrt{T\log N}\right).
\end{aligned} \tag{A.68}$$

For $\boldsymbol{g}_5$, by Lemma 5 we have

$$\max_{1\leq j\leq N}|g_{5,j}| = r_cT\cdot\max_{1\leq j\leq N}|\widehat{\alpha}_j-\alpha_j| = O_p\left(\sqrt{T\log N}\right). \tag{A.69}$$

For $\boldsymbol{g}_6$, by (A.59) we have

$$\max_{1\leq j\leq N}|g_{6,j}| = r_c\max_{1\leq j\leq N}|q_j| = O_p\left(\sqrt{T\log N}\right). \tag{A.70}$$

Combining (A.64) and (A.65)∼(A.70) and noting that $||\boldsymbol{w}_u^*||_1\leq\lambda$ and $||\widetilde{\boldsymbol{w}_u}||_1\leq\lambda$, we obtain

$$\left|(\boldsymbol{w}_u^*-\widetilde{\boldsymbol{w}_u})'\left(\widehat{\boldsymbol{U}}-\boldsymbol{U}\right)'\left(\widehat{\boldsymbol{U}}\widetilde{\boldsymbol{w}_u}-\boldsymbol{r_c}\right)\right| \leq 2\lambda\max_{1\leq j\leq N}|g_j| = O_p\left(\sqrt{T\log N}\right).$$

Combining (A.52), (A.53), (A.54) and (A.63) yields the desired bound (A.49).

Next we prove (A.46). Let $\mathcal{F}$ be the σ-algebra generated by $(U_{tj})_{1\leq t\leq T,1\leq j\leq N}$. Let $\widetilde{Q}=\sum_{j=1}^N\widetilde{w_{u,j}}u_j$, and $Q^*=\sum_{j=1}^N w_{u,j}^*u_j$, where $(u_1,\ldots,u_N)'$ denotes a future idiosyncratic return. Since $\widetilde{\boldsymbol{w}_u}$ is estimated from observed data, it is independent of $(u_1,\ldots,u_N)'$. It follows that

$$\begin{aligned}
&E\left((Q^*-\widetilde{Q})^2\mid\mathcal{F}\right) \\
=\ &\sum_{j,k=1}^N(w_{u,j}^*-\widetilde{w_{u,j}})(w_{u,k}^*-\widetilde{w_{u,k}})E(u_ju_k) \\
=\ &\sum_{j,k=1}^N(w_{u,j}^*-\widetilde{w_{u,j}})(w_{u,k}^*-\widetilde{w_{u,k}})\left(E\left[(u_j-\alpha_j)(u_k-\alpha_k)\right]+\alpha_j\alpha_k\right) \\
=\ &(\boldsymbol{w}_u^*-\widetilde{\boldsymbol{w}_u})'\boldsymbol{\Sigma}_u(\boldsymbol{w}_u^*-\widetilde{\boldsymbol{w}_u})+\left(\sum_{j=1}^N(w_{u,j}^*-\widetilde{w_{u,j}})\alpha_j\right)^2 \\
\geqslant\ &||\boldsymbol{w}_u^*-\widetilde{\boldsymbol{w}_u}||_{\boldsymbol{\Sigma}}^2.
\end{aligned} \tag{A.71}$$

Also note that

$$\frac{1}{T}||\boldsymbol{U}\boldsymbol{w}_u^*-\boldsymbol{U}\widetilde{\boldsymbol{w}_u}||_2^2=\frac{1}{T}\sum_{t=1}^T\sum_{j,k=1}^N(w_{u,j}^*-\widetilde{w_{u,j}})(w_{u,k}^*-\widetilde{w_{u,k}})U_{tj}U_{tk}.$$

Define $Y_{j,k} = E(u_j u_k) - \frac{1}{T}\sum\limits_{t=1}^{T} U_{tj}U_{tk}$. Because $||\boldsymbol{w}_u^*||_1 \leq \lambda$ and $||\widetilde{\boldsymbol{w}_u}||_1 \leq \lambda$, we obtain

$$\begin{aligned} E\left((Q^* - \widetilde{Q})^2 \mid \mathcal{F}\right) - \frac{1}{T}||\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{U}\widetilde{\boldsymbol{w}_u}||_2^2 &= \sum_{j,k=1}^{N}(w_{u,j}^* - \widetilde{w_{u,j}})(w_{u,k}^* - \widetilde{w_{u,k}})Y_{j,k} \\ &\leq 4\lambda^2 \max_{1\leq j,k\leq N}|Y_{j,k}|. \end{aligned} \tag{A.72}$$

Therefore by inequality (A.71),

$$\begin{aligned} ||\boldsymbol{w}_u^* - \widetilde{\boldsymbol{w}_u}||_{\boldsymbol{\Sigma}}^2 &\leq E\left(\left(Q^* - \widetilde{Q}\right)^2 \,\middle|\, \widetilde{\boldsymbol{F}}\right) \\ &\leq \frac{1}{T}||\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{U}\widetilde{\boldsymbol{w}_u}||_2^2 + 4\lambda^2 \cdot \max_{1\leq j,k\leq N}|Y_{j,k}|, \end{aligned}$$

which, combined with (A.49) and Lemma 4, yields the desired bound (A.46). □

2. Proof of Proposition 8

Proposition 8 can be proved in exactly the same way as Proposition 7 using Assumptions B2, C3 and the bound (A.46).

G.2 Proof of Proposition 6

Proof of Proposition 6. Consider again the convex set $\boldsymbol{S_N}$ in (A.50). Using $\widehat{r_c}$ as response implies that $\widehat{\boldsymbol{R}^*} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{w}_u^*}$ is the projection of $\widehat{\boldsymbol{r_c}}$ onto the set $\boldsymbol{S_N}$. By the same reasoning as the derivation for (A.52), we have

$$\begin{aligned} &||\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{U}\widehat{\boldsymbol{w}_u^*}||_2^2 \\ &\leq (\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{U}\widehat{\boldsymbol{w}_u^*})'(\boldsymbol{U}\boldsymbol{w}_u^* - \widehat{\boldsymbol{r_c}}) + \left(\boldsymbol{w}_u^* - \widehat{\boldsymbol{w}_u^*}\right)'\boldsymbol{U}'\left(\widehat{\boldsymbol{U}} - \boldsymbol{U}\right)\widehat{\boldsymbol{w}_u^*} \\ &\quad + \left(\boldsymbol{w}_u^* - \widehat{\boldsymbol{w}_u^*}\right)'\left(\widehat{\boldsymbol{U}} - \boldsymbol{U}\right)'\left(\widehat{\boldsymbol{U}}\widehat{\boldsymbol{w}_u^*} - \widehat{\boldsymbol{r_c}}\right). \end{aligned} \tag{A.73}$$

For the first term on the RHS of (A.73), we have

$$\begin{aligned} \frac{1}{T}\left|(\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{U}\widehat{\boldsymbol{w}_u^*})'(\boldsymbol{U}\boldsymbol{w}_u^* - \widehat{\boldsymbol{r_c}})\right| &= \frac{1}{T}\left|(\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{U}\widehat{\boldsymbol{w}_u^*})'(\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{r_c} + \boldsymbol{r_c} - \widehat{\boldsymbol{r_c}})\right| \\ &= \frac{1}{T}\left|\widehat{G} + (r_c - \widehat{r_c})\sum_{j=1}^{N}(w_{u,j}^* - \widehat{w_{u,j}^*})\left(\sum_{t=1}^{T}U_{tj}\right)\right| \\ &\leq \frac{1}{T}|\widehat{G}| + 2\lambda|r_c - \widehat{r_c}|\max_{1\leq j\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}U_{tj}\right|, \end{aligned} \tag{A.74}$$

where $\widehat{G} = (\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{U}\widehat{\boldsymbol{w}_u^*})'(\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{r_c})$, which admits a similar decomposition to (A.22). By the same reasoning as for (A.32) and Assumption C2, we have

$$\frac{1}{T}|\widehat{G}| = O_p\left(\sqrt{\frac{\log N}{T}}\right). \tag{A.75}$$

Furthermore, by Lemma 1 and Assumption B2, we get

$$\max_{1\le j\le N}\left|\frac{1}{T}\sum_{t=1}^{T} U_{tj}\right| = O_p\left(1+\sqrt{\frac{\log N}{T}}\right). \tag{A.76}$$

Combining (A.74)∼(A.76) and noting that $\log N/T \to 0$, we get that

$$\frac{1}{T}\left|(\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{U}\widehat{\boldsymbol{w}_u^*})'(\boldsymbol{U}\boldsymbol{w}_u^* - \widehat{\boldsymbol{r}_{\boldsymbol{c}}})\right| = O_p\left(|r_c - \widehat{r_c}|\right). \tag{A.77}$$

For the second term on the RHS of (A.73), by the same arguments as for (A.54), we have

$$\frac{1}{T}\left(\boldsymbol{w}_u^* - \widehat{\boldsymbol{w}_u^*}\right)'\boldsymbol{U}'\left(\widehat{\boldsymbol{U}} - \boldsymbol{U}\right)\widehat{\boldsymbol{w}_u^*} = O_p\left(\sqrt{\frac{\log N}{T}}\right) \xrightarrow{P} 0. \tag{A.78}$$

As to the third term on the RHS of (A.73), decompose $\widehat{\boldsymbol{g}} := (\widehat{\boldsymbol{U}} - \boldsymbol{U})'(\widehat{\boldsymbol{U}}\widehat{\boldsymbol{w}_u^*} - \widehat{\boldsymbol{r}_{\boldsymbol{c}}})$ similarly to (A.64) with $\widetilde{\boldsymbol{w}_u}$ and $\boldsymbol{r_c}$ replaced by $\widehat{\boldsymbol{w}_u^*}$ and $\widehat{\boldsymbol{r}_{\boldsymbol{c}}}$, respectively. The first four components can be bounded in the same way as in (A.65), (A.66), (A.67) and (A.68). As to $\widehat{\boldsymbol{g}}_5 = \widehat{r_c}T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$ and $\widehat{\boldsymbol{g}}_6 = \widehat{r_c}\boldsymbol{e}'\boldsymbol{H}\boldsymbol{1}$, by Lemma 5 and (A.59) and using the assumption that $|r_c - \widehat{r_c}| \xrightarrow{P} 0$, we obtain

$$\max_{1\le j\le N}|\widehat{g}_{5,j}| \le T\left(|\widehat{r_c} - r_c| + r_c\right)\cdot \max_{1\le j\le N}|\widehat{\alpha}_j - \alpha_j| = O_p\left(\sqrt{T\log N}\right),$$

and

$$\max_{1\le j\le N}|\widehat{g}_{6,j}| \le \left(|\widehat{r_c} - r_c| + r_c\right)\cdot \max_{1\le j\le N}|q_j| = O_p\left(\sqrt{T\log N}\right).$$

It follows that

$$\left|\frac{1}{T}\left(\boldsymbol{w}_u^* - \widehat{\boldsymbol{w}_u^*}\right)'\left(\widehat{\boldsymbol{U}} - \boldsymbol{U}\right)'\left(\widehat{\boldsymbol{U}}\widehat{\boldsymbol{w}_u^*} - \widehat{\boldsymbol{r}_{\boldsymbol{c}}}\right)\right| \le 2\lambda \max_{1\le j\le N}|\widehat{g}_j| = O_p\left(\sqrt{\frac{\log N}{T}}\right). \tag{A.79}$$

Combining (A.73), (A.77), (A.78) and (A.79), and using the assumption that $|r_c - \widehat{r_c}| \xrightarrow{P} 0$, we obtain

$$\frac{1}{T}||\boldsymbol{U}\boldsymbol{w}_u^* - \boldsymbol{U}\widehat{\boldsymbol{w}_u^*}||_2^2 \xrightarrow{P} 0. \tag{A.80}$$

Now define $\widehat{Q^*} = \sum_{j=1}^{N} \widehat{w^*_{u,j}} u_j$, and recall that $Q^* = \sum_{j=1}^{N} w^*_{u,j} u_j$, where $(u_1, \dots, u_N)'$ denotes a future idiosyncratic return. By the same arguments as in (A.71) and (A.72) and using (A.76), we have

$$||\boldsymbol{w}^*_u - \widehat{\boldsymbol{w}^*_u}||^2_{\boldsymbol{\Sigma}_u} \leq E\left((Q^* - \widehat{Q^*})^2 \mid \mathcal{F}\right) \leq \frac{1}{T}||\boldsymbol{U}\boldsymbol{w}^*_u - \boldsymbol{U}\widehat{\boldsymbol{w}^*_u}||^2_2 + 4\lambda^2 \max_{1\leq j,k\leq N} |Y_{j,k}|. \quad \text{(A.81)}$$

Combining (A.80), (A.81) and Lemma 3, we get

$$||\boldsymbol{w}^*_u - \widehat{\boldsymbol{w}^*_u}||^2_{\boldsymbol{\Sigma}_u} \xrightarrow{P} 0. \quad \text{(A.82)}$$

Therefore,

$$|\boldsymbol{\alpha}'\boldsymbol{w}^*_u - \boldsymbol{\alpha}'\widehat{\boldsymbol{w}^*_u}| = \left|\boldsymbol{\alpha}'\boldsymbol{\Sigma}_u^{-\frac{1}{2}} \cdot \boldsymbol{\Sigma}_u^{\frac{1}{2}}(\boldsymbol{w}^*_u - \widehat{\boldsymbol{w}^*_u})\right| \leq \sqrt{\boldsymbol{\alpha}'\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\alpha} \cdot (\boldsymbol{w}^*_u - \widehat{\boldsymbol{w}^*_u})'\boldsymbol{\Sigma}_u(\boldsymbol{w}^*_u - \widehat{\boldsymbol{w}^*_u})} \xrightarrow{P} 0,$$

and

$$\left|\sqrt{\widehat{\boldsymbol{w}^*_u}'\boldsymbol{\Sigma}_u\widehat{\boldsymbol{w}^*_u}} - 1\right| \leq ||\boldsymbol{w}^*_u - \widehat{\boldsymbol{w}^*_u}||_{\boldsymbol{\Sigma}_u} \xrightarrow{P} 0.$$

□

Appendix H Proof of Theorem 2

Proof. By the factor model (2.16), the return of portfolio $\widehat{\boldsymbol{w}_{all}}$ equals

$$\begin{aligned} R &= \sigma\left(\sqrt{\frac{\widehat{\theta}_f}{\widehat{\theta}_{all}}}\widehat{\boldsymbol{w}^*_f} - \sqrt{\frac{\widehat{\theta}_u}{\widehat{\theta}_{all}}}\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{w}^*_u}\right)'\boldsymbol{f} + \sigma\sqrt{\frac{\widehat{\theta}_u}{\widehat{\theta}_{all}}}\widehat{\boldsymbol{w}^*_u}'\boldsymbol{r} \\ &= \sigma\left(\sqrt{\frac{\widehat{\theta}_f}{\widehat{\theta}_{all}}}\widehat{\boldsymbol{w}^*_f} - \sqrt{\frac{\widehat{\theta}_u}{\widehat{\theta}_{all}}}\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{w}^*_u} + \sqrt{\frac{\widehat{\theta}_u}{\widehat{\theta}_{all}}}\boldsymbol{\beta}'\widehat{\boldsymbol{w}^*_u}\right)'\boldsymbol{f} + \sigma\sqrt{\frac{\widehat{\theta}_u}{\widehat{\theta}_{all}}}\widehat{\boldsymbol{w}^*_u}'\boldsymbol{u}. \end{aligned}$$

We first prove that

$$\left|\left|\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{w}^*_u} - \boldsymbol{\beta}'\widehat{\boldsymbol{w}^*_u}\right|\right|_2 \xrightarrow{P} 0. \quad \text{(A.83)}$$

Note that $\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{w}^*_u} - \boldsymbol{\beta}'\widehat{\boldsymbol{w}^*_u}$ is a vector of dimension K. Because K is bounded, we only need to prove that for each $j = 1, \dots, K$,

$$\left|\sum_{i=1}^{N}\left(\widehat{\beta_{ij}} - \beta_{ij}\right) \cdot \widehat{w^*_{u,i}}\right| \xrightarrow{P} 0. \quad \text{(A.84)}$$

Recall that $||\widehat{\boldsymbol{w}_u^*}||_1 \leq \lambda$. Therefore

$$\left|\sum_{i=1}^{N}\left(\widehat{\beta_{ij}}-\beta_{ij}\right)\cdot\widehat{w_{u,i}^*}\right| \leq \lambda\cdot\max_{1\leq i\leq N}\left|\widehat{\beta_{ij}}-\beta_{ij}\right|.$$

By using a similar proof to Lemma 5, one can show that

$$\max_{1\leq i\leq N}\left|\widehat{\beta_{ij}}-\beta_{ij}\right| = O_p\left(\sqrt{\frac{\log N}{T}}\right)\xrightarrow{P} 0.$$

Consequently, (A.84) and (A.83) hold.

We now prove the convergences of both mean and variance of R. We have

$$E\left(R\right) = \sigma\left(\sqrt{\frac{\widehat{\theta}_f}{\widehat{\theta}_{all}}}\widehat{\boldsymbol{w}_f^*} - \sqrt{\frac{\widehat{\theta}_u}{\widehat{\theta}_{all}}}\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{w}_u^*} + \sqrt{\frac{\widehat{\theta}_u}{\widehat{\theta}_{all}}}\boldsymbol{\beta}'\widehat{\boldsymbol{w}_u^*}\right)'\boldsymbol{\mu}_f + \sigma\sqrt{\frac{\widehat{\theta}_u}{\widehat{\theta}_{all}}}\widehat{\boldsymbol{w}_u^*}'\boldsymbol{\alpha}.$$

By the consistencies of $\widehat{\theta}_f$, $\widehat{\theta}_u$, $\widehat{\theta}_{all}$, $\widehat{\boldsymbol{w}_f^*}$, ,the convergences (2.28) and (A.83), and the relationship (2.22), we obtain

$$E\left(R\right)\xrightarrow{P}\sigma\sqrt{\frac{\theta_f}{\theta_{all}}}\sqrt{\theta_f} + \sigma\sqrt{\frac{\theta_u}{\theta_{all}}}\sqrt{\theta_u} = \sigma\sqrt{\theta_{all}} = r^*.$$

As to its variance, we have

$$\begin{aligned}&\mathrm{Var}\left(R\right)\\ =&\ \sigma^2\left(\sqrt{\frac{\widehat{\theta}_f}{\widehat{\theta}_{all}}}\widehat{\boldsymbol{w}_f^*} - \sqrt{\frac{\widehat{\theta}_u}{\widehat{\theta}_{all}}}\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{w}_u^*} + \sqrt{\frac{\widehat{\theta}_u}{\widehat{\theta}_{all}}}\boldsymbol{\beta}'\widehat{\boldsymbol{w}_u^*}\right)'\boldsymbol{\Sigma}_f\left(\sqrt{\frac{\widehat{\theta}_f}{\widehat{\theta}_{all}}}\widehat{\boldsymbol{w}_f^*} - \sqrt{\frac{\widehat{\theta}_u}{\widehat{\theta}_{all}}}\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{w}_u^*} + \sqrt{\frac{\widehat{\theta}_u}{\widehat{\theta}_{all}}}\boldsymbol{\beta}'\widehat{\boldsymbol{w}_u^*}\right)\\ &+\sigma^2\frac{\widehat{\theta}_u}{\widehat{\theta}_{all}}\widehat{\boldsymbol{w}_u^*}'\boldsymbol{\Sigma}_u\widehat{\boldsymbol{w}_u^*}.\end{aligned}$$

Using again the consistencies of $\widehat{\theta}_f$, $\widehat{\theta}_u$, $\widehat{\theta}_{all}$, $\widehat{\boldsymbol{w}_f^*}$, the convergences (2.29) and (A.83), and the relationship (2.22), we have

$$\mathrm{Var}\left(R\right)\xrightarrow{P}\sigma^2\frac{\theta_f}{\theta_{all}}\cdot 1 + \sigma^2\frac{\theta_u}{\theta_{all}}\cdot 1 = \sigma^2.$$

□

References

Bai, Z., Li, H., and Wong, W.-K. "The best estimation for high-dimensional Markowitz mean-variance optimization." (2013).

Bai, Z., Liu, H., and Wong, W.-K. "Enhancement of the applicability of Markowitz's portfolio optimization by utilizing random matrix theory." *Mathematical Finance*, 19(4):639–667 (2009).
URL http://dx.doi.org/10.1111/j.1467-9965.2009.00383.x

Basak, G. K., Jagannathan, R., and Ma, T. "Jackknife estimator for tracking error variance of optimal portfolios." *Management Science*, 55(6):990–1002 (2009).

Best, M. J. and Grauer, R. R. "On the sensitivity of mean-variance-efficient portfolios to changes in asset means: some analytical and computational results." *Review of Financial Studies*, 4(2):315–342 (1991).

Black, F. and Litterman, R. B. "Asset allocation: combining investor views with market equilibrium." *The Journal of Fixed Income*, 1(2):7–18 (1991).

Britten-Jones, M. "The Sampling Error in Estimates of Mean-Variance Efficient Portfolio Weights." *The Journal of Finance*, 54(2):655–671 (1999).

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

Chatterjee, S. "Assumptionless consistency of the Lasso." *arXiv preprint arXiv:1303.5817* (2013).

Chopra, V. K. and Ziemba, W. T. "The effect of errors in means, variances, and covariances on optimal portfolio choice." *The Journal of Portfolio Management*, 19(2):6–11 (1993).

DeMiguel, V., Garlappi, L., Nogales, F. J., and Uppal, R. "A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms." *Management Science*, 55(5):798–812 (2009).

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. "Least angle regression." *The Annals of Statistics*, 32(2):407–499 (2004). With discussion, and a rejoinder by the authors.
URL http://dx.doi.org/10.1214/009053604000000067

El Karoui, N. "High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: Risk underestimation." *The Annals of Statistics*, 38(6):3487–

3566 (2010).
URL http://dx.doi.org/10.1214/10-AOS795

—. “On the realized risk of high-dimensional Markowitz portfolios.” *SIAM Journal on Financial Mathematics*, 4(1):737–783 (2013).
URL http://dx.doi.org/10.1137/090774926

Engle, R., Ferstenberg, R., and Russell, J. “Measuring and Modeling Execution Cost and Risk.” *The Journal of Portfolio Management*, 38(2):14–28 (2012).

Fan, J., Li, Y., and Yu, K. “Vast volatility matrix estimation using high-frequency data for portfolio selection.” *Journal of the American Statistical Association*, 107(497):412–428 (2012a).
URL http://dx.doi.org/10.1080/01621459.2012.656041

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

Fastrich, B., Paterlini, S., and Winker, P. “Constructing Optimal Sparse Portfolios Using Regularization Methods.” *Available at SSRN 2169062* (2012).

Feng, G., Giglio, S., and Xiu, D. “Taming the Factor Zoo.” *Fama-Miller Working Paper Forthcoming; Chicago Booth Research Paper No. 17-04.* (2017).
URL https://ssrn.com/abstract=2934020

Garlappi, L., Uppal, R., and Wang, T. “Portfolio selection with parameter and model uncertainty: A multi-prior approach.” *Review of Financial Studies*, 20(1):41–81 (2007).

Green, R. C. and Hollifield, B. “When Will Mean-Variance Efficient Portfolios Be Well Diversified?” *The Journal of Finance*, 47(5):1785–1809 (1992).

Jagannathan, R. and Ma, T. “Risk reduction in large portfolios: Why imposing the wrong constraints helps.” *The Journal of Finance*, 58(4):1651–1684 (2003).

Jegadeesh, N. and Titman, S. “Profitability of momentum strategies: An evaluation of alternative explanations.” *The Journal of Finance*, 56(2):699–720 (2001).

Jobson, J. D. and Korkie, B. M. “Performance hypothesis testing with the Sharpe and Treynor measures.” *The Journal of Finance*, 36(4):889–908 (1981).

Kan, R. and Zhou, G. “Optimal portfolio choice with parameter uncertainty.” *Journal of Financial and Quantitative Analysis*, 42(3):621 (2007).

Korajczyk, R. A. and Sadka, R. “Pricing the commonality across alternative measures of liquidity.” *Journal of Financial Economics*, 87(1):45–72 (2008).

Lai, T. L., Xing, H., and Chen, Z. “Mean–variance portfolio optimization when means and covariances are unknown.” *The Annals of Applied Statistics*, 5(2A):798–823 (2011).
URL `http://dx.doi.org/10.1214/10-AOAS422`

Ledoit, O. and Wolf, M. “Improved estimation of the covariance matrix of stock returns with an application to portfolio selection.” *Journal of Empirical Finance*, 10(5):603–621 (2003).

—. “A well-conditioned estimator for large-dimensional covariance matrices.” *Journal of Multivariate Analysis*, 88(2):365–411 (2004).
URL `http://dx.doi.org/10.1016/S0047-259X(03)00096-4`

—. “Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks.” *The Review of Financial Studies* (2017).

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

Memmel, C. “Performance hypothesis testing with the Sharpe ratio.” *Finance Letters*, 1(1) (2003).

Michaud, R. O. “The Markowitz optimization enigma: is “optimized” optimal?” *Financial Analysts Journal*, 31–42 (1989).

Rothman, A. J. “Positive definite estimators of large covariance matrices.” *Biometrika*, 99(3):733–740 (2012).

Tibshirani, R. “Regression shrinkage and selection via the lasso.” *Journal of the Royal Statistical Society. Series B. Methodological*, 58(1):267–288 (1996).

Tu, J. and Zhou, G. “Markowitz meets Talmud: A combination of sophisticated and naive diversification strategies.” *Journal of Financial Economics*, 99(1):204–215 (2011).