Efficient Portfolio Selection in a Large Market†

Jiaqin Chen and Ming Yuan§

Bank of America and University of Wisconsin-Madison

(March 27, 2016)

Abstract

Recent empirical studies show that the estimated Markowitz mean-variance portfolios oftentimes perform rather poorly when there are more than several assets in the investment universe. In this article, we argue that such disappointing performance can be largely attributed to the estimation error incurred in sample mean-variance portfolios, and therefore could be improved by utilizing more efficient estimating strategies. In particular, we show that this "Markowitz optimization enigma" (Michaud, 1998) could be resolved by carefully balancing the tradeoff between the estimation error and systematic error through the so-called subspace mean-variance analysis. In addition to the consistent improvement observed on real and simulated data sets, we prove that, under an approximate factor model, it is possible to use this strategy to construct portfolio rules whose performance closely resemble that of theoretical mean-variance efficient portfolios in a large market.

Keywords: Approximate factor model, asymptotic efficiency, beta pricing model, estimation error, mean-variance analysis, Sharpe ratio. (*JEL* G11)

---

† This research was supported in part by NSF CAREER Award DMS-1321692 and FRG Grant DMS-1265202.

§ Address for correspondence: Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706. (Email: myuan@stat.wisc.edu).

1 Introduction

Although the importance of diversification in investment has long been recognized (see, e.g., Rubinstein, 2006), how to optimally allocate wealth across a universe of risky assets in a principled way remained elusive until the seminal work of Markowitz (1952; 1959). See also Roy (1952). The mean-variance efficient portfolios derived from such analysis are determined by the mean and variance of the returns of the underlying assets. In practice, both moments need to be estimated from historical data, by either the usual sample moments or more sophisticated estimators. Examples include Barry (1974), Brown (1976), Bawa, Brown and Klein (1979), Jobson, Korkie, and Ratti (1979), Jobson and Korkie (1980), Jorion (1985; 1986), McKinlay and Pastor (2000), Goldfarb and Iyengar (2003), Ledoit and Wolf (2004), Garlappi, Uppal, and Wang (2007), and Kan and Zhou (2007) among many others. See also Brandt (2010) for a recent review of the developments.

The practical merits of these estimated mean-variance efficient portfolio strategies, however, have come under close scrutiny in recent years. In a thought provoking article, DeMiguel, Garlappi and Uppal (2009) demonstrated through extensive empirical studies that neither the sample moment based efficient portfolios nor their many sophisticated extensions perform well in a realistic setting where there are more than several assets in the market. Moreover, all of these mean-variance analysis inspired portfolio rules fail to outperform on a consistent basis the naïve diversification which simply assigns an equal weight to each of the assets. Similar observations were also made by Michaud (1998), Benartzi and Thaler (2001), Behr, Güttler and Miebs (2008), and Duchin and Levy (2009) among many others.

The lackluster performance of these estimated mean-variance efficient portfolio rules can be attributed to the estimation error associated with the estimated moments from historical data. The impact of estimation error on the estimated efficient portfolio is well documented even in the classical case when there are only several assets. See, e.g., Jobson and Korkie (1980). Such a problem oftentimes can be alleviated with a larger estimation window. Classical large sample theory suggests that when the number of assets is small and the historical data are abundant, the mean and variance of the returns can be consistently estimated. This property can be readily translated into the nearness between the theoretical mean-variance efficient portfolio and its estimates by means of the usual delta method, and in turn, the approximate optimality of the estimated mean-variance efficient portfolio rules. The effect of estimation error, however, quickly grows out of control when the number of assets increases; and an unrealistic amount of historical data are needed in order for the aforementioned large sample theory to be relevant. For example, as pointed out by DeMiguel, Garlappi and Uppal (2009), for the estimated mean-variance efficient portfolios to outperform the naïve diversification, i.e, to have a higher Sharpe ratio, we will need around 3000 months worth of historical data even for an investment universe of 25 assets. In other words, when the market is not small one cannot realistically expect to have enough historical data to do better than naïve diversification.

These findings inevitably cast doubts on the practical value of the mean-variance efficient portfolios. Mindful of such challenges, many researchers have opted for other portfolios that are traditionally perceived as suboptimal in the pursuit of a higher Sharpe ratio. Examples include minimum variance portfolio (Jorion, 1985; 1986), short-sell constrained minimum variance portfolios (Jagannathan and Ma, 2003), the equally weighted diversification (DeMiguel, Garlappi and Uppal, 2009), and other modifications to the mean-variance portfolio rule (see, e.g., Pesaran and Zaffaroni, 2010; and Antoine, 2012). The empirical successes of these supposedly suboptimal portfolios undoubtedly further fuel the debate about the meaning of optimality of the mean-variance analysis; or as Michaud (1998) put it – "*Is 'optimized' optimal?*" We shall argue in this article that the answer is affirmative, and the "optimized" can be optimal after all. The disappointing performance of the estimated efficient portfolios reported earlier merely reflects our inability to accurately estimate the moments of asset returns in a large market, which does not necessarily prevent us from constructing a feasible yet efficient portfolio rule.

In particular, we show that the optimal Sharpe ratio could be achieved through the so-called subspace mean-variance analysis. Instead of seeking the optimal allocation of wealth across the whole universe, the subspace mean-variance analysis restricts investment in a set of portfolios including, for example, portfolios corresponding to the leading eigenvectors of the covariance matrix of asset returns, and determines the optimal investment with such restrictions in the usual Markowitz fashion. The resulting subspace mean-variance efficient portfolios can then be estimated in the same fashion as the global mean-variance efficient portfolios. We argue that the loss of efficiency in restricting the investment options, referred to as systematic error hereafter, can be compensated through reduced estimation error. We demonstrate on both real and simulated data sets, that the two sources of errors can be balanced to yield performance consistently superior to both naïve diversification and estimated global mean-variance portfolio.

More specifically, we study the performance of the estimated subspace mean-variance portfolio when restricting investment in the leading principle components of the sample covariance matrix of historical returns. To further explain the empirical successes, we investigate the econometric properties of the sample subspace mean-variance efficient portfolios. Our analysis makes use of the notion that asset returns are driven by systematic risks represented by marketwide factors, a widely held belief that is corroborated by numerous studies (see, e.g., Connor, Goldberg and Korajczyk, 2010). Mainstream asset pricing models such as the capital asset pricing model (CAPM, for short; Sharpe, 1964), the intertemporal CAPM (Merton, 1973), and the arbitrage pricing theory (APT, for short; Ross, 1976) are all based on such principles. A fairly general framework to characterize these factor structures is the approximate factor model of Chamberlain and Rothchild (1983). Within this framework, we show that in a large market, the estimation error can be well-controlled using the proposed strategy. More specifically, we show that estimated subspace mean-variance portfolio is asymptotically efficient in that its Sharpe ratio approaches that of the global mean-variance portfolio as the estimation window and the size of the market increase.

The rest of this paper is organized as follows. Section 2 introduces the concept of subspace mean-variance analysis and explains how the resulting portfolios can be estimated in practice. In Section 3, we study the econometric properties of the estimated subspace mean-variance efficient portfolios and establish their asymptotic efficiency in a large market. These theoretical analyses are further supported by numerical experiments presented in Section 4. We close with some concluding remarks and discussions in Section 5.

2 Subspace Mean-Variance Analysis

Let $\mathbf{r} \in \mathbb{R}^N$ be the return of N risky assets in excess of a risk-free rate. Then the Markowitz mean-variance efficient portfolio is given as the solution to

$$\min_{\mathbf{w}\in\mathbb{R}^N} \left\{ \frac{\gamma}{2}\mathbf{w}^\mathsf{T}\Sigma\mathbf{w} - \mathbf{w}^\mathsf{T}E \right\}, \tag{1}$$

where $E \in \mathbb{R}^N$ and $\Sigma \in \mathbb{R}^{N\times N}$ are the mean and variance of $\mathbf{r}$ respectively, and γ is the coefficient of relative risk aversion. The solution of (1), denoted by,

$$\mathbf{w}_{\text{mv}} = \frac{1}{\gamma}\Sigma^{-1}E, \tag{2}$$

gives the optimal proportion of wealth that should be invested among the risky assets. It is well-known that the mean-variance efficient portfolio $\mathbf{w}_{\text{mv}}$ achieves the highest possible Sharpe ratio. Recall that the Sharpe ratio of a portfolio allocation $\mathbf{w} \in \mathbb{R}^N$ is given by

$$s(\mathbf{w}) = \frac{\mathbf{w}^\mathsf{T} E}{(\mathbf{w}^\mathsf{T}\Sigma\mathbf{w})^{1/2}},$$

so that for any $\mathbf{w}$,

$$s(\mathbf{w}) \le s(\mathbf{w}_{\text{mv}}) = (E^\mathsf{T}\Sigma^{-1}E)^{1/2}. \tag{3}$$

Obviously investors cannot hold the mean-variance portfolio in practice because neither E nor Σ is known in advance. Therefore, $\mathbf{w}_{\text{mv}}$ can only serve as a gold standard, and the goal is instead to construct feasible portfolio rules that can reproduce its level of performance. The most common strategy towards this goal is to first estimate E and Σ from historical return data, typically by their sample counterparts; and then plug the estimates in (2) to give an estimated mean-variance portfolio. More specifically, let $\mathbf{r}_t$, $t = 1, 2, \ldots, T$ be the excess returns of month t. Then the estimated mean-variance portfolio is:

$$\widehat{\mathbf{w}}_{\text{mv}} = \frac{1}{\gamma}\widehat{\Sigma}^{-1}\widehat{E},$$

where $\widehat{E}$ and $\widehat{\Sigma}$ are the sample moments:

$$\widehat{E} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_t, \qquad \text{and} \qquad \widehat{\Sigma} = \frac{1}{T-1}\sum_{t=1}^{T}(\mathbf{r}_t - \bar{\mathbf{r}}_t)(\mathbf{r}_t - \bar{\mathbf{r}}_t)^\mathsf{T}.$$

In the case when $\widehat{\Sigma}$ is singular, for example when $N \ge T$, its Moore-Penrose inverse can be used in place of $\widehat{\Sigma}^{-1}$ in defining $\widehat{\mathbf{w}}_{\text{mv}}$.

The sample mean-variance portfolio $\widehat{\mathbf{w}}_{\text{mv}}$ works fairly well when the number of assets is small whereas the estimation window is large. By law of large numbers, both $\widehat{E}$ and $\widehat{\Sigma}$ are consistent in that they converge to the respective true moments when the estimation window is large enough. The consistency of the moment estimators can be translated to the resulting estimated portfolio allocation $\widehat{\mathbf{w}}_{\text{mv}}$ using the usual delta method, leading to

$$s(\widehat{\mathbf{w}}_{\text{mv}}) \to_p s(\mathbf{w}_{\text{mv}}), \tag{4}$$

as the estimation window enlarges. In other words, the estimated portfolio is asymptotically efficient. The practical importance of (4) is also clear as it suggests that, even though the mean-variance portfolio itself typically is not practically available, there are feasible portfolio rules capable of reaching similar level of performance for large enough estimation window.

The plug-in tactics for constructing portfolio rules, however, quickly falter as the size of market grows as many have observed empirically. The suboptimal performance of estimated mean-variance portfolios can be attributed to the estimation error associated with the estimated moments (e.g., Merton, 1980). Although a classical and straightforward problem when the number of assets is small, estimating the mean and covariance matrix in a large market with relatively limited amount of data is notoriously difficult. As the number of parameters involved in the mean and covariances of the asset returns increases with the size of the market, an unrealistic amount of historical data are often needed to yield any meaningful moment estimator. As a result, the sample mean-variance portfolio may perform rather poorly in a large investment universe. More specifically, the following theorem shows that when the number of assets N is much larger than the estimation window T, the sample mean-variance portfolio may become useless in that its Sharpe ratio can be arbitrarily small.

Theorem 1. *Assume that the excess returns $\mathbf{r}_t$, $t = 1, 2, \ldots, T$ are independently and normally distributed with mean E and covariance Σ. Then the Sharpe ratio of the sample mean-variance portfolio satisfies*

$$\frac{s(\widehat{\mathbf{w}}_{\rm mv})}{s(\mathbf{w}_{\rm mv})} = O_p(\sqrt{T/N}).$$

In particular, with a fixed estimation window, the relative efficiency of the sample mean-variance portfolio converges to zero in probability as the number of assets increases.

To overcome this challenge, we propose here to trade the optimality of the mean-variance portfolio with estimability, an idea reminiscent of the all-important bias-variance tradeoff in statistics.

More specifically, instead of searching through all possible investment options, we consider restricting investment in a carefully chosen linear subspace. From a mathematical point of view, this means that we restrict the asset allocation vector $\mathbf{w}$ to be in a linear subspace $\mathcal{P}$ of $\mathbb{R}^N$. Efficient wealth allocation within $\mathcal{P}$ can be determined by Markowitz style analysis, i.e., by solving the following optimization problem similar to (1):

$$\min_{\mathbf{w}\in\mathcal{P}}\left\{\frac{\gamma}{2}\mathbf{w}^\intercal\Sigma\mathbf{w}-\mathbf{w}^\intercal E\right\}. \tag{5}$$

Similar to the usual mean-variance analysis, the solution of (5) can be expressed explicitly.

Proposition 1. *Let $\mathcal{P}$ be a d dimensional linear subspace of $\mathbb{R}^N$. Then the solution of (5) is given by*

$$\mathbf{w}_{\text{mv}}^{\mathcal{P}} := \frac{1}{\gamma}P_{\mathcal{P}}\left(P_{\mathcal{P}}^\intercal\Sigma P_{\mathcal{P}}\right)^{-1}P_{\mathcal{P}}^\intercal E, \tag{6}$$

where $P_{\mathcal{P}}$ is a $N\times d$ matrix whose columns are an orthonormal basis of $\mathcal{P}$.

We note that the expression (6) is invariant to the choice of $P_{\mathcal{P}}$, or equivalently the basis of $\mathcal{P}$. We shall in what follows refer to $\mathbf{w}_{\text{mv}}^{\mathcal{P}}$ as the subspace mean-variance portfolio. Same as the global mean-variance portfolio $\mathbf{w}_{\text{mv}}$, the subspace mean-variance portfolio cannot be implemented in practice because it still depends on the unknown mean and covariances of asset returns. We shall consider instead the estimated subspace mean-variance portfolio:

$$\widehat{\mathbf{w}}_{\text{mv}}^{\mathcal{P}} = \frac{1}{\gamma}P_{\mathcal{P}}\left(P_{\mathcal{P}}^\intercal\widehat{\Sigma}P_{\mathcal{P}}\right)^{-1}P_{\mathcal{P}}^\intercal\widehat{E}.$$

Recall that $\widehat{E}$ and $\widehat{\Sigma}$ are the sample mean and covariance matrix respectively.

Both naïve diversification and sample mean-variance portfolio can be viewed as special cases of the estimated subspace mean-variance portfolio with different choices of $\mathcal{P}$. In particular, naïve diversification corresponds to the choice of a one dimensional linear subspace $\mathcal{P}=\{a\mathbf{1}:a\in\mathbb{R}\}$ whereas the sample mean-variance portfolio can be identified with $\mathcal{P}=\mathbb{R}^N$. What is of interest here are, however, linear subspaces between these two extreme choices. Let $\widehat{\mathbf{w}}_{\text{mv}}^{\mathcal{P}}$ be the estimated subspace mean-variance within linear subspace $\mathcal{P}$. Its "suboptimality" can be measured by

$$s(\mathbf{w}_{\text{mv}})-s(\widehat{\mathbf{w}}_{\text{mv}}^{\mathcal{P}})=\left[s(\mathbf{w}_{\text{mv}}^{\mathcal{P}})-s(\widehat{\mathbf{w}}_{\text{mv}}^{\mathcal{P}})\right]+\left[s(\mathbf{w}_{\text{mv}})-s(\mathbf{w}_{\text{mv}}^{\mathcal{P}})\right]. \tag{7}$$

We shall refer to the two terms on the right hand side of (7) as estimation error and systematic error respectively because the first term measures the loss of optimality due to the estimated moments whereas the second term corresponds to the deficiency caused by restricting ourselves to a smaller investment space. Typically, as the dimension $\dim(\mathcal{P})$ increases, estimation error increases since there are more parameters to be estimated; whereas the systematic error decreases since there are less missed investment opportunities. As a result, improved performance could be achieved by balancing the tradeoff between the two sources of errors. Such a tradeoff can be illustrated by Figure 1.

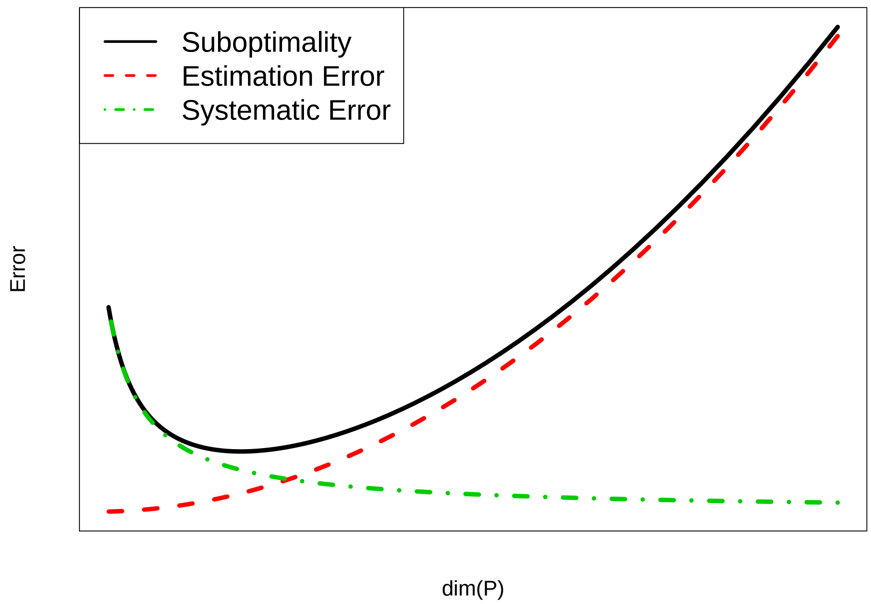

Figure 1: Illustration of tradeoff between estimation error and systematic error in subspace mean-variance analysis.

The choice of subspace $\mathcal{P}$ clearly plays an important role in the proposed strategy. In principle, we can choose any of our favorite portfolio rules and form $\mathcal{P}$ as the linear subspace spanned by these rules. For example, one can choose $\mathcal{P}$ to be the linear subspace spanned by the leading eigenvectors of the second moment $\mathbb{E}(\mathbf{r}\mathbf{r}^\top)$ as suggested by Carrasco and Noumon (2013), or the Fama-French factors, leading to a strategy similar to those adopted by Fan, Fan and Lv (2008) among others. The more specific choice we made here is motivated by our understanding of the approximate factor model, a fairly general framework commonly used to describe the belief that the individual returns are driven by marketwide factors. As our econometric analysis from Section 3 reveals, when it comes to portfolio selection in a large market, there is little loss in restricting investment in the eigenportfolios, i.e., portfolios corresponding to the leading eigenvectors of the covariance matrix of the asset returns. In other words, we can take $\mathcal{P}$ to be the linear space spanned by the first d eigenvectors of Σ.

From Proposition 1, by taking $P_{\mathcal{P}} = [\eta_1, \ldots, \eta_d]$, the subspace mean-variance portfolio is

$$\frac{1}{\gamma}\sum_{k=1}^{d} \theta_k^{-1} \eta_k \eta_k^{\mathsf{T}} E,$$

where $\theta_1 \geq \theta_2 \geq \ldots$ are the eigenvalues of Σ, and η_ks are their respective eigenvectors. Of course, $\mathcal{P}$ is not known in advance and can be naturally estimated by the linear space spanned by the first d eigenvectors of the sample covariance matrix $\widehat{\Sigma}$, denoted by $\widehat{\mathcal{P}}$. The corresponding estimated subspace mean variance portfolio is then given by

$$\widehat{\mathbf{w}}_d = \frac{1}{\gamma}\sum_{k=1}^{d} \widehat{\theta}_k^{-1} \widehat{\eta}_k \widehat{\eta}_k^{\mathsf{T}} \widehat{E},$$

where $\widehat{\theta}_1 \geq \widehat{\theta}_2 \geq \ldots$ are the eigenvalues of $\widehat{\Sigma}$, and $\widehat{\eta}_k$s are their corresponding eigenvectors. We shall now show that, in the context of the approximate factor model, $\widehat{\mathbf{w}}_d$ can achieve a Sharpe ratio similar to that of the global mean-variance portfolio whenever d is appropriately chosen.

3 Econometric Properties

In this section, we study the econometric properties of the estimated subspace mean-variance portfolio rules $\widehat{\mathbf{w}}_d$. In particular, we show here that it can achieve asymptotically the same level of performance as the global mean-variance portfolio in a large market. Following the most commonly used asset pricing models such as the CAPM and APT among others, we assume that the systematic risks are represented by a small number of marketwide factors and expected returns on individual securities are linear functions of their standardized covariances, or betas, with these factors.

Recall that $\mathbf{r}_t = (r_{1t}, \ldots, r_{Nt})^{\mathsf{T}}$ is the excess return of time t. It is assumed to follow the following approximate factor model:

$$r_{jt} = E_j + \beta_{j1} f_{1t} + \ldots + \beta_{jK} f_{Kt} + \varepsilon_{jt}, \qquad j = 1, \ldots, N; \quad t = 1, \ldots, T, \tag{8}$$

where $\mathbf{f}_t = (f_{1t}, \ldots, f_{Kt})^{\mathsf{T}}$ is a vector of common factors, $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jK})^{\mathsf{T}}$ is a vector of factor loadings associated with the jth asset, and ε_{jt} is the idiosyncratic component of r_{jt} satisfying $\mathbb{E}(\varepsilon_{jt}|\mathbf{f}_t) = \mathbf{0}$. The factors may be unobserved with mean zero and variance Σ_f.

Without loss of generality, we assume that Σ_f is strictly positive definite so that there is no redundancy among the factors. The idiosyncratic risks have a variance Σ_ε. As a result, the covariance matrix of $\mathbf{r}_t$ is $\Sigma = B\Sigma_f B^\mathsf{T} + \Sigma_\varepsilon$ where B is a $N \times K$ matrix whose jth row is $\boldsymbol{\beta}_j^\mathsf{T}$.

Let

$$E_j = \alpha_j + \beta_{j1}\mu_1^f + \ldots + \beta_{jK}\mu_K^f, \qquad j = 1, 2, \ldots, N, \tag{9}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)^\mathsf{T}$ is the so-called Jensen's alpha (Jensen, 1968), and μ_k^f is the risk premium of the kth factor, that is, the excess return of an asset whose beta for the kth factor is one and zero otherwise. Without loss of generality, we shall assume that the risk premium $\mu_k^f > 0$. Exact arbitrage pricing dictates that the pricing error $\boldsymbol{\alpha} = \mathbf{0}$. Assuming that there is no cross-sectional dependence, i.e., Σ_ε is diagonal, Huberman (1982) showed that no-arbitrage implies that $\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{\alpha}$ is bounded. Similar results were also established by Chamberlain and Rothschild (1983) who relaxed the assumption of uncorrelated idiosyncratic noise. Such an asymptotic APT model has been studied extensively in the literature and there are ample evidences that they provide an adequate portrait of the real-world market (see, e.g., Zhang, 2009). Although our method of portfolio construction does not rely on the validity of the asymptotic APT, for brevity, in what follows we shall assume it holds nonetheless because of its popularity. In addition, we shall also consider the following assumptions for the approximate factor model (8).

Assumption A (Factors) The factors have finite fourth moments such that there exists a positive constant $C_1 < \infty$ satisfying

$$\max_{1\le k\le K} \mathbb{E} f_{kt}^4 \le C_1.$$

Assumption B (Factors Loadings) There is a strictly positive definite matrix Σ_B such that

$$B^\mathsf{T} B/N \to \Sigma_B \qquad \text{as } N \to \infty.$$

Assumption C (Idiosyncratic Risks) The covariance matrix of $\boldsymbol{\epsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{Nt})^\mathsf{T}$, Σ_ε has eigenvalues bounded away from both zero and infinity. Moreover, the idiosyncratic risks have finite fourth moments such that there exists a positive constant $C_2 < \infty$ satisfying

$$\max_{1 \le j \le N} \mathbb{E}\varepsilon_{jt}^4 \le C_2.$$

These assumptions are fairly general and met by multi-factor asset pricing models appeared in the literature (see, e.g., Bai and Ng, 2002). We emphasize that we do not require the normality of either the factors or idiosyncratic risks. In addition, cross-sectional dependence and heterogeneity are allowed for the idiosyncratic risks.

Within the framework of the approximate factor model, it is convenient to explain how the proposed portfolio rule $\widehat{\mathbf{w}}_d$ works. It is instructive to consider the special case when there is no pricing error, i.e., $\boldsymbol{\alpha} = \mathbf{0}$; and the idiosyncratic noise ε_{jt}s are uncorrelated and with a common variance σ_ε^2, i.e., $\Sigma_\varepsilon = \sigma_\varepsilon^2 I$. By Sherman-Morrison-Woodbury identity, the global mean-variance portfolio can be given by

$$\begin{aligned}
\mathbf{w}_{\mathrm{mv}} &= \frac{1}{\gamma}\Sigma^{-1}E \\
&= \frac{1}{\gamma}\left(B\Sigma_f B^\mathsf{T} + \sigma_\varepsilon^2 I\right)^{-1} B\boldsymbol{\mu}_f \\
&= \frac{1}{\gamma}\left(\frac{1}{\sigma_\varepsilon^2} I - \frac{1}{\sigma_\varepsilon^2} B\left(\sigma_\varepsilon^2\Sigma_f^{-1} + B^\mathsf{T}B\right)^{-1} B^\mathsf{T}\right) B\boldsymbol{\mu}_f \\
&= \frac{1}{\gamma} B\left[\frac{1}{\sigma_\varepsilon^2}\boldsymbol{\mu}_f - \frac{1}{\sigma_\varepsilon^2}\left(\sigma_\varepsilon^2\Sigma_f^{-1} + B^\mathsf{T}B\right)^{-1} B^\mathsf{T}B\boldsymbol{\mu}_f\right]
\end{aligned}$$

Clearly, $\mathbf{w}_{\mathrm{mv}}$ belongs to the linear subspace spanned by the column vectors of B, which in this case also coincides with the linear space spanned by the first K eigenvectors of Σ. In other words, there is no loss of efficiency in restricting investment in the eigenportfolios. For the more general approximate factor model, such a relationship no longer holds, but as we shall see, the efficiency of $\widehat{\mathbf{w}}_d$ remains.

Note that $\widehat{\mathbf{w}}_d$ relies on the historical data which we shall allow for temporal dependence for the factors and idiosyncratic risks. As a result, the factors could be either static or dynamic (see, e.g., Forni, Hallin, Lippi and Reichlin, 2000; and Forni and Lippi, 2001). More specifically, we shall assume that the factors satisfy the following condition.

Assumption D The factors $\mathbf{f}_1, \ldots, \mathbf{f}_T$ are weakly dependent in that there exists a positive constant $C_3 < \infty$ satisfying

$$\max_{1 \le k_1, k_2 \le K} \mathbb{E}\left\{ \sum_{t=1}^{T} \left(f_{k_1 t} f_{k_2 t} - \Sigma^f_{k_1 k_2} \right)^2 \right\} \le C_3 T$$

where $\Sigma^f_{k_1 k_2}$ is the (k_1, k_2)th entry of Σ_f.

Assumption E The idiosyncratic risks $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_T$ are weakly dependent in that there exists a positive constant $C_4 < \infty$ satisfying

$$\max_{1 \le j_1, j_2 \le N} \mathbb{E}\left\{ \sum_{t=1}^{T} \left(\varepsilon_{j_1 t} \varepsilon_{j_2 t} - \Sigma^{\varepsilon}_{j_1 j_2} \right)^2 \right\} \le C_4 T$$

where $\Sigma^{\varepsilon}_{j_1 j_2}$ is the (j_1, j_2)th entry of Σ_{ε}.

Assumption F The factors and idiosyncratic risks are jointly weakly dependent in that there exists a positive constant $C_5 < \infty$ satisfying

$$\max_{\substack{1 \le j \le N \\ 1 \le k \le K}} \mathbb{E}\left(\sum_{t=1}^{T} \varepsilon_{jt}^2 f_{kt}^2 \right) \le C_5 T.$$

We are now in position to state our main result whose proof is relegated to the appendix.

Theorem 2. *Let $d = K$ be fixed and finity. Then under asymptotic APT model satisfying Assumptions A-F,*

$$s^2(\widehat{\mathbf{w}}_d) = s^2(\mathbf{w}_{\rm mv}) + O_p(T^{-1/2} + N^{-1/2}).$$

As suggested by Theorem 2, when the market is large and the estimation window is long, the Sharpe ratio of $\widehat{\mathbf{w}}_d$ will approximately equal to that of $\mathbf{w}_{\rm mv}$, so long as $d = K$. It is worth noting that the principal components generally differ from the common factors in a linear factor model. But Theorem 2 shows that, as far as portfolio selection is concerned, the principal components may serve the same purpose as the common factors in a large market.

The number of factors K, of course, is typically unknown in practice. But they can be consistently estimated from the historical data as well. In particular, Bai and Ng (2002) develop a class of criteria for estimating K when both N and T are large. One possible choice, as they suggested, is to estimate K by

$$\widehat{d} = \underset{1\le k\le k_{\max}}{\arg\min}\left\{\log\left(\sum_{j>k}\widehat{\theta}_k^2\right) + \frac{k(N+T)}{NT}\log\left(\frac{NT}{N+T}\right)\right\}, \tag{10}$$

where $k_{\max}$ is a prespecified maximum possible number of factors, and $\widehat{\theta}$s are the eigenvalues of $\widehat{\Sigma}$. Following Bai and Ng (2002), we shall set $k_{\max} = 8$ in the numerical experiments. Under some regularity conditions, Bai and Ng (2002) showed that when both N and T are large, $\mathbb{P}\{\widehat{d} = K\} \to 1$. As a corollary to Theorem 2, we have

Corollary 3. *For any $\widehat{d}$ such that $\mathbb{P}\{\widehat{d} = K\} \to 1$ as $N, T \to \infty$, under the asymptotic APT model satisfying Assumptions A-F,*

$$s^2(\widehat{\mathbf{w}}_{\widehat{d}}) = s^2(\mathbf{w}_{\rm mv}) + O_p(T^{-1/2} + N^{-1/2}).$$

Typically K is relatively small, and the numerical illustrations in the next section suggest that fairly good and robust performance can be achieved generally for a wide range of choices of d.

4 Experimental Studies

In this section, we evaluate the performance of the portfolio selection rules introduced in the previous section on both simulated and real data sets.

4.1 Simulation Results

For illustration purposes, we begin with a simulation study. Our simulation setup is similar to MacKinlay and Pastor (2000), DeMiguel, Garlappi, and Uppal (2009), and Tu and Zhou (2011) among others. More specifically, we simulated the monthly returns of N risky assets from a three-factor model. To investigate the effects of market size, we considered $N = 25$ or 100. Following Tu and Zhou (2011), we simulated the three factors from a multivariate normal distribution with mean and covariances calibrated from July 1963 to August 2007 monthly data on the market portfolio, the Fama-Frenchs size and book-to-market portfolios respectively (available at `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/`

data_library.html). The factor loadings of the risky assets were randomly sampled from a uniform distribution between 0.9 and 1.2 for the market βs, -0.3 and 1.4 for the size portfolio βs, -0.5 and 0.9 for the book-to-market portfolio βs. In addition, the residual variance-covariance matrix Σ_ε is taken to be diagonal, with the diagonal elements sampled from a uniform distribution between 0.10 and 0.30 to yield an average cross-sectional volatility of 20%. To appreciate how the length of estimation window may affect the performance of portfolio rules, we considered $T = 60$, 120, or 240 months.

First we examine the effect of the $d = \dim(\mathcal{P})$ on the corresponding estimated subspace mean-variance portfolio. To this end, we computed the proposed portfolio rules $\widehat{\mathbf{w}}_d$ with $d = 1, 2, \ldots$ for each simulated datasets. The Sharpe ratio of $\widehat{\mathbf{w}}_d$ was then calculated. We report in Figure 2 the results averaged over 1000 runs.

The effect of the systematic error and estimation error is clear from this exercise. When $d = 1$ or 2, the leading eigenportfolios fail to capture all market-wide factors and as a result, incur non-negligible systematic error. On the other hand when $d > 3$, the systematic error is fairly small whereas the estimation error becomes increasingly prominent as d increases. The ideal balance between systematic error and estimation error is achieved when $d = 3$. These observations confirm the econometric analysis presented earlier.

For contrast, in each plot of Figure 2, the Sharpe ratio of the naïve diversification is also reported, as represented by the gray horizontal lines. It is worth noticing that although the optimal tradeoff between the two sources of error is achieved when $d = 3$, for a wide range of choices for d, the proposed subspace mean-variance portfolio can still produce significant improvement over the naïve diversification.

To further illustrate the practical merits of the proposed methodology, we now consider the performance of $\widehat{w}_d$ with d estimated using $\widehat{d}$ defined in (10). For each simulation run, $\widehat{d}$ may take a different value. But typically $\widehat{d} = 3$ which occured 953 out of the 1000 runs. Figure 3 reports the comparison between the estimated subspace mean-variance portfolio, sample mean-variance portfolio and the naïve diversification along with the true yet in practice infeasible mean-variance efficient portfolio. It is clear that the estimated subspace mean-variance portfolio consistently outperforms the naïve diversification.

Another desirable feature of the estimated subspace mean-variance portfolio is its stability. Estimated mean-variance portfolios based on plug-in principle often produce extremely

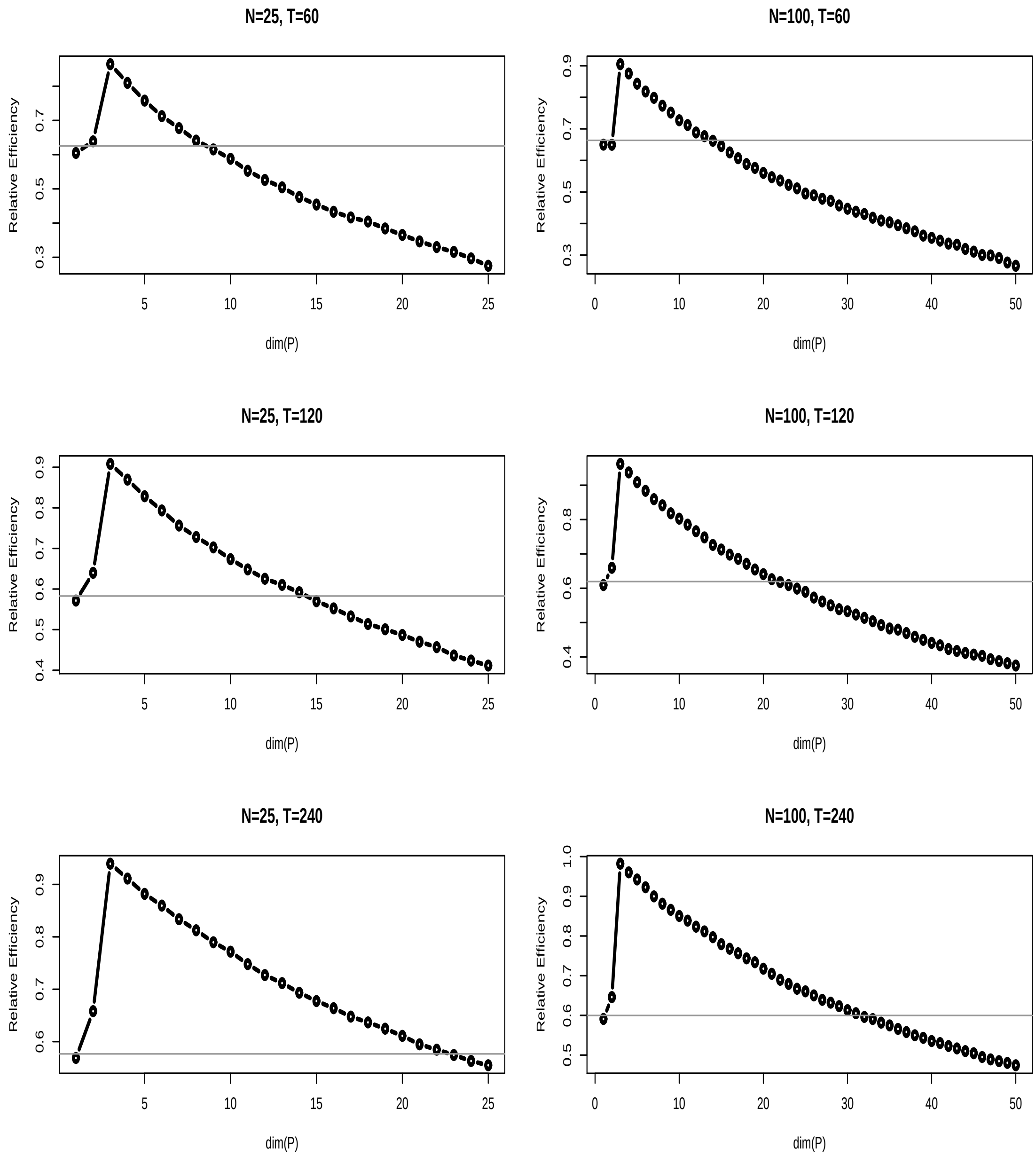

Figure 2: Simulation results from a three-factor model. For each combination of market size $N = 25$ or 100, and estimation window $T = 60$, 120 or 240 months, the relative efficiency, measured by the ratio between the Sharp ratio relative to that of the true (infeasible in practice) mean-variance portfolio, of $\widehat{\mathbf{w}}_d$ is reported here for different choices of d. The results are averaged over 1000 simulated datasets for each plot. The gray horizontal lines correspond the averaged relative efficiency for the naïve diversification.

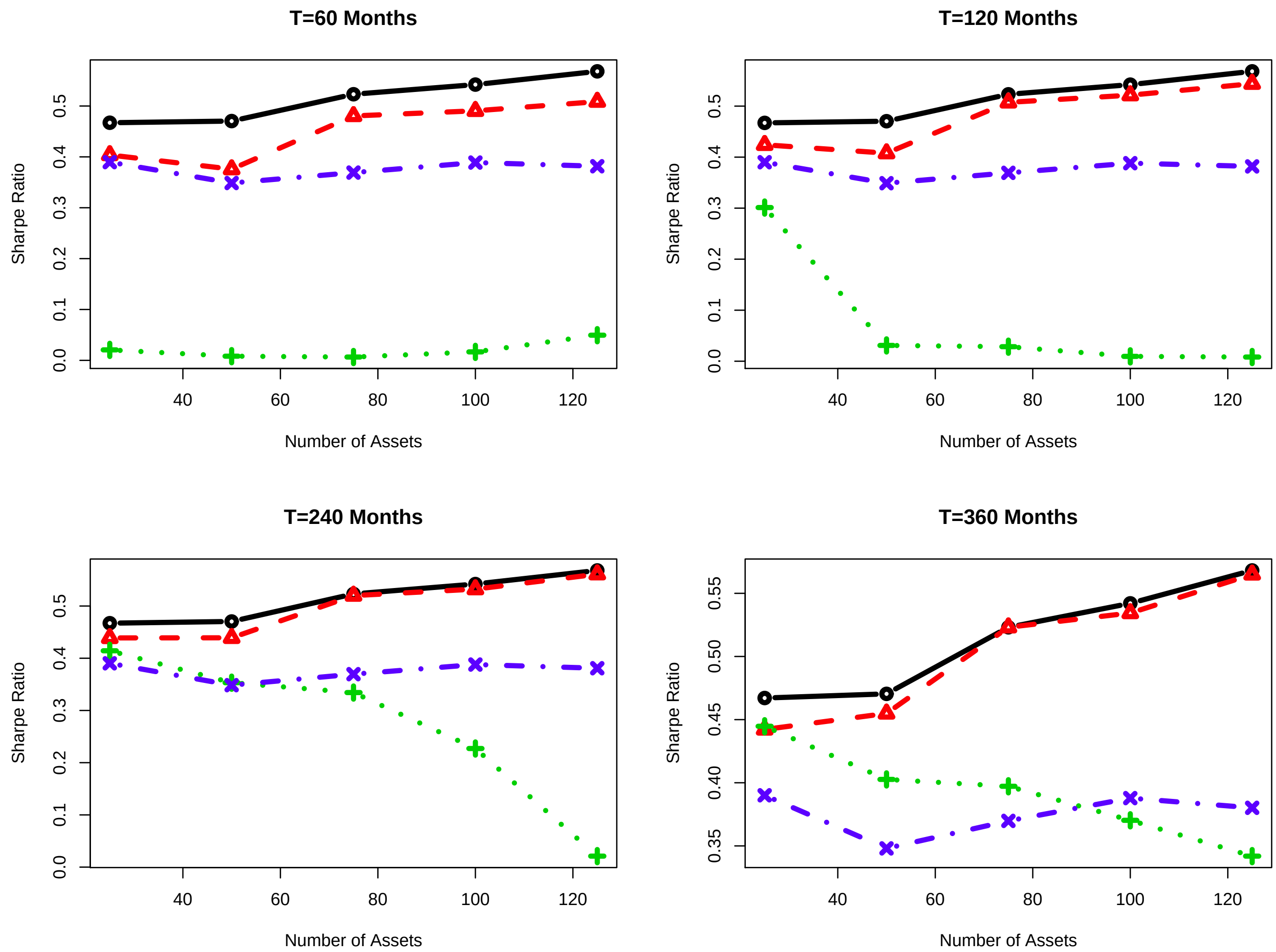

Figure 3: Simulation results based on three factor model: Data were simulated from a three factor model. Sharpe ratio of the true mean-variance portfolio (circles), estimated subspace mean-variance portfolio (triangles), sample mean-variance portfolio (pluses), and naïve portfolio rule (crosses) are presented for different estimating window size and market size. Note that the true mean-variance portfolio is not feasible and it is added for reference only.

large long and short positions that fluctuate substantially over time. As noted by Black and Litterman (1992), "when investors have tried to use quantitative models to help optimize the critical allocation decision, the unreasonable nature of the results has often thwarted their efforts." This problem can be alleviated when we restrict our investment in a low dimensional subspace. To demonstrate such advantage of the estimated subspace mean-variance portfolio, we look at a typical example with 100 assets and $T = 120$ months historical data. Figure 4 depicts the holdings of the risky asset in comparison with those of the sample mean-variance portfolio and true population mean-variance portfolio.

It is clear that holdings of the estimated subspace mean-variance portfolio is much more stable than those of the sample mean-variance portfolio and they track very well with the optimal holdings represented by the population mean-variance portfolio. To further assess the stability, we evaluated the amount of monthly rebalancing required by both portfolio rules. Let $\mathbf{w}_t$ and $\mathbf{w}_{t+1}$ be the portfolio weights with 100 assets constructed at Month t and $t+1$ respectively. Then rebalancing cost at Month $t+1$ can be naturally measured by

$$\text{Turnover}_t := \sum_{j=1}^{N} |\mathbf{w}_{t+1,j} - \mathbf{w}_{t,j}|.$$

Figure 5 provides the box plots of Turnover_t occurred in a 50 year period when the estimating window is 120 months and 240 months respectively, which again suggests that the subspace mean-variance portfolio is much more stable than usual sample mean-variance portfolio.

4.2 Empirical Illustrations

We now study the empirical performance of the proposed portfolio rules. The data set used in our analysis is the Fama-French 25 (5×5) and 100 (10×10) portfolios formed on size and book-to-market, each containing equal-weighted returns for the intersections of size portfolios and book-to-market portfolios (available at `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`). Following DeMiguel, Garlappi and Uppal (2007), we use a rolling-window approach to monthly returns collected over a fifty year period, from January 1961 to December 2010. More specifically, for any given estimation window T, we determine the risky asset holdings at time t using mean and covariances estimated using $\mathbf{r}_{t-T}, \ldots, \mathbf{r}_{t-1}$. We record the returns of a portfolio rule over the fifty year period and compute its Sharpe ratio by taking the ratio of the average of these recorded returns over their sample standard

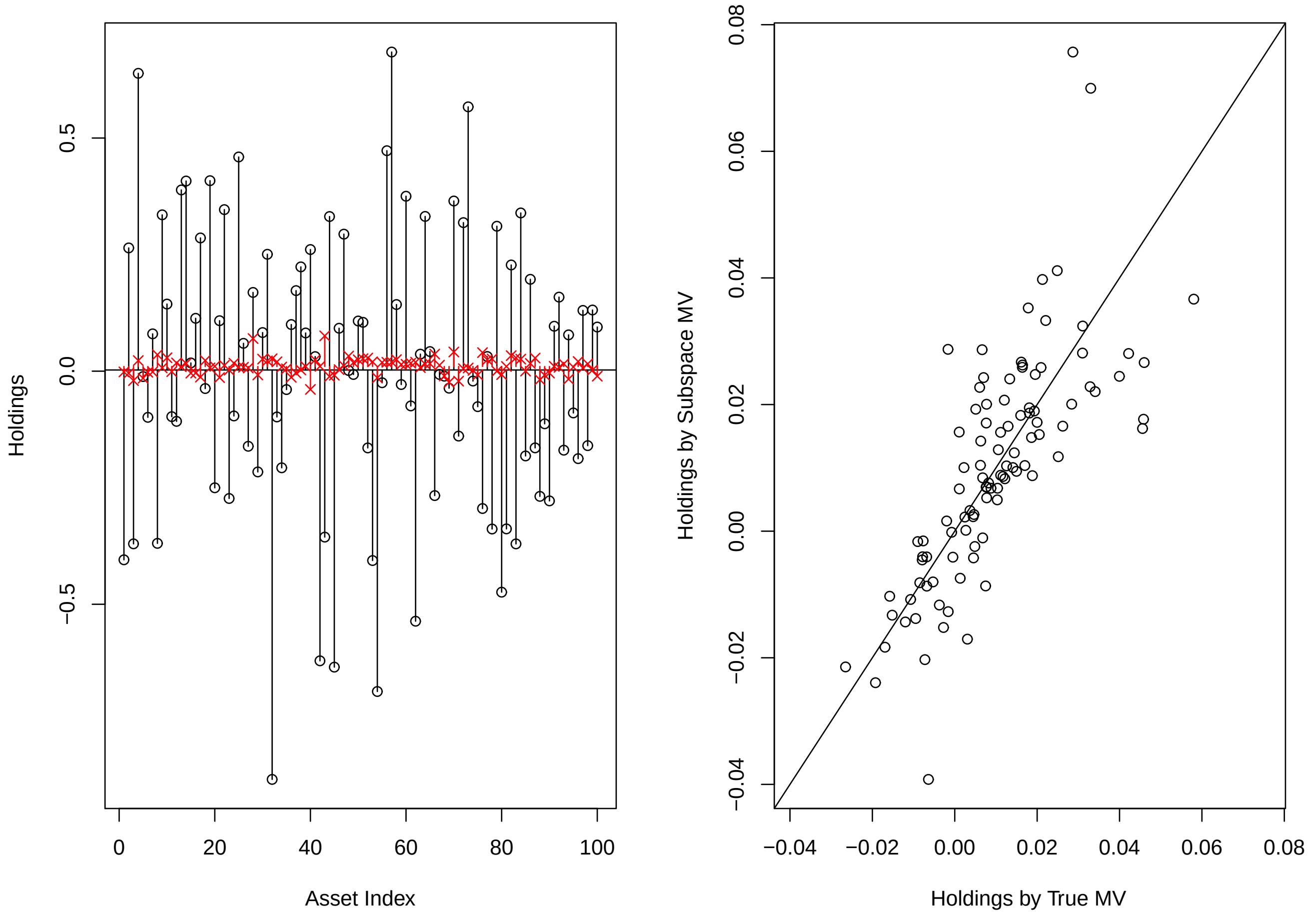

Figure 4: Stability – The left panel compares the holdings of the estimated mean-variance portfolio, represented by the circles, and subspace mean-variance portfolio, represented by the crosses. The right panel compares the holdings of the estimated subspace mean-variance portfolio with those of the true global mean-variance portfolio.

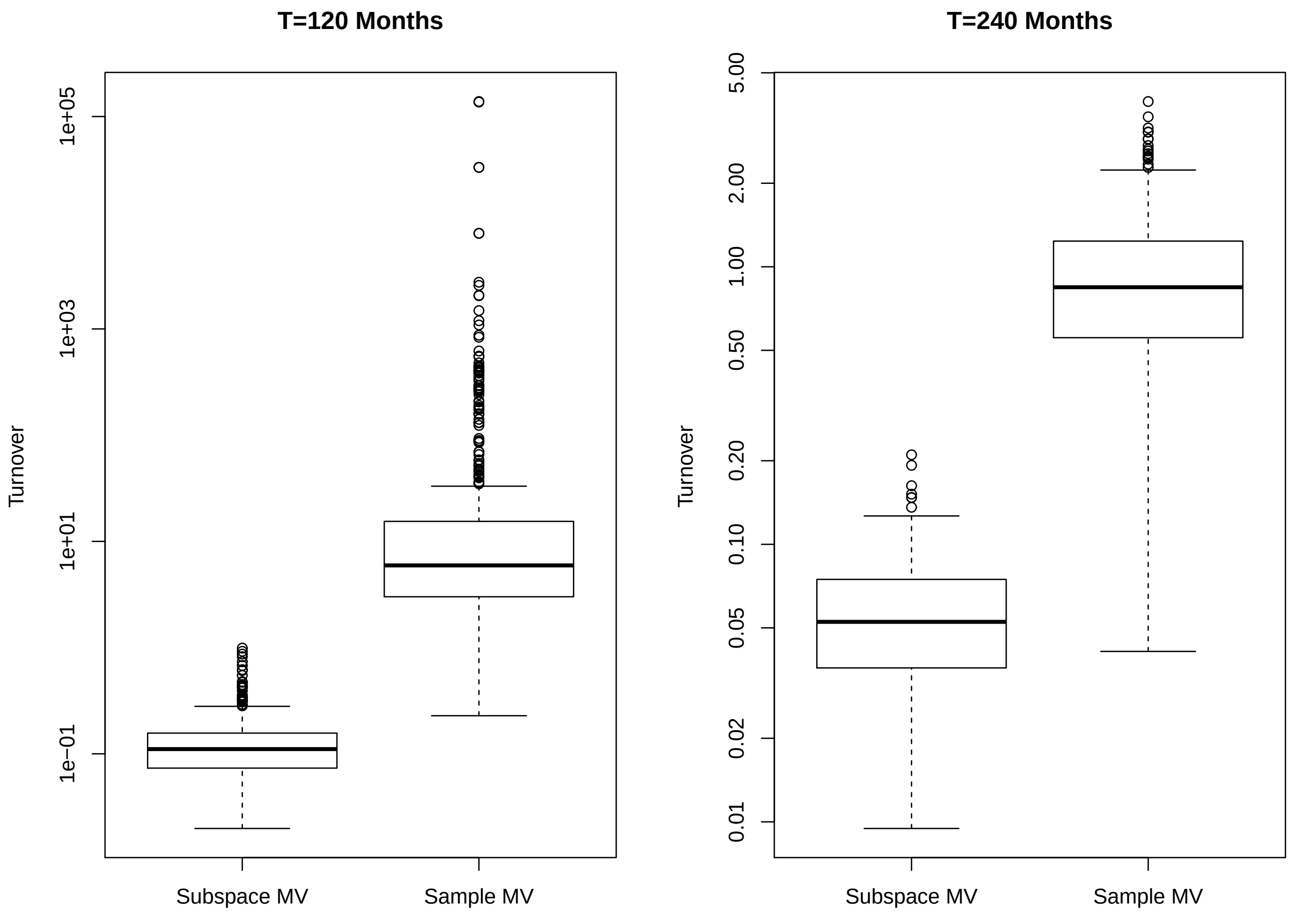

Figure 5: Rebalancing cost – Boxplots of the monthly rebalancing cost for the subspace mean-variance portfolio and sample mean-variance portfolio with 100 assets over a period of 50 years. The left panel corresponds to an estimating window of 120 months whereas the right panel 240 months.The Y-axes in both panels are in log scale for better contrast between the two portfolio rules.

deviation. We report in Figure 6 the Sharpe ratio computed in this fashion for different choices of estimation window and d for $N = 25$ and 100.

The horizontal lines in the figure correspond to the Sharpe ratio for the estimated subspace mean-variance with $\dim(\mathcal{P})$ determined by the information criterion given in (10). Not surprisingly, the pattern now becomes more profound than the simulation results presented earlier, which could be a result of the small-firm effect, calendar effects, momentum, and other anomalies often reported in the literature. At each time t, we calibrated d using the information criterion (10) as before. Typical choices of d are between five and seven, and $\widehat{d} = 5$ occurs nearly 50% of the time.

To demonstrate the merits of the subspace mean-variance portfolio, we also applied several other common portfolio rules to these data. Many existing rules have been recently examined and compared by DeMiguel, Garlappi and Uppal (2007), and Tu and Zhou (2011). In addition to the sample mean-variance portfolio and naïve diversification, we included in our comparison four of the better ones they identified: the three-fund rule of Jorion (1986; `PJ` for short), the rule from Kan and Zhou (2007; `KZ` for short), combination rule based on sample mean-variance portfolio and naïve diversification (Tu and Zhou, 2011, `S&N` for short), and combination rule based on `KZ` and naïve diversification (Tu and Zhou, 2011, `KZ&N` for short). The Sharpe ratio attained by these methods with estimation window $T = 60$ months, 120 months and 360 months are reported in Table 1.

Data	T	Sample	Naïve	S&N	KZ	KZ&N	PJ	Subspace MV
	60	0.21	0.13	0.21	0.21	0.21	0.22	0.25
25 Portfolios	120	0.32	0.14	0.33	0.33	0.33	0.33	0.30
	240	0.36	0.15	0.36	0.36	0.36	0.35	0.35
	60	0.13	0.13	NA	NA	NA	NA	0.16
100 Portfolios	120	0.10	0.14	0.14	0.10	0.12	0.08	0.23
	240	0.22	0.15	0.24	0.25	0.26	0.25	0.25

Table 1: Comparison between the estimated subspace mean-variance portfolio and several other popular alternatives on the Fama-French data sets. Reported here are the Sharpe ratio achieved over a period of 50 years. The portfolio rules S&N, KZ, KZ&N and PJ require $T > N$ and therefore are not applicable for $N = 100$ portfolios when $T = 60$.

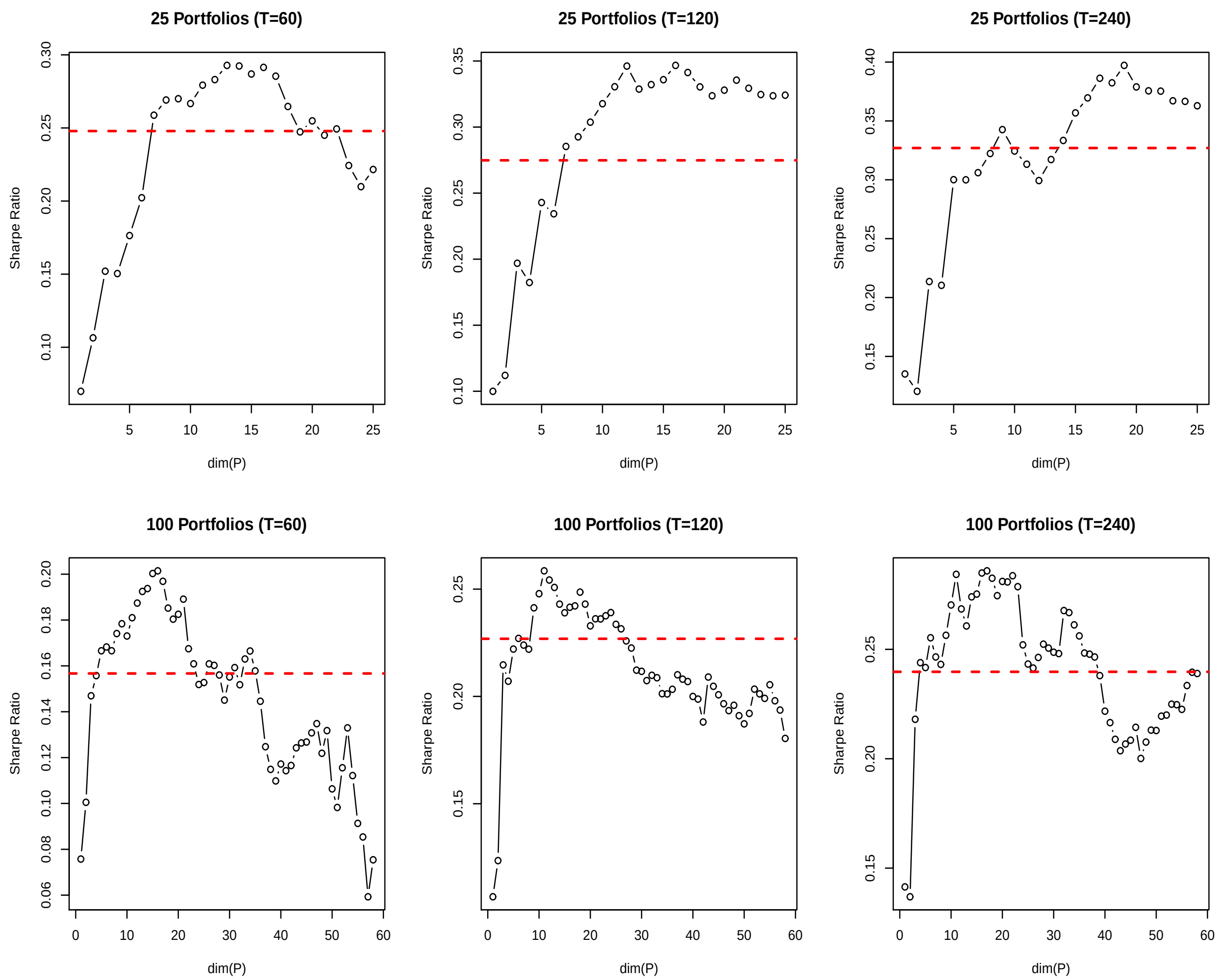

Figure 6: Fama-French portfolio examples – historical performance of $\widehat{\mathbf{w}}_d$ for different choices of d over a fifty year period. The dashed horizontal lines represent the Sharpe ratio of the estimated subspace mean-variance portfolio with $\dim(\mathcal{P})$ determined using the information criterion (10).

As shown by Table 1, the estimated subspace mean-variance portfolio compares favorably with the other methods. In most cases, it offers substantial improvement over sample mean-variance portfolio and naïve diversification. It is particularly attractive when the ratio N/T is relatively large.

5 Conclusions and Discussions

Recent empirical studies have ignited a series of debate regarding the practical merits of mean-variance analysis when it comes to a large investment universe. At the center of these discussions is the fundamental question: to what extent can the promised optimality of the mean-variance portfolio be replicated in practice? We argue that it is possible to achieve similar performance as the population mean-variance portfolio in practice although evidences from recent empirical studies seemingly point to another direction. In particular, we show that under the approximate factor model, optimal Sharpe ratio can be achieved by restricting the investment in a number of leading eigen-portfolios.

Appendix A – Proof of Theorem 1

It suffices to consider the case when $T/N \to 0$. Recall that

$$\widehat{\mathbf{w}}_{\mathrm{mv}} = \frac{\widehat{\Sigma}^{+}\widehat{E}}{\mathbf{1}^{\mathsf{T}}\widehat{\Sigma}^{+}\widehat{E}},$$

where $\widehat{\Sigma}^{+}$ is the Moore-Penrose inverse of $\widehat{\Sigma}$ to account for the fact that $\widehat{\Sigma}$ is not of full rank. Its Sharpe ratio is given by

$$s(\widehat{\mathbf{w}}_{\mathrm{mv}}) = \frac{E^{\mathsf{T}}\widehat{\Sigma}^{+}\widehat{E}}{\left(\widehat{E}^{\mathsf{T}}\widehat{\Sigma}^{+}\Sigma\widehat{\Sigma}^{+}\widehat{E}\right)^{1/2}}.$$

Let $Z_i = \Sigma^{-1/2}\mathbf{r}_i$, then $\widehat{\Sigma} = \Sigma^{1/2}S_Z\Sigma^{1/2}$ where S_Z is the sample covariance matrix of Z_is. As a result, $\widehat{\Sigma}^{+} = \Sigma^{-1/2}S_Z^{+}\Sigma^{-1/2}$, and moreover

$$s(\widehat{\mathbf{w}}_{\mathrm{mv}}) = \frac{E^{\mathsf{T}}\Sigma^{-1/2}S_Z^{+}\Sigma^{-1/2}\widehat{E}}{\left(\widehat{E}^{\mathsf{T}}\Sigma^{-1/2}S_Z^{+}S_Z^{+}\Sigma^{-1/2}\widehat{E}\right)^{1/2}}$$

Because $T/N \to 0$, with probability tending to one, S_Z has exactly rank $T-1$. See, e.g., Anderson (2003). In the rest of the proof, we shall proceed under this event. Write the eigenvalue decomposition of S_Z by

$$S_Z = \sum_{k=1}^{T-1} \widehat{\alpha}_k \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T}.$$

Here, the summation is taken up to $T-1$ because S_Z has rank $T-1$. It then follows that

$$S_Z^+ = \sum_{k=1}^{T-1} \frac{1}{\widehat{\alpha}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T}.$$

Write also

$$P_Z = \sum_{k=1}^{T-1} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T},$$

the projection matrix onto the linear space spanned by $Z_i - \bar{Z}$, $i = 1, \dots, T$.

By Cauchy-Schwartz inequality

$$\begin{aligned} s(\widehat{\mathbf{w}}_{\mathrm{mv}}) &= \frac{E^\mathsf{T} \Sigma^{-1/2} S_Z^+ \Sigma^{-1/2} \widehat{E}}{\left(\widehat{E}^\mathsf{T} \Sigma^{-1/2} S_Z^+ S_Z^+ \Sigma^{-1/2} \widehat{E}\right)^{1/2}} \\ &= \frac{E^\mathsf{T} \Sigma^{-1/2} P_Z S_Z^+ \Sigma^{-1/2} \widehat{E}}{\left(\widehat{E}^\mathsf{T} \Sigma^{-1/2} S_Z^+ S_Z^+ \Sigma^{-1/2} \widehat{E}\right)^{1/2}} \\ &\leq \left(E^\mathsf{T} \Sigma^{-1/2} P_Z P_Z \Sigma^{-1/2} E\right)^{1/2} \\ &=: \|P_Z \mathbf{u}\|, \end{aligned}$$

where $\mathbf{u} = \Sigma^{-1/2} E$. Note that P_Z is Haar distributed on the orthogonal group. Therefore,

$$\mathbb{E}\|P_Z \mathbf{u}\|^2 = \frac{T-1}{N} \|\mathbf{u}\|^2.$$

By law of large numbers,

$$\|P_Z \mathbf{u}\|^2 = \frac{T}{N} \|\mathbf{u}\|^2 (1 + o_p(1)).$$

See, e.g., Chikuse (2003). The proof is now completed by noting that

$$\|\mathbf{u}\|^2 = E^\mathsf{T} \Sigma^{-1} E = s^2(\mathbf{w}_{\mathrm{mv}}).$$

Appendix B – Proof of Proposition 1

Observe that the map $\mathbb{R}^d \to \mathcal{P} : \mathbf{x} \mapsto P_{\mathcal{P}}\mathbf{x}$ is bijective. Let $\mathbf{x}_{\mathcal{P}} \in \mathbb{R}^d$ be such that $\mathbf{w}^{\mathcal{P}}(\mu) = P_{\mathcal{P}}\mathbf{x}_{\mathcal{P}}$ solves

$$\min_{\mathbf{w}\in\mathcal{P}} \left\{ \frac{\gamma}{2}\mathbf{w}^\mathsf{T}\Sigma\mathbf{w} - \mathbf{w}^\mathsf{T}E \right\}.$$

Then $\mathbf{x}_{\mathcal{P}}$ solves

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left\{ \frac{\gamma}{2}\mathbf{x}^\mathsf{T}P_{\mathcal{P}}^\mathsf{T}\Sigma P_{\mathcal{P}}\mathbf{x} - \mathbf{x}^\mathsf{T}P_{\mathcal{P}}^\mathsf{T}E \right\}.$$

Therefore

$$\mathbf{x}_{\mathcal{P}} = \frac{1}{\gamma}(P_{\mathcal{P}}^\mathsf{T}\Sigma P_{\mathcal{P}})^{-1}P_{\mathcal{P}}^\mathsf{T}E,$$

which yields the claimed formula for $\mathbf{w}_{\mathrm{mv}}^{\mathcal{P}}$. ∎

Appendix C – Proof of Theorem 2

The proof is somewhat lengthy and we break it into several steps for clarity.

C.1 – Convergence of leading eigen-pairs of $\widehat{\Sigma}$

One of the main technical tools needed is the asymptotic properties of the eigenvalue and eigenvector pairs of $\widehat{\Sigma}$. It is well known that the eigen-pairs of $\widehat{\Sigma}$ may not converge to those of Σ when the number N of assets is large when compared with the length T of estimation window. See, e.g., Johnstone and Lu (2009). We shall show here that under the approximate factor model, the leading ones nonetheless behave reasonably with appropriate scaling.

To this end, denote by

$$\bar{\mathbf{f}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t, \qquad \bar{\boldsymbol{\epsilon}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\epsilon}_t,$$

and

$$\begin{aligned} \widehat{\Sigma}_f &= \frac{1}{T-1}\sum_{t=1}^{T}(\mathbf{f}_t - \bar{\mathbf{f}})(\mathbf{f}_t - \bar{\mathbf{f}})^\mathsf{T}, \\ \widehat{\Sigma}_\varepsilon &= \frac{1}{T-1}\sum_{t=1}^{T}(\boldsymbol{\epsilon}_t - \bar{\boldsymbol{\epsilon}})(\boldsymbol{\epsilon}_t - \bar{\boldsymbol{\epsilon}})^\mathsf{T}. \end{aligned}$$

Then

$$\widehat{\Sigma} = B\widehat{\Sigma}_f B^\mathsf{T} + \widehat{\Sigma}_\varepsilon + \frac{1}{T-1}\sum_{t=1}^T B(\mathbf{f}_t - \bar{\mathbf{f}})(\boldsymbol{\epsilon}_t - \bar{\boldsymbol{\epsilon}})^\mathsf{T} + \frac{1}{T-1}\sum_{t=1}^T (\boldsymbol{\epsilon}_t - \bar{\boldsymbol{\epsilon}})(\mathbf{f}_t - \bar{\mathbf{f}})^\mathsf{T} B^\mathsf{T}.$$

Let

$$\frac{1}{N}\widehat{\Sigma} = \sum_k \widehat{\lambda}_k \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T}, \qquad \text{and} \qquad \frac{1}{N} B\Sigma_f B^\mathsf{T} = \sum_{k=1}^K \lambda_k \eta_k \eta_k^\mathsf{T}$$

be their respective eigenvalue decomposition. For brevity, we shall assume that all positive eigenvalues of $B\Sigma_f B^\mathsf{T}/N$ has multiplicity one, i.e., $\lambda_1 > \lambda_2 > \ldots > \lambda_K$. The more general case can be treated in an identical fashion. We first show that

$$\widehat{\lambda}_k - \lambda_k = O_p\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{N}}\right), \qquad \text{and} \qquad \|\widehat{\eta}_k - \eta_k\| = O_p\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{N}}\right), \tag{11}$$

for $k = 1, \ldots, K$. In the light of the classical results from Bhatia, Davis and McIntosh (1983), it suffices to show that

$$\left\|\frac{1}{N}\left(\widehat{\Sigma} - B\Sigma_f B^\mathsf{T}\right)\right\|_\mathrm{F} = O_p\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{N}}\right). \tag{12}$$

To this end, we first observe that

$$\widehat{\Sigma} - B\Sigma_f B^\mathsf{T} = B(\widehat{\Sigma}_f - \Sigma_f)B^\mathsf{T} + \widehat{\Sigma}_{\boldsymbol{\epsilon}} + \frac{1}{T-1}\sum_{t=1}^T B(\mathbf{f}_t - \bar{\mathbf{f}})(\boldsymbol{\epsilon}_t - \bar{\boldsymbol{\epsilon}})^\mathsf{T} + \frac{1}{T-1}\sum_{t=1}^T (\boldsymbol{\epsilon}_t - \bar{\boldsymbol{\epsilon}})(\mathbf{f}_t - \bar{\mathbf{f}})^\mathsf{T} B^\mathsf{T}.$$

We shall now bound the four terms on the right hand side individually. It is clear that

$$\frac{1}{N^2}\left\|B(\widehat{\Sigma}_f - \Sigma_f)B^\mathsf{T}\right\|_\mathrm{F}^2 \le \frac{1}{N^2}\|B\|_2^4 \|\widehat{\Sigma}_f - \Sigma_f\|_2 = O_p\left(\frac{1}{T}\right), \tag{13}$$

We then consider the second term.

$$\begin{aligned}
\mathbb{E}\|\widehat{\Sigma}_\varepsilon\|_\mathrm{F}^2 &= \mathbb{E}\left\|\frac{1}{T-1}\sum_{t=1}^T (\boldsymbol{\epsilon}_t - \bar{\boldsymbol{\epsilon}})(\boldsymbol{\epsilon}_t - \bar{\boldsymbol{\epsilon}})^\mathsf{T}\right\|_\mathrm{F}^2 \\
&= \frac{1}{(T-1)^2}\mathbb{E}\left\|\sum_{t=1}^T \boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^\mathsf{T} - T\bar{\boldsymbol{\epsilon}}\bar{\boldsymbol{\epsilon}}^\mathsf{T}\right\|_\mathrm{F}^2.
\end{aligned}$$

Recall that if A_1, A_2 are both positive definite and $A_1 - A_2$ is also positive definite, then $\|A_1\|_\mathrm{F} \ge \|A_2\|_\mathrm{F}$ where $\|\cdot\|_\mathrm{F}$ is the Frobenius norm. Using this fact, we have

$$\begin{aligned}
\frac{1}{(T-1)^2}\mathbb{E}\left\|\sum_{t=1}^T \boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^\mathsf{T} - T\bar{\boldsymbol{\epsilon}}\bar{\boldsymbol{\epsilon}}^\mathsf{T}\right\|_\mathrm{F}^2 &\le \frac{1}{(T-1)^2}\mathbb{E}\left\|\sum_{t=1}^T \boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^\mathsf{T}\right\|_\mathrm{F}^2 \\
&= \frac{1}{(T-1)^2}\mathbb{E}\left\|\sum_{t=1}^T \left(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^\mathsf{T} - \Sigma_\varepsilon\right)\right\|_\mathrm{F}^2 + \frac{T^2}{(T-1)^2}\|\Sigma_\varepsilon\|_\mathrm{F}^2.
\end{aligned}$$

By Assumption E, we get

$$\mathbb{E}\left\|\sum_{t=1}^{T}\left(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^{\mathsf{T}}-\Sigma_{\varepsilon}\right)\right\|_{\mathrm{F}}^{2}\le C_4N^2T.$$

On the other hand,

$$\|\Sigma_{\varepsilon}\|_{\mathrm{F}}^{2}\le N\lambda_{\max}(\Sigma_{\varepsilon})=O(N)$$

following Assumption C, where $\lambda_{\max}(A)$ is the largest eigenvalue of A. Together, we have

$$\mathbb{E}\left\|\frac{1}{N}\widehat{\Sigma}_{\varepsilon}\right\|_{\mathrm{F}}^{2}=O\left(\frac{1}{N}+\frac{1}{T}\right). \tag{14}$$

We finally bound the last two terms. Note that

$$\mathbb{E}\left\|\sum_{t=1}^{T}B(\mathbf{f}_t-\bar{\mathbf{f}})(\boldsymbol{\epsilon}_t-\bar{\boldsymbol{\epsilon}})^{\mathsf{T}}\right\|_{\mathrm{F}}^{2}\le\|B\|_{\mathrm{F}}^{2}\mathbb{E}\left\|\sum_{t=1}^{T}(\mathbf{f}_t-\bar{\mathbf{f}})(\boldsymbol{\epsilon}_t-\bar{\boldsymbol{\epsilon}})^{\mathsf{T}}\right\|_{\mathrm{F}}^{2}.$$

By Assumption B,

$$\|B\|_2^2\le\|B\|_{\mathrm{F}}^2=O(N).$$

On the other hand, by Assumption F,

$$\mathbb{E}\left\|\sum_{t=1}^{T}(\mathbf{f}_t-\bar{\mathbf{f}})(\boldsymbol{\epsilon}_t-\bar{\boldsymbol{\epsilon}})^{\mathsf{T}}\right\|_{\mathrm{F}}^{2}=\sum_{k=1}^{K}\sum_{j=1}^{N}\mathbb{E}\left(\sum_{t=1}^{T}(f_{kt}-\bar{\mathbf{f}}_k)(\varepsilon_{jt}-\bar{\boldsymbol{\epsilon}}_j)\right)^2=O(NT).$$

Thus,

$$\mathbb{E}\left\|\frac{1}{N}\frac{1}{T-1}\sum_{t=1}^{T}B(\mathbf{f}_t-\bar{\mathbf{f}})(\boldsymbol{\epsilon}_t-\bar{\boldsymbol{\epsilon}})^{\mathsf{T}}\right\|_{\mathrm{F}}^{2}=O\left(\frac{1}{T}\right). \tag{15}$$

Equation (12), and subsequently (11), follow immediately from (13), (14) and (15).

C.2 – Out-of-sample mean of $\widehat{\mathbf{w}}_d$

Equipped with (11), we now investigate the out-of-sample mean of the estimated subspace mean-variance portfolio $\tilde{\mathbf{w}}$. Recall that

$$\begin{aligned}\widehat{\mathbf{w}}_d &= \frac{1}{\gamma}\sum_{k=1}^{K}\frac{1}{N\widehat{\lambda}_k}\left(\widehat{\eta}_k^{\mathsf{T}}\widehat{E}\right)\widehat{\eta}_k\\ &= \frac{1}{\gamma}\sum_{k=1}^{K}\frac{1}{N\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^{\mathsf{T}}B\boldsymbol{\mu}_f+\frac{1}{\gamma}\sum_{k=1}^{K}\frac{1}{N\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^{\mathsf{T}}B\bar{\mathbf{f}}+\frac{1}{\gamma}\sum_{k=1}^{K}\frac{1}{N\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^{\mathsf{T}}\bar{\boldsymbol{\epsilon}},\end{aligned}$$

where

$$\widehat{E} = B\boldsymbol{\mu}_f + B\bar{\mathbf{f}} + \bar{\boldsymbol{\epsilon}}.$$

is the mean returns estimated from historical data. It is not hard to see that

$$\begin{aligned} E^\mathsf{T}\widehat{\mathbf{w}}_d &= \frac{1}{N\gamma}\boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T}\left(\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^\mathsf{T}\right) B\boldsymbol{\mu}_f + \frac{1}{N\gamma}\boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T}\left(\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^\mathsf{T}\right) B\bar{\mathbf{f}} \\ &\quad + \frac{1}{N\gamma}\boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T}\left(\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^\mathsf{T}\right)\bar{\boldsymbol{\epsilon}}. \end{aligned}$$

We now analyze the three terms on the right hand side separately.

We begin with the first term.

$$\begin{aligned} &\frac{1}{N}\boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T}\left(\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^\mathsf{T}\right) B\boldsymbol{\mu}_f \\ =\ &\frac{1}{N}\boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T}\left(\frac{1}{N}B\Sigma_f B^\mathsf{T}\right)^+ B\boldsymbol{\mu}_f + \frac{1}{N}\boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T}\left\{\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^\mathsf{T} - \left(\frac{1}{N}B\Sigma_f B^\mathsf{T}\right)^+\right\} B\boldsymbol{\mu}_f \\ =\ &\boldsymbol{\mu}_f^\mathsf{T}\Sigma_f^{-1}\boldsymbol{\mu}_f + \frac{1}{N}\boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T}\left\{\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^\mathsf{T} - \left(\frac{1}{N}B\Sigma_f B^\mathsf{T}\right)^+\right\} B\boldsymbol{\mu}_f. \end{aligned}$$

Note that

$$\begin{aligned} &\left|\frac{1}{N}\boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T}\left\{\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^\mathsf{T} - \left(\frac{1}{N}B\Sigma_f B^\mathsf{T}\right)^+\right\} B\boldsymbol{\mu}_f\right| \\ \leq\ &\frac{1}{N}\left\|\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^\mathsf{T} - \left(\frac{1}{N}B\Sigma_f B^\mathsf{T}\right)^+\right\|_2 \boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T} B\boldsymbol{\mu}_f. \end{aligned}$$

where $\|\cdot\|_2$ stands for the usual matrix spectral norm. Observe that

$$\left(\frac{1}{N}B\Sigma_f B^\mathsf{T}\right)^+ = \sum_{k=1}^K \frac{1}{\lambda_k}\eta_k\eta_k^\mathsf{T}.$$

By (11),

$$\left\|\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^\mathsf{T} - \left(\frac{1}{N}B\Sigma_f B^\mathsf{T}\right)^+\right\|_2 = O_p\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{N}}\right). \tag{16}$$

Therefore,

$$\frac{1}{N}\boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T}\left(\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^\mathsf{T}\right) B\boldsymbol{\mu}_f = \boldsymbol{\mu}_f^\mathsf{T}\Sigma_f^{-1}\boldsymbol{\mu}_f + O_p\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{N}}\right).$$

Next consider the second term in the expression of $E^{\mathsf{T}}\widehat{\mathbf{w}}_d$.

$$\begin{aligned}\left|\frac{1}{N}\boldsymbol{\mu}_f^{\mathsf{T}}B^{\mathsf{T}}\left(\sum_{k=1}^{K}\frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^{\mathsf{T}}\right)B\bar{\mathbf{f}}\right| &\leq \frac{1}{N}\left\|\sum_{k=1}^{K}\frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^{\mathsf{T}}\right\|_2\|B\boldsymbol{\mu}_f\|\|B\bar{\mathbf{f}}\| \\ &\leq \frac{1}{N}\left\|\sum_{k=1}^{K}\frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^{\mathsf{T}}\right\|_2\|B\|_2^2\|\boldsymbol{\mu}_f\|\|\bar{\mathbf{f}}\|.\end{aligned}$$

By triangular inequality,

$$\begin{aligned}\left\|\sum_{k=1}^{K}\frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^{\mathsf{T}}\right\|_2 &\leq \left\|\left(\frac{1}{N}B\Sigma_f B^{\mathsf{T}}\right)^+\right\|_2+\left\|\sum_{k=1}^{K}\frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^{\mathsf{T}}-\left(\frac{1}{N}B\Sigma_f B^{\mathsf{T}}\right)^+\right\|_2 \\ &= \lambda_K^{-1}+O_p\left(\frac{1}{\sqrt{T}}+\frac{1}{\sqrt{N}}\right).\end{aligned}$$

By Assumption B,

$$\|B\|_2^2\leq\|B\|_{\mathrm{F}}^2=O(N).$$

Together with the fact that

$$\|\bar{\mathbf{f}}\|=O_p(T^{-1/2}),$$

we get

$$\left|\frac{1}{N}\boldsymbol{\mu}_f^{\mathsf{T}}B^{\mathsf{T}}\left(\sum_{k=1}^{K}\frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^{\mathsf{T}}\right)B\bar{\mathbf{f}}\right|=O_p\left(\frac{1}{T^{1/2}}\right).$$

Similarly the last term can also be bounded.

$$\begin{aligned}\left|\frac{1}{N}\boldsymbol{\mu}_f^{\mathsf{T}}B^{\mathsf{T}}\left(\sum_{k=1}^{K}\frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^{\mathsf{T}}\right)\bar{\boldsymbol{\epsilon}}\right| &\leq \frac{1}{N}\left\|\sum_{k=1}^{K}\frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^{\mathsf{T}}\right\|_2\|B\boldsymbol{\mu}_f\|\|\bar{\boldsymbol{\epsilon}}\| \\ &\leq \frac{1}{N}\left\|\sum_{k=1}^{K}\frac{1}{\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^{\mathsf{T}}\right\|_2\|B\|_2\|\boldsymbol{\mu}_f\|\|\bar{\boldsymbol{\epsilon}}\| \\ &\leq O_p\left(\frac{1}{T^{1/2}}\right).\end{aligned}$$

In conclusion, we have

$$E^{\mathsf{T}}\widehat{\mathbf{w}}_d=\boldsymbol{\mu}_f^{\mathsf{T}}\Sigma_f^{-1}\boldsymbol{\mu}_f+O_p\left(\frac{1}{\sqrt{T}}+\frac{1}{\sqrt{N}}\right). \tag{17}$$

C.3 – Out-of-sample Variance of $\widehat{\mathbf{w}}_d$

It remains to consider the out-of-sample variance of $\widehat{\mathbf{w}}_d$:

$$\widehat{\mathbf{w}}_d^{\mathsf{T}}\Sigma\widehat{\mathbf{w}}_d=\widehat{\mathbf{w}}_d^{\mathsf{T}}\Sigma_\varepsilon\widehat{\mathbf{w}}_d+\widehat{\mathbf{w}}_d^{\mathsf{T}}B\Sigma_f B^{\mathsf{T}}\widehat{\mathbf{w}}_d. \tag{18}$$

We shall bound the two terms on the right hand side separately.

C.3.1 – Bounding $\widehat{\mathbf{w}}_d^{\mathsf{T}} \Sigma_\varepsilon \widehat{\mathbf{w}}_d$

Observe that

$$\begin{aligned}\widehat{\mathbf{w}}_d^{\mathsf{T}} \Sigma_\varepsilon \widehat{\mathbf{w}}_d &= \widehat{E}^{\mathsf{T}} \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^{\mathsf{T}} \right) \Sigma_\varepsilon \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^{\mathsf{T}} \right) \widehat{E} \\ &\leq \frac{1}{N^2} \left\| \left(\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^{\mathsf{T}} \right) \Sigma_\varepsilon \left(\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^{\mathsf{T}} \right) \right\|_2 \|\widehat{E}\|^2.\end{aligned}$$

By triangular inequality,

$$\begin{aligned}&\left\| \left(\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^{\mathsf{T}} \right) \Sigma_\varepsilon \left(\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^{\mathsf{T}} \right) \right\|_2 \\ \leq\ & \left\| \left(B\Sigma_f B^{\mathsf{T}}\right)^+ \Sigma_\varepsilon \left(B\Sigma_f B^{\mathsf{T}}\right)^+ \right\|_2 \\ &+ \left\| \left(B\Sigma_f B^{\mathsf{T}}\right)^+ \Sigma_\varepsilon \left(B\Sigma_f B^{\mathsf{T}}\right)^+ - \left(B\Sigma_f B^{\mathsf{T}}\right)^+ \Sigma_\varepsilon \left(\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^{\mathsf{T}} \right) \right\|_2 \\ &+ \left\| \left(\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^{\mathsf{T}} \right) \Sigma_\varepsilon \left(B\Sigma_f B^{\mathsf{T}}\right)^+ - \left(\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^{\mathsf{T}} \right) \Sigma_\varepsilon \left(\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^{\mathsf{T}} \right) \right\|_2\end{aligned}$$

In the light of (16), we get

$$\left\| \left(\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^{\mathsf{T}} \right) \Sigma_\varepsilon \left(\sum_{k=1}^K \frac{1}{\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^{\mathsf{T}} \right) \right\|_2 \to_p \left\| \left(B\Sigma_f B^{\mathsf{T}}\right)^+ \Sigma_\varepsilon \left(B\Sigma_f B^{\mathsf{T}}\right)^+ \right\|_2 \leq \lambda_K^{-2} \|\Sigma_\varepsilon\|_2 = O(1).$$

Next, we bound $\|\widehat{E}\|^2$. Note that

$$\|\widehat{E}\|^2 = \boldsymbol{\mu}_f^{\mathsf{T}} B^{\mathsf{T}} B \boldsymbol{\mu}_f + \bar{\mathbf{f}}^{\mathsf{T}} B^{\mathsf{T}} B \bar{\mathbf{f}} + \bar{\boldsymbol{\epsilon}}^{\mathsf{T}} \bar{\boldsymbol{\epsilon}} + 2\boldsymbol{\mu}_f^{\mathsf{T}} B^{\mathsf{T}} B \bar{\mathbf{f}} + \boldsymbol{\mu}_f^{\mathsf{T}} B^{\mathsf{T}} \bar{\boldsymbol{\epsilon}} + 2\bar{\mathbf{f}}^{\mathsf{T}} B^{\mathsf{T}} \bar{\boldsymbol{\epsilon}}.$$

Recall that $\mathbf{f}$ is the sample mean of T zero mean random vectors. Therefore $\mathbb{E}\mathbf{f}_k^2 = O(T^{-1})$ for any $1 \leq k \leq K$. As K is finite, we get

$$\mathbb{E}\|\bar{\mathbf{f}}\|^2 = O(T^{-1}).$$

Similarly, $\bar{\boldsymbol{\epsilon}}$ is the sample mean of T zero mean random vectors of dimension N. Thus,

$$\mathbb{E}\|\bar{\boldsymbol{\epsilon}}\|^2 = O(NT^{-1}).$$

Together with the facts that

$$\|B\|_2^2 = O(N), \qquad \text{and} \qquad \|\boldsymbol{\mu}_f\| = O(1),$$

we get

$$
\begin{aligned}
\boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T} B \boldsymbol{\mu}_f \le \|B\|_2^2 \|\boldsymbol{\mu}_f\|^2 &= O(N); \\
\bar{\mathbf{f}}^\mathsf{T} B^\mathsf{T} B \bar{\mathbf{f}} \le \|B\|_2^2 \|\bar{\mathbf{f}}\|^2 &= O_p(NT^{-1}); \\
\left|\boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T} B \bar{\mathbf{f}}\right| \le \|B\|_2^2 \|\boldsymbol{\mu}_f\| \|\bar{\mathbf{f}}\| &= O_p(NT^{-1/2}); \\
\left|\boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T} \bar{\boldsymbol{\epsilon}}\right| \le \|B\|_2 \|\boldsymbol{\mu}_f\| \|\bar{\boldsymbol{\epsilon}}\| &= O_p(NT^{-1/2}); \\
\left|\bar{\mathbf{f}}^\mathsf{T} B^\mathsf{T} \bar{\boldsymbol{\epsilon}}\right| \le \|B\|_2 \|\bar{\mathbf{f}}\| \|\bar{\boldsymbol{\epsilon}}\| &= O_p(NT^{-1}),
\end{aligned}
$$

which imply that $\|\widehat{E}\|^2 = O_p(N)$. Thus,

$$
\widehat{\mathbf{w}}_d^\mathsf{T} \Sigma_\varepsilon \widehat{\mathbf{w}}_d = O_p\left(\frac{1}{N}\right).
$$

C.3.2 – Bounding $\widehat{\mathbf{w}}_d^\mathsf{T} B \Sigma_f B^\mathsf{T} \widehat{\mathbf{w}}_d$

It remains to bound the second term on the right hand side of (18). Write

$$
\begin{aligned}
\widehat{\mathbf{w}}_d^\mathsf{T} B \Sigma_f B^\mathsf{T} \widehat{\mathbf{w}}_d &= \boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T} \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T}\right) B \Sigma_f B^\mathsf{T} \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T}\right) B \boldsymbol{\mu}_f \\
&\quad + (\widehat{E} - B\boldsymbol{\mu}_f)^\mathsf{T} \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T}\right) B \Sigma_f B^\mathsf{T} \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T}\right) (\widehat{E} - B\boldsymbol{\mu}_f) \\
&\quad + 2(\widehat{E} - B\boldsymbol{\mu}_f)^\mathsf{T} \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T}\right) B \Sigma_f B^\mathsf{T} \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T}\right) B \boldsymbol{\mu}_f.
\end{aligned}
$$

We now analyze the three terms on the right hand side separately.

$$
\begin{aligned}
& \boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T} \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T}\right) B \Sigma_f B^\mathsf{T} \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T}\right) B \boldsymbol{\mu}_f \\
= \ & \boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T} \left(B \Sigma_f B^\mathsf{T}\right)^+ B \Sigma_f B^\mathsf{T} \left(B \Sigma_f B^\mathsf{T}\right)^+ B \boldsymbol{\mu}_f \\
& \quad + \boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T} \left\{\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T} - \left(B \Sigma_f B^\mathsf{T}\right)^+\right\} B \Sigma_f B^\mathsf{T} \left\{\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T} - \left(B \Sigma_f B^\mathsf{T}\right)^+\right\} B \boldsymbol{\mu}_f \\
& \quad + 2\boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T} \left\{\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T} - \left(B \Sigma_f B^\mathsf{T}\right)^+\right\} B \Sigma_f B^\mathsf{T} \left(B \Sigma_f B^\mathsf{T}\right)^+ B \boldsymbol{\mu}_f
\end{aligned}
$$

Observe that

$$
\begin{aligned}
& \boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T} \left(B \Sigma_f B^\mathsf{T}\right)^+ B \Sigma_f B^\mathsf{T} \left(B \Sigma_f B^\mathsf{T}\right)^+ B \boldsymbol{\mu}_f \\
= \ & \boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T} \left(B \Sigma_f B^\mathsf{T}\right)^+ B \boldsymbol{\mu}_f \\
= \ & \boldsymbol{\mu}_f^\mathsf{T} \Sigma_f^{-1} \boldsymbol{\mu}_f.
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
& \boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T} \left\{ \sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T} - \left(B\Sigma_f B^\mathsf{T}\right)^+ \right\} B\Sigma_f B^\mathsf{T} \left\{ \sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T} - \left(B\Sigma_f B^\mathsf{T}\right)^+ \right\} B\boldsymbol{\mu}_f \\
\leq \; & \left\| \sum_{k=1}^K \frac{1}{\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T} - \left(\frac{1}{N} B\Sigma_f B^\mathsf{T}\right)^+ \right\|_2^2 \left\| \frac{1}{N} B\Sigma_f B^\mathsf{T} \right\|_2 \left(\frac{1}{N} \|B\boldsymbol{\mu}_f\|^2 \right) \\
= \; & O_p\left(\frac{1}{N} + \frac{1}{T} \right).
\end{aligned}
$$

By Cauchy-Schwartz inequality,

$$
\begin{aligned}
& \left| \boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T} \left\{ \sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T} - \left(B\Sigma_f B^\mathsf{T}\right)^+ \right\} B\Sigma_f B^\mathsf{T} \left(B\Sigma_f B^\mathsf{T}\right)^+ B\boldsymbol{\mu}_f \right| \\
\leq \; & \left(\boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T} \left(B\Sigma_f B^\mathsf{T}\right)^+ B\Sigma_f B^\mathsf{T} \left(B\Sigma_f B^\mathsf{T}\right)^+ B\boldsymbol{\mu}_f \right)^{1/2} \\
& \times \left(\boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T} \left\{ \sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T} - \left(B\Sigma_f B^\mathsf{T}\right)^+ \right\} B\Sigma_f B^\mathsf{T} \left\{ \sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T} - \left(B\Sigma_f B^\mathsf{T}\right)^+ \right\} B\boldsymbol{\mu}_f \right)^{1/2} \\
= \; & O_p\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{N}} \right).
\end{aligned}
$$

This shows that

$$
\boldsymbol{\mu}_f^\mathsf{T} B^\mathsf{T} \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T} \right) B\Sigma_f B^\mathsf{T} \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T} \right) B\boldsymbol{\mu}_f = \boldsymbol{\mu}_f^\mathsf{T} \Sigma_f^{-1} \boldsymbol{\mu}_f + O_p\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{N}} \right).
$$

Similarly, the second term can be bounded by

$$
\begin{aligned}
& (\widehat{E} - B\boldsymbol{\mu}_f)^\mathsf{T} \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T} \right) B\Sigma_f B^\mathsf{T} \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T} \right) (\widehat{E} - B\boldsymbol{\mu}_f) \\
= \; & (\widehat{E} - B\boldsymbol{\mu}_f)^\mathsf{T} \left(B\Sigma_f B^\mathsf{T}\right)^+ B\Sigma_f B^\mathsf{T} \left(B\Sigma_f B^\mathsf{T}\right)^+ (\widehat{E} - B\boldsymbol{\mu}_f) \\
& \quad + (\widehat{E} - B\boldsymbol{\mu}_f)^\mathsf{T} \left\{ \sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T} - \left(B\Sigma_f B^\mathsf{T}\right)^+ \right\} B\Sigma_f B^\mathsf{T} \\
& \qquad \times \left\{ \sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T} - \left(B\Sigma_f B^\mathsf{T}\right)^+ \right\} (\widehat{E} - B\boldsymbol{\mu}_f) \\
& \quad + 2(\widehat{E} - B\boldsymbol{\mu}_f)^\mathsf{T} \left\{ \sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k} \widehat{\eta}_k \widehat{\eta}_k^\mathsf{T} - \left(B\Sigma_f B^\mathsf{T}\right)^+ \right\} B\Sigma_f B^\mathsf{T} \left(B\Sigma_f B^\mathsf{T}\right)^+ (\widehat{E} - B\boldsymbol{\mu}_f) \\
= \; & O_p\left(\frac{1}{T} \right) + O_p\left(\frac{1}{NT} + \frac{1}{T^2} \right) + O_p\left(\frac{1}{N^{1/2}T} + \frac{1}{T^{3/2}} \right).
\end{aligned}
$$

Now, the third term can be bounded using Cauchy Schwartz inequality:

$$\begin{aligned}
& \left|(\widehat{E} - B\boldsymbol{\mu}_f)^\top \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^\top\right) B\Sigma_f B^\top \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^\top\right) B\boldsymbol{\mu}_f\right| \\
\leq & \left\{\boldsymbol{\mu}_f^\top B^\top \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^\top\right) B\Sigma_f B^\top \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^\top\right) B\boldsymbol{\mu}_f\right\}^{1/2} \\
& \times \left\{(\widehat{E} - B\boldsymbol{\mu}_f)^\top \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^\top\right) B\Sigma_f B^\top \left(\sum_{k=1}^K \frac{1}{N\widehat{\lambda}_k}\widehat{\eta}_k\widehat{\eta}_k^\top\right) (\widehat{E} - B\boldsymbol{\mu}_f)\right\}^{1/2} \\
= & \; O_p\left(\frac{1}{\sqrt{T}}\right)
\end{aligned}$$

To sum up, we have

$$\widehat{\mathbf{w}}_d^\top B\Sigma_f B^\top \widehat{\mathbf{w}}_d = \boldsymbol{\mu}_f^\top \Sigma_f^{-1}\boldsymbol{\mu}_f + O_p\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{N}}\right). \tag{19}$$

Together with (17), we get

$$\begin{aligned}
s^2(\widehat{\mathbf{w}}_d) &= \frac{(E^\top \widehat{\mathbf{w}}_d)^2}{\widehat{\mathbf{w}}_d^\top B\Sigma_f B^\top \widehat{\mathbf{w}}_d} \\
&= (\boldsymbol{\mu}_f^\top \Sigma_f^{-1}\boldsymbol{\mu}_f)^{-1} + O_p\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{N}}\right) \\
&= s^2(\mathbf{w}_{\mathrm{mv}}) + O_p\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{N}}\right),
\end{aligned}$$

which completes the proof.

Appendix D – Alternative Portfolio Rules

For completeness, we list here the alternative portfolio rules included in the numerical experiments from Section 4.

- Naïve diversification

$$\mathbf{w}_{\mathrm{e}} = \frac{1}{N}\mathbf{1}.$$

- The portfolio rule of Jorion (1986) is based on Bayes-Stein estimates of E and Σ:

$$\widehat{E}^{\mathrm{BS}} = (1-\nu)\widehat{E} + \nu\widehat{E}_g\mathbf{1}$$

and

$$\widehat{\Sigma}_{\text{BS}} = \left(1 + \frac{1}{T+\lambda}\right)\tilde{\Sigma} + \frac{\lambda}{T(T+1+\lambda)}\frac{\mathbf{1}\mathbf{1}^{\mathsf{T}}}{\mathbf{1}^{\mathsf{T}}\tilde{\Sigma}^{-1}\mathbf{1}}$$

where

$$\widehat{E}_g = \frac{\mathbf{1}^{\mathsf{T}}\widehat{\Sigma}^{-1}\widehat{E}}{\mathbf{1}^{\mathsf{T}}\widehat{\Sigma}^{-1}\mathbf{1}}, \qquad \nu = \frac{N+2}{N+2+T(\widehat{E}-\widehat{E}_g\mathbf{1})^{\mathsf{T}}\tilde{\Sigma}^{-1}(\widehat{E}-\widehat{E}_g\mathbf{1})},$$

and

$$\tilde{\Sigma} = \frac{T-1}{T-N-2}\widehat{\Sigma}, \qquad \lambda = (N+2)/[(\widehat{E}-\widehat{E}_g\mathbf{1})^{\mathsf{T}}\tilde{\Sigma}^{-1}(\widehat{E}-\widehat{E}_g\mathbf{1})].$$

The portfolio weights are then given by

$$\widehat{\mathbf{w}}^{\text{PJ}} = \frac{1}{\gamma}(\widehat{\Sigma}^{\text{BS}})^{-1}\widehat{E}^{\text{BS}}.$$

- The rule from Kan and Zhou (2007):

$$\widehat{\mathbf{w}}_{\text{KZ}} = \frac{1}{\gamma}\frac{T-N-2}{c(T-1)}\left[\eta\widehat{\Sigma}^{-1}\widehat{E} + (1-\eta)\widehat{E}_g\widehat{\Sigma}^{-1}\mathbf{1}\right],$$

where

$$\eta = \frac{\psi^2}{\psi^2 + N/T}, \qquad \psi^2 = (\widehat{E}-\widehat{E}_g\mathbf{1})^{\mathsf{T}}\widehat{\Sigma}^{-1}(\widehat{E}-\widehat{E}_g\mathbf{1}).$$

- Combination rule based on sample mean-variance portfolio and naïve diversification (Tu and Zhou, 2011):

$$\widehat{\mathbf{w}}_{\text{S\&N}} = (1-\delta)\mathbf{w}_{\text{e}} + \delta\frac{T-N-2}{T-1}\widehat{\mathbf{w}}_{\text{mv}},$$

where

$$\delta = \pi_1/(\pi_1+\pi_2)$$

and

$$\pi_1 = \mathbf{w}_{\text{e}}^{\mathsf{T}}\widehat{\Sigma}\mathbf{w}_{\text{e}} - \frac{2}{\gamma}\mathbf{w}_{\text{e}}^{\mathsf{T}}\widehat{E} + \frac{1}{\gamma^2}\tilde{\theta}^2,$$

$$\pi_2 = \frac{1}{\gamma^2}\left[\frac{(T-2)(T-N-2)}{(T-N-1)(T-N-4)} - 1\right]\tilde{\theta}^2 + \frac{(T-2)(T-N-2)}{\gamma^2(T-N-1)(T-N-4)}\frac{N}{T}.$$

Here

$$\tilde{\theta}^2 = \frac{(T-N-2)\theta^2 - N}{T} + \frac{2\theta^N(1+\theta^2)^{-(T-2)/2}}{T\int_0^{\theta^2/(1+\theta^2)} x^{N/2}(1-x)^{(T-N-2)/2}dx}$$

and

$$\theta^2 = \widehat{E}^{\mathsf{T}}\widehat{\Sigma}^{-1}\widehat{E}.$$

- Combination rule based on `KZ` and naïve diversification (Tu and Zhou, 2011)

$$\widehat{\mathbf{w}}^{\mathrm{KZ\&N}} = (1-\zeta)\mathbf{w}_{\mathrm{e}} + \zeta \widehat{\mathbf{w}}_{\mathrm{KZ}},$$

where

$$\zeta = \frac{\pi_1 - \pi_{13}}{\pi_1 - 2\pi_{13} + \pi_3},$$

and

$$\begin{aligned}\pi_{13} = \frac{1}{\gamma^2}\tilde{\theta}^2 - \frac{1}{\gamma}\mathbf{w}_{\mathrm{e}}^{\mathsf{T}}\widehat{E} + \frac{(T-N-1)(T-N-4)}{\gamma(T-2)(T-N-2)}\left[\eta \mathbf{w}_{\mathrm{e}}^{\mathsf{T}}\widehat{E} + (1-\eta)\widehat{E}_g \mathbf{w}_{\mathrm{e}}^{\mathsf{T}}\mathbf{1}\right] \\ -\frac{(T-N-1)(T-N-4)}{\gamma^2(T-2)(T-N-2)}\left[\eta \widehat{E}^{\mathsf{T}}\tilde{\Sigma}^{-1}\widehat{E} + (1-\eta)\widehat{E}_g \widehat{E}^{\mathsf{T}}\tilde{\Sigma}^{-1}\mathbf{1}\right],\end{aligned}$$

$$\pi_3 = \frac{1}{\gamma^2}\tilde{\theta}^2 - \frac{(T-N-1)(T-N-4)}{\gamma^2(T-2)(T-N-2)}\left(\tilde{\theta}^2 - \frac{N}{T}\eta\right).$$

References

[1] Anderson, T.W. (2003), *An Introduction to Multivariate Statistical Analysis*, 3rd Edition, New York: Wiley.

[2] Antoine, B. (2012), Portfolio selection with estimation risk: a test-based approach, *Journal of Financial Econometrics*, **10**, 164-197.

[3] Bai, J. and S. Ng (2002), Determining the number of factors in approximate factor models, *Econometrica*, **70**, 191-221.

[4] Barry, C. B. (1974), Portfolio analysis under uncertain means, variances, and covariances, *Journal of Finance*, **29**, 515-522.

[5] Bawa, V. S., S. Brown, and R. Klein (1979), *Estimation Risk and Optimal Portfolio Choice*, Amsterdam, The Netherlands: North Holland.

[6] Behr, P., A. Güttler and F. Miebs (2008), Is minimum-variance investing really worth the while? An analysis with robust performance inference.

[7] Benartzi, S. and R.H. Thaler (2001), Naïve diversification strategies in retirement saving plans, *American Economic Review*, **91(1)**, 79-98.