The Dispersion Bias

Lisa Goldberg* Alex Papanicolaou† Alex Shkolnik‡

December 14, 2017§

Abstract

Estimation error has plagued quantitative finance since Harry Markowitz launched modern portfolio theory in 1952. Using random matrix theory, we characterize a source of bias in the sample eigenvectors of financial covariance matrices. Unchecked, the bias distorts weights of minimum variance portfolios and leads to risk forecasts that are severely biased downward. To address these issues, we develop an eigenvector bias correction. Our approach is distinct from the regularization and eigenvalue shrinkage methods found in the literature. We provide theoretical guarantees on the improvement our correction provides as well as estimation methods for computing the optimal correction from data.

*Departments of Economics and Statistics and Consortium for Data Analytics in Risk, University of California, Berkeley, CA 94720 and Aperio Group, `lrg@berkeley.edu`.

†Consortium for Data Analytics in Risk, University of California, Berkeley, CA 94720, `apapanicolaou@berkeley.edu`.

‡Consortium for Data Analytics in Risk, University of California, Berkeley, CA 94720, `ads2@berkeley.edu`.

§We thank the Center for Risk Management Research, the Consortium for Data Analytics in Risk, and the Coleman Fung Chair for financial support. We thank Marco Avellaneda, Bob Anderson, Kay Giesecke, Guy Miller, George Papanicolaou, participants at the the 3rd Annual CDAR Symposium in Berkeley, participants at the Swissquote Conference 2017 on FinTech,q and participants at the UC Santa Barbara Seminar in Statistics and Applied Probability for discussion and comments. We are grateful to Stephen Bianchi, whose incisive experiment showing that it is errors in eigenvectors, and not in eigenvalues, that corrupt large minimum variance portfolios, pointed us in a good direction.

1 Introduction

Harry Markowitz transformed finance in 1952 by framing portfolio construction as a tradeoff between mean and variance of return. This application of mean-variance optimization is the basis of theoretical breakthroughs as fundamental as the Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT), as well as practical innovations as impactful as Exchange Traded Funds.[1] Still, all financial applications of mean-variance optimization suffer from estimation error in covariance matrices, and we highlight two difficulties. First, a portfolio that is optimized using an estimated covariance matrix is never the true Markowitz portfolio. Second, in current practice, the forecasted risk of the optimized portfolio is typically too low, sometimes by a wide margin. Thus, investors end up with the wrong portfolio, one that is riskier, perhaps a lot riskier, than anticipated.

In this article, we address these difficulties by correcting a systematic bias in the first eigenvector of a sample covariance matrix. Our setting is that of a typical factor model,[2] but our statistical setup differs from most recent literature. In the last two decades, theoretical and empirical emphasis has been on the case when the number of assets N and number of observations T are both large. In this regime, consistency of principal component analysis (PCA) estimates may be established (Bai & Ng 2008). Motivated by many applications, we consider the setting of relatively few observations (in asymptotic theory: N grows and T is fixed). Indeed, an investor often has a portfolio of thousands of securities but only hundreds of observations.[3] PCA is applied in this environment in the early, pioneering work by Connor & Korajczyk (1986) and Connor & Korajczyk (1988), but also very recently (Wang & Fan 2017). In this high dimension, low sample-size regime, PCA factor estimates necessarily carry a finite-sample bias.

An elementary simulation experiment reveals that in a large minimum variance portfolio, errors in portfolio weights are driven by the first principal component, not its variance.[4] The fact that the eigenvalues of the sample covariance matrix are not important requires some nontrivial analysis, which we carry out. In particular, we show (in our asymptotic regime) that the bias in the dominant sample eigenvalue does not effect the performance of the estimated minimum variance portfolio. Only the bias in the dominant sample eigenvector needs to be addressed. We measure portfolio performance using two metrics. Tracking error, the workhorse of financial practitioners, measures deviations in weights

[1] The seminal paper is Markowitz (1952). See Treynor (1962) and Sharpe (1964) for the Capital Asset Pricing Model and Ross (1976) for the Arbitrage Pricing Theory.

[2] In this setting, the eigenvalues of the covariance matrix corresponding to the factors grow linearly in the dimension. This is not the traditional random matrix theory setting in which all eigenvalues are bounded, nor that of "weak" factors, e.g., Onatski (2012).

[3] While high frequency data are available in some markets, many securities are observed only at a daily horizon or less frequently. Moreover, markets are non-stationary, so even when there is a long history of data available, its relevance to some problems is questionable.

[4] This experiment was first communicated to us by Stephen Bianchi.

between the estimated (optimized) and optimal portfolios. We use the variance forecast ratio, familiar to both academics and practitioners, to measure the accuracy of the risk forecast of the portfolio, however right or wrong that portfolio may be.

To develop some intuition for the results to come, consider a simplistic world where all security exposures to the dominant (market) factor are identical. With probability one, a PCA estimate of our idealized, dominant factor will have higher dispersion (variation in its entries). Decreasing this dispersion, obviously, mitigates the estimation error. We denote our idealized, dominant factor by z. We prove that the same argument applies to any other dominant factor along the direction of z with high probablity for N large. Thus moving our PCA estimate towards z, by some amount, is very likely to decrease estimation error. In the limit ($N \uparrow \infty$), the estimation error is reduced with probability one. The larger the component of the true dominant factor along z is, the more we can decrease the estimation error.

While a careful proof of our result relies on some recent theory on sample eigenvector asymptotics, rule of thumb versions have been known to practitioners since the 1970s (see footnote 10 and the corresponding discussion). Indeed, the dominant risk factor consistently found in the US and many other developed public equity markets has most (if not all) individual equities positively exposed to it. In other words, empirically, the dominant risk factor has a significant component in z. Our characterization of the dispersion bias may then be viewed as a formalization of standard operation procedure.

Related Literature. The impact of estimation error on optimized portfolios has been investigated thoroughly in simulation and emprical settings. For example, see Jobson & Korkie (1980), Britten-Jones (1999), Bianchi, Goldberg & Rosenberg (2017) and the references therein. DeMiguel, Garlappi & Uppal (2007) compare a variety of methods for mitigating estimation error, benchmarking againt the equally weigted portfolio in out-of-sample tests. They conclude that unreasonably long estimation windows are required for current methods to consistently outperform the benchmark. We review some methods most closely related to our approach.[5]

Early work on estimation error and the Markowitz problem was focused on Bayesian approaches. Vasicek (1973) and Frost & Savarino (1986) were perhaps the first to impose informative priors on the model parameters.[6] More realistic priors incorporating multi-factor modeling are analyzed in Pástor (2000) (sample

[5]The literature on this topic is extensive. We briefly mention a few important references that do not overlap at all with out work. Michaud & Michaud (2008) recommends the use of bootstap resampling. Lai, Xing & Chen (2011) reformulate the Markowitz problem as one of stochastic optimization with unknown moments. Goldfarb & Iyengar (2003) develop a robust optimization procedure for the Markowitz problem by embedding a factor structure in the constraint set.

[6]Preceeding work analyzed diffuse priors and was shown to be inneficient (Frost & Savarino 1986). The latter, instead, presumes all stocks are identical and have the same correlations. Vasicek (1973) specified a normal prior on the cross-sectional market betas (dominant factor).

mean) and Gillen (2014) (sample covariance). Formulae for Bayes' estimates of the return mean and covariance matrix based on normal and inverted Wishart priors may be found in Lai & Xing (2008, Chapter 4, Section 4.4.1).

A related approach to the Bayesian framework is that of shrinkage or regularization of the sample covariance matrix.[7] Shrinkage methods have been proposed in contexts where little underlying structure is present (Bickel & Levina 2008) as well as those in which a factor or other correlation structure is presumed to exist (e.g. Ledoit & Wolf (2003), Ledoit & Wolf (2004), Fan, Liao & Mincheva (2013) and Bun, Bouchaud & Potters (2016)). Perhaps surprisingly, shrinkage methods turn out to be related to placing constraints on the portfolio weights in the Markowitz optimization. Jagannathan & Ma (2003) show that imposing a positivity constraint typically shrinks the large entries of the sample covariance downward.[8]

As already mentioned, factor analysis and PCA in particular play a prominent role in the literature. It appears that while eigenvector bias is acknowledged, direct[9] bias corrections are made only to the eigenvalues corresponding to the principal components (e.g. Ledoit & Péché (2011) and Wang & Fan (2017)). Some work on characterizing the behavior of sample eigenvectors may be found in Paul (2007) and Shen, Shen, Zhu & Marron (2016). In the setting of Markowitz portfolios, the impact of eigenvalue bias and optimal corrections are investigated in in El Karoui et al. (2010) and El Karoui (2013).

Our approach also builds upon several profound contributions in the literature on portfolio composition. In an influential paper, Green & Hollifield (1992) observe the importance of the structure of the dominant factor to the composition of minimum variance portfolios. In particular, the "dispersion" of the dominant factor exposures drives the extreme positions in the portfolio composition. This dispersion is further amplified by estimation error, as pointed out in earlier work by Blume (1975) (see also Vasicek (1973)). These early efforts have led to a number of heuristics[10] to correct the sample bias of dominant factor estimates.

Contributions. We contribute to the literature by providing a method that significantly improves the performance of PCA-estimated minimum-variance portfolios. Our approach and perspective appear to be new. We summarize some of the main points.

Several authors (see above) have noted that sample eigenvectors carry a bias in the statistical and model setting we adopt. We contribute in this direction by, first, recognizing that it is the bias in the first sample eigenvector that

[7] In the Bayesian setup, the sample estimates are "shrunk" toward the prior (Lai & Xing 2008).

[8] This is generalized and analyzed further in DeMiguel, Garlappi, Nogales & Uppal (2009).

[9] Several approaches to alter the sample eigenvectors indirectly (e.g. shrinking the sample towards some structured covariance) do exist. However, the analysis of these approaches is not focused on characterizing the bias inherent to the sample eigenvectors themselves.

[10] For example, the Blume and Vasicek (beta) adjustments. See the discussion of Exhibit 3 and footnote 7 in Clarke, De Silva & Thorley (2011).

drives the performance of PCA-based, minimum-variance portfolios. Second, we show that this bias may in fact be corrected to some degree (cf. discussion below (3.7) in Wang & Fan (2017)). In our domain of application this degree is material. We point out that eigenvalue bias, which appears to be the focus in recent literature, does not have a material impact on minimum-variance portfolio performance. This motivates lines of research into more general Markowitz optimization problems. Finally, our correction can be framed geometrically in terms of the spherical law of cosines. This perspective illuminates possible extensions of our work. We discuss this further in our concluding remarks.

We also develop a bias correction and show that it outperforms standard PCA. Minimum variance portfolios constructed with our corrected covariance matrix are materially closer to optimal, and their risk forecasts are materially more accurate. In an idealized one-factor setting, we provide theoretical guarantees for the size of the improvement. Our theory also identifies some limitations. We demonstrate the efficacy of the method with an entirely data-driven correction. In an empirically calibrated simulation, its performance is far closer to the theoretically optimal than to standard PCA.

Organization. The article proceeds as follows: in Section 2 we describe in detail the problem and fundamental results around the sample covariance matrix and PCA. In Section 3, we present our main results on producing a bias corrected covariance estimate. Finally, in Section 4 we present numerical results illustrating the performance of our method in improving the estimated portfolio and risk forecasts.

2 Model specification and estimation error

2.1 Problem formulation

We consider a one-factor, linear model for returns to $N \in \mathbb{N}$ securities. In this setting, a generating process for the N-vector of exess returns R takes the form

$$R = \phi\beta + \epsilon \tag{1}$$

where ϕ is the return to the factor, $\beta = (\beta_1, \dots, \beta_N)$ is the vector of factor exposures, and $\epsilon = (\epsilon^1, \dots, \epsilon^N)$ is the vector of diversifiable specific returns. While the returns $(\phi, \epsilon) \in \mathbb{R} \times \mathbb{R}^N$ are random variables, we treat each exposure $\beta_n \in \mathbb{R}$ as a constant to be estimated. Assuming ϕ and the $\{\epsilon_n\}$ are mean zero and pairwise uncorrelated, the $N \times N$ covariance matrix of R is expressed as

$$\mathbf{\Sigma} = \sigma^2 \beta\beta^\top + \mathbf{\Delta}\,. \tag{2}$$

Here, σ^2 is the variance of ϕ and $\mathbf{\Delta}$ is a diagonal matrix with nth entry $\mathbf{\Delta}_{nn} = \delta_n^2$, the variance of ϵ_n. Estimation of $\mathbf{\Sigma}$ is central to numerous applications.

We consider a setting with $T \ll N$ observations $\{R_t\}_{t=1}^T$ of the vector R, generated by a latent time-series $\{\phi_t, \epsilon_t\}_{t=1}^T$ of (ϕ, ϵ). The usual assumption is that the observations are i.i.d. and we assume it throughout. Sample error distorts estimates of the model parameters $(\sigma^2, \beta, \delta^2)$ leading to the estimate,

$$\widehat{\mathbf{\Sigma}} = \hat{\sigma}^2 \hat{\beta} \hat{\beta}^\top + \widehat{\mathbf{\Delta}} \tag{3}$$

which approximates (2) by employing an estimated model $(\hat{\sigma}^2, \hat{\beta}, \hat{\delta}^2)$.

We focus on the impact of estimation error on Markowitz portfolio optimization[11]. In particular, given an estimated covariance matrix $\widehat{\mathbf{\Sigma}}$, we consider the estimated minimum variance portfolio, $\widehat{w}$, which is the solution to,

$$\begin{aligned} &\min_{w \in \mathbb{R}^N} w^\top \widehat{\mathbf{\Sigma}} w \\ &w^\top 1_N = 1. \end{aligned} \tag{4}$$

The weights, $\widehat{w}$, are extremely sensitive to errors in the estimated model, and risk forecasts for the optimized portfolio tend to be too low. We aim to address these issues in a high-dimension, low sample-size setting, $T \ll N$.

We are also interested in the equally weigthed portfolio where $w = \frac{1}{N} 1_N$. It is a very simple non-optimized portfolio, which we use to test whether corrections we make for improving the minimum variance portfolio are not offset by degraded performance elsewhere.

2.2 PCA and the $T \ll N$ regime

We consider principal component analysis (PCA) as the starting point for our analysis. PCA's use for risk factor identification is widespread in the literature, and it is appropriate when σ^2 is much larger than $\|\mathbf{\Delta}\|$. Without loss of generality, we take $\|\beta\|_2 = 1$.

Forming a data matrix $\mathbf{R} = (R_1, \ldots, R_T)$, we denote the data covariance matrix by $\mathbf{S} = \mathbf{R}\mathbf{R}^\top / T$. PCA identifies $\hat{\sigma}^2$ with the largest eigenvalue of $\mathbf{S}$ and $\hat{\beta}$ with the corresponding eigenvector. The diagonal matrix of specific risks is estimated as $\widehat{\mathbf{\Delta}} = \mathbf{diag}(\mathbf{S} - \hat{\sigma}^2 \hat{\beta} \hat{\beta}^\top)$ corresponding to least-squares regression of $\mathbf{R}$ onto the estimated factors. The estimate $\widehat{\mathbf{\Sigma}}$ of the covariance $\mathbf{\Sigma}$ is assembled as in (3).

Bias in this basic PCA estimator above arises from the use of the sample[12] covariance matrix $\mathbf{S}$. Many variants of the estimator exist to correct this bias at various scales of N/T. We focus on the T fixed and $N \uparrow \infty$ regime, which is appropriate for the type of applications we consider (see Section 1). This

[11] The standard Markowitz portfolio optimization problem is given by $\max_w w^T \mathbf{\Sigma} w$ subject to $w^T 1_N = 1$ and $w^T \mu \geq r$ where μ are the expected returns and r is the target portfolio return.

[12] If $\mathbf{\Sigma}$ replaces $\mathbf{S}$, sample bias vanishes and the estimator is asymptotically exact as $N \uparrow \infty$.

asymptotic regime has been analyzed in recent work. We summarize some results below.

To characterize the error in the estimated model $(\hat{\sigma}^2, \hat{\beta}, \hat{\delta}^2)$ asymptotically, we let $\hat{\theta}_{\beta,\hat{\beta}}$ denote the angle between β and the PCA-estimate $\hat{\beta}$ and $\gamma_{\beta,\hat{\beta}} = \cos \hat{\theta}_{\beta,\hat{\beta}}$. To highlight the dependence on N, we write $\hat{\sigma}^2 = \hat{\sigma}^2_N$ and $\hat{\theta} = \hat{\theta}_N$. Under general factor models, $\sigma^2_N/N \to \sigma^2_\infty$ and $\{\delta^2_n\}_{n\in\mathbb{N}}$ are bounded. In this context, (Shen et al. 2016) show that there is a non-degenerate positive random variable Ψ such that

$$\gamma_{\beta,\hat{\beta}} \xrightarrow{a.s.} \Psi^{-1} \tag{5}$$

$$\frac{\hat{\sigma}^2_N}{\sigma^2_N} \xrightarrow{a.s.} \xi\Psi^2 \tag{6}$$

as $N \to \infty$ for T fixed where ξ and Ψ are random variables dependent on T and $\Psi > 1$. We will elaborate more on these in Section 3. The inflation factor Ψ creates an upward bias for eigenvalue estimation and also defines a cone near which the sample eigenvector will lie with high probability, while ξ creates a random fluctuation in the eigenvalue inflation. Under normality, $\xi = \chi^2_T/T$. (Wang & Fan 2017) gives some results on the asymptotic distribution of the sample eigenvector on the cone. These limits fully characterize the error in the estimated model asymptotically. For $T \to \infty$, it follows that $\xi, \Psi \to 1$, indicating that for the factor model with growing eigenvalues in the dimension N, the sample eigenvector and eigenvalue are consistent.

Much has been studied in the random matrix theory literature on how to correct eigenvalues to improve the performance of the portfolio constructed via (4). In the next section, we introduce metrics to evaluate estimation error in portfolios, and we show how errors in estimates of factor exposures and factor volatility contribute to these metrics for a minimum variance portfolio.

2.3 Errors in optimized portfolios

Estimation error causes two types of difficulties in optimized portfolios. First, estimation error distorts portfolio weights. Second, estimation error generally causes the risk of optimized portfolios to be biased downward. Both effects are present for a minimum variance portfolio (4) constructed with the PCA-estimate $\widehat{\mathbf{\Sigma}}$ of Section 2.2. We now define the metrics for assessing the magnitude of these two errors.

In Section 2.1, we introduced the estimated minimum variance portfolio, $\widehat{w}$ as the solution to (4). We denote by w the optimal portfolio, i.e. the solution of (4) with $\widehat{\mathbf{\Sigma}} = \mathbf{\Sigma}$. Since $\mathbf{\Sigma}$ is positive definite, the optimal portfolio weights $\{w_n\}_{n=1}^N$ may be given explicitly.[13]

[13]Indeed, the Sherman- Morrison-Woodbury formula yields an explicit solution even for a

We define,

$$\mathcal{T}_{\widehat{w}}^2 = (w - \widehat{w})^\top \mathbf{\Sigma} \, (w - \widehat{w}) \, , \tag{7}$$

the (squared) tracking error of $\widehat{w}$. $\mathcal{T}_{\widehat{w}}^2$ measures the distance between the optimal and estimated portfolios, w and $\widehat{w}$. Specifically, it is the square of the width of the distribution of return differences $w - \widehat{w}$.

The variance of portfolio $\widehat{w}$ is given by $\widehat{w}^\top \widehat{\mathbf{\Sigma}} \, \widehat{w}$ and its true variance is $\widehat{w}^\top \mathbf{\Sigma} \, \widehat{w}$. We define,

$$\mathcal{R}_{\widehat{w}} = \frac{\widehat{w}^\top \widehat{\mathbf{\Sigma}} \, \widehat{w}}{\widehat{w}^\top \mathbf{\Sigma} \, \widehat{w}} \, , \tag{8}$$

the variance forecast ratio. Ratio (8) is less than one when the risk of the portfolio $\widehat{w}$ is underforecast. With respect to the equally weighted portfolio, we only consider the variance forecast ratio.

Metrics (7) and (8) quantify the errors in portfolio weights and risk forecasts induced by estimation error.[14] We state the following result concerning PCA bias, given a regularity assumption on the $\{\beta_n\}_{n \in \mathbb{N}}$ and for the homogeneous case where $\widehat{\mathbf{\Delta}} = \hat{\delta}^2 \mathbf{I}$, i.e., a scalar matrix. The more general heteregeneous risk case has minor changes involving weighted inner products with respect to $\widehat{\mathbf{\Delta}}^{-1}$.

Define $z = \mathbf{1}_N / \sqrt{N}$ and the angle cosines $\gamma_{\hat{\beta},\beta} := \cos \theta_{\hat{\beta},\beta} = \hat{\beta}^T \beta$, $\gamma_{\beta,z} := \cos \theta_{\beta,z} = \beta^T z$, and $\gamma_{\hat{\beta},z} := \cos \theta_{\hat{\beta},z} = \hat{\beta}^T z$. We note that since β, $\hat{\beta}$, and z are all growing in dimension as $N \to \infty$, the angle cosine quantities are all sequences in N. We do not require the deterministic quantity $\gamma_{\beta,z}$ to converge to anything since all results will be stated in terms of asymptotic equivalence but we typically take $\gamma_{\beta,z}$ to be a fixed value regardless of N.

Theorem 2.1 (Estimator and PCA Bias)**.** *For any estimate* $(\hat{\sigma}^2, \hat{\beta}, \hat{\delta}^2)$*, as* $N \to \infty$ *for* T *fixed, the estimate has a squared tracking error satisfying,*

$$\mathcal{T}_{\widehat{w}}^2 \overset{a.s.}{\sim} \frac{\sigma_N^2}{N} \widehat{m}_{\hat{\beta}}^2 (r_{\hat{\beta}} - \gamma_{\beta,\hat{\beta}})^2 + \frac{1}{N} \delta^2 \frac{\gamma_{\hat{\beta},z}^2 - \gamma_{\beta,z}^2}{(1 - \gamma_{\hat{\beta},z}^2)(1 - \gamma_{\beta,z}^2)} \tag{9}$$

---

multi- factor model and with a guaranteed mean return constraint (El Karoui). In our setting,

$$w_n \propto \frac{1}{\delta_n^2} (\beta_{\mathrm{MV}} - \beta_n) \, ; \qquad \beta_{\mathrm{MV}} = \frac{1 + \sigma^2 \sum_{\ell=1}^N \beta_\ell^2 / \delta_\ell^2}{\sigma^2 \sum_{\ell=1}^N \beta_\ell / \delta_\ell^2} \, .$$

[14] For a relationship to more standard error norms see (Wang & Fan 2017).

and overall satisfying,

$$\mathscr{R}_{\widehat{w}} \overset{a.s.}{\sim} \frac{\hat{\delta}^2}{\sigma_N^2 \widehat{m}_{\hat{\beta}} (r_{\hat{\beta}} - \gamma_{\beta,\hat{\beta}})^2 + \delta^2}, \tag{10}$$

where

$$\widehat{m}_{\hat{\beta}} = \frac{\gamma_{\hat{\beta},z}}{1 - \gamma_{\hat{\beta},z}^2}, \quad r_{\hat{\beta}} = \frac{\gamma_{\beta,z}}{\gamma_{\hat{\beta},z}}. \tag{11}$$

For the PCA estimator of Section 2.2, the asymptotic limits of tracking error and variance forecast ratios are given by

$$\mathscr{T}_{\widehat{w}}^2 \overset{a.s.}{\sim} \frac{\sigma_N^2}{N} \frac{\Psi^{-2}\gamma_{\beta,z}^2}{\left(1 - \Psi^{-2}\gamma_{\beta,z}^2\right)^2} \left(\Psi - \Psi^{-1}\right), \quad \mathscr{R}_{\widehat{w}} \overset{a.s.}{\to} 0, \tag{12}$$

where Ψ *is the random variable found in* (5) *and* (6).

The result states that the variance forecast ratio of the portfolio from the PCA estimator is asymptotically 0, since the estimated risk will be increasingly over-optimistic as $N \to \infty$, and this is entirely due to estimation error between the sample eigenvector and the population eigenvector. As N grows, the forecast risk becomes negligible relative to the true risk, rendering the PCA estimated minimum variance portfolio worse and worse. The tracking error is also driven by the error of the sample eigenvector and for increasing dimension, its proximity to the true minimum variance portfolio as measured by tracking error is asymptotically bounded below.

The key drivers of the error are the terms $r_{\hat{\beta}} - \gamma_{\beta,\hat{\beta}}$ and $\gamma_{\hat{\beta},z}^2 - \gamma_{\beta,z}^2$. Theorem 2.1 leads to a follow-up proposition.

Proposition 2.2. *Define the space,*

$$\mathcal{S}_\beta = \{x : 0 = r_x - \gamma_{\beta,x}\}. \tag{13}$$

If $\hat{\beta} \in \mathcal{S}_\beta$*,*

$$\mathscr{T}_{\widehat{w}}^2 \overset{a.s.}{\sim} \frac{1}{N}\delta^2 \frac{\gamma_{\hat{\beta},z}^2 - \gamma_{\beta,z}^2}{(1 - \gamma_{\hat{\beta},z}^2)(1 - \gamma_{\beta,z}^2)}, \quad \mathscr{R}_{\widehat{w}} \overset{a.s.}{\sim} \frac{\hat{\delta}^2}{\delta^2}, \tag{14}$$

The proposition shows that there exists a null space, $\mathcal{S}_\beta$, where true factor risk with respect to β of the estimated minimum variance portfolio, expressed as $\sigma^2(\widehat{w}^T\beta)^2$, is diversifiable. That is, only for an estimator $\hat{\beta} \in \mathcal{S}_\beta$ will true factor risk of the estimated minimum variance satisfy $\sigma^2(\widehat{w}^T\beta)^2 \to 0$ as $N \to \infty$.

Since $\gamma_{\hat{\beta},z} \geq \gamma_{\beta,z}$ is guaranteed on $\mathcal{S}_\beta$, the term $\gamma^2_{\hat{\beta},z} - \gamma^2_{\beta,z}$ in (9) indicates our objective is to minimize $\gamma_{\hat{\beta},z}$ on $\mathcal{S}_\beta$. The true vector of factor exposures β represent an optimal choice from $\mathcal{S}_\beta$ since that gives exact recovery of the factor exposures. More practically, for a set of *bona fide* estimators represented by $\mathcal{E}$, our task for achieving a non-degenerate variance ratio and reduced tracking error can be expressed as an optimization problem of the form,

$$\underset{x \in \mathcal{E}, \|x\|_2 = 1}{\text{minimize}} \quad \gamma_{x,z} + \mu d(x, \mathcal{S}_\beta) \tag{15}$$

where $d(x, \mathcal{S}_\beta)$ is a measure of distance from x to the space $\mathcal{S}_\beta$ and μ is a parameter governing the trade-off between proximity to $\mathcal{S}_\beta$ and $\gamma_{x,z}$.

In the following, we introduce a correction to the eigenvector $\hat{\beta}$ with the aim of minimizing $d(x, \mathcal{S}_\beta)$ to 0.

3 Bias correction for the one-factor model

We give a mathematical argument showing that in a PCA estimate of a one-factor model, estimation error tends to increase the dispersion in a dominant factor. The intuition behind this result comes from a simple example. Suppose the vector of exposures to the market factor, β, is the constant N-vector, $z = 1_N/\sqrt{N}$, whose entries are all equal to $1/\sqrt{N}$. Then the dispersion of the entries of β is 0.With probability 1, the PCA estimator, $\hat{\beta}$, will have higher dispersion than β when $\beta = z$. In practice, β is often close enough to $1_N/\sqrt{N}$ so the probability that $\hat{\beta}$ will have lower dispersion than β is small. The overdispersion of the entries of $\hat{\beta}$ manifests itself in the quantity $\gamma_{\hat{\beta},z}$, which asymptotically is almost surely less than $\gamma_{\beta,z}$.

Our main objective is to offer a correction to the first sample eigenvector $\hat{\beta}$ that improves the estimated model so that squared tracking errors of optimized portfolios are as low as they can be, and their variance forecast ratios are as close as possible to 1. As can be seen from Proposition 2.2, if we produce an estimate $\hat{\beta} \in \mathcal{S}_\beta$, we can greatly improve portfolio metrics. In particular, dominant terms that drive the poor behavior of the variance forecast ratio and tracking error are removed and lead to stable convergence to more ideal quantities. It can be readily seen from the Spherical Law of Cosines[15] that this objective is equivalent to maximizing $\gamma_{\beta,\hat{\beta}} = \cos\hat{\theta}_{\beta,\hat{\beta}}$, the cosine of the angle between β and $\hat{\beta}$ from the choice of estimators we consider.

[15] https://en.wikipedia.org/wiki/Spherical_law_of_cosines

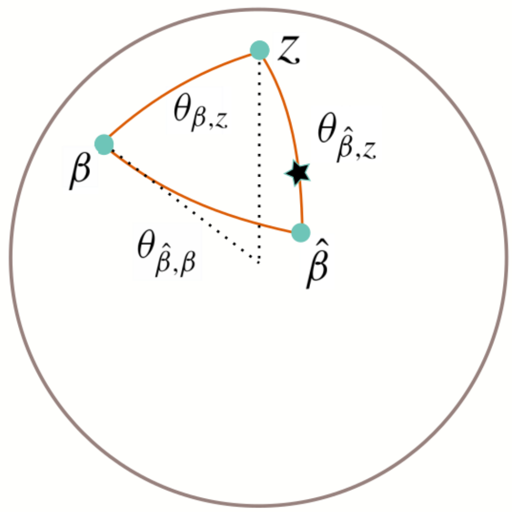

Figure 1: The true and PCA-estimated exposure vectors, β and $\hat{\beta}$, along with the dispersionless vector, z, are marked on the unit sphere. For large N and small T, with high probability, it is beneficial to replace $\hat{\beta}$ with $\hat{\beta}_*$ along the geodesic between $\hat{\beta}$ and z prescribed by (16). A plausible replacement vector $\hat{\beta}_*$ that leads to an improved estimate is marked with a star.

3.1 Oracle estimator

Since $\hat{\beta}$ lies on an $N-1$ sphere and $\mathcal{S}_\beta$ defines an $N-2$ dimensional space, we have too many potential directions along which we can move to try to produce an improvement. We consider an estimator along a geodesic between $\hat{\beta}$ and z. It is given in the following definition and illustrated in Figure 1.

Definition 3.1. *Define the set of estimators $\hat{\beta}_\rho$ under consideration as by the geodesic curve given by the formula,*

$$\hat{\beta}_\rho := \hat{\beta}(\rho) = \frac{\hat{\beta} + \rho z}{\|\hat{\beta} + \rho z\|_2}. \tag{16}$$

High-dimensional asymptotic results for factor models rest on a few basic assumptions for spiked covariance models that naturally arise from our factor specification. General results from Shen et al. (2016) and Wang & Fan (2017) utilize some strict ordering and boundedness of the top eigenvalue and the remaining eigenvalues. These results also require linear growth of the top eigenvalue as the dimension grows as well as convergence of the average of the remaining eigenvalues. We state them expressed in terms of the factor and specific risk quantities σ^2 and $\mathbf{\Delta}$ and refer the reader to Shen et al. (2016) and Wang & Fan (2017) for the technical assumptions as they appear in the spike model literature.

Assumption 3.2. *The factor variance $\sigma^2 = \sigma_N^2$ satisfies $\frac{N}{T\sigma_N^2} \to c_1$ as $N \to \infty$ where $0 < c_1 < \infty$. Specific risk is homogeneous such that $\mathbf{\Delta} = \delta^2 \mathbf{I}$ and bounded with $\delta^2 \in [c_0, C_0]$ for $c_0, C_0 > 0$.*

Assumption 3.3. *The returns are distributed as $\mathbf{R}_t \sim \mathcal{N}(0, \mathbf{\Sigma})$.*

We further assume a particular orientation with respect to z given below.

Assumption 3.4. *The vector β never coincides with z, i.e. $\gamma_{\beta,z} < 1$, and population and sample eigenvectors β and $\hat{\beta}$ are orientated such that $\gamma_{\beta,z}, \gamma_{\hat{\beta},z} \geq 0$.*

The first assumption provides an exploding eigenvalue as well as boundedness of the specific variance making inference in the high dimensional limit feasible. The constants c and c_1 depend on T though as throughout we leave the dependence implicit since we consider T fixed. The distribution assumption is mainly for providing a sufficient condition to facilitate proofs. The last assumption is not much of a restriction. The reasoning is simple: either we do not actually care about the orientation and therefore our bias correction approach will seek a vector that is "close" but possibly with the wrong orientation. Or, we do care and we have enough information to know that $\gamma_{\hat{\beta},z} > 0$ is the correct orientation.

As noted above, we can improve tracking error and variance forecast ratio by reducing the angle between the estimated eigenvector and the true underlying eigenvector, β, equivalently, by replacing $\hat{\beta}$ with an appropriate choice of $\hat{\beta}_\rho$. We find an optimal value ρ_N^* for a particular N. We present our method and its impact on tracking error and variance forecast ratio for a minimum variance portfolio below. We restrict to the case of homogeneous specific risk where $\mathbf{\Delta} = \delta^2 \mathbf{I}$ for expositional purposes but consider the full- fledged case in empirical results in Section 4.

In the following theorem, we also provide a correction to the sample eigenvalue. Our bias correction for the sample eigenvector introduces a bias in the variance forecast ratio for the equally weighted portfolio. We shrink the sample eigenvalue, treated as the variance of the estimated factor, to debias the variance forecast ratio for the equally weighted portfolio.

Theorem 3.5. *Under 3.2, 3.3, 3.4, let $\mathbf{R} = (R_1, \ldots, R_T)$ be a set of T independent observations from a normal distribution with covariance matrix . Let $\widehat{\mathbf{\Delta}} = \hat{\delta}^2 \mathbf{I}$ where $\hat{\delta}^2$ is an arbitrary estimator of specific risk.*

Define the oracle corrected estimate as $\hat{\beta}_{\rho^*} := \hat{\beta}_{\rho_N^*}$ where the finite sample (fixed N and T) optimal value ρ_N^* solves $0 = r_{\hat{\beta}_\rho} - \gamma_{\beta,\hat{\beta}_\rho}$ or equivalently maximizes $\gamma_{\beta,\hat{\beta}_\rho}$.

i. The finite sample optimal oracle ρ_N^* is given by,

$$\rho_N^* = \frac{\gamma_{\beta,z} - \gamma_{\beta,\hat{\beta}}\gamma_{\hat{\beta},z}}{\gamma_{\beta,\hat{\beta}} - \gamma_{\beta,z}\gamma_{\hat{\beta},z}}. \tag{17}$$

ii. *For T fixed, $N \to \infty$, we have,*

$$\rho_N^* \overset{a.s.}{\sim} \bar{\rho}_N = \frac{\gamma_{\beta,z}}{1-\gamma_{\beta,z}^2}\left(\Psi - \Psi^{-1}\right). \tag{18}$$

where $\Psi^2 = 1+\delta^2 c_1/\xi$, $\xi = \chi_T^2/T$, and χ_T^2 is the chi-squared distribution with T degrees of freedom. Also, $\bar{\rho}_N > 0$ almost surely if $\gamma_{\beta,z} > 0$. And the asymptotic improvement of the optimal angles $\theta_{\beta,\hat{\beta}_{\rho^*}}$ and $\theta_{\beta,\hat{\beta}_{\bar{\rho}_N}}$ over the original angle $\theta_{\beta,\hat{\beta}}$ as $N \to \infty$ is,

$$\frac{\sin^2\theta_{\beta,\hat{\beta}_{\rho_N^*}}}{\sin^2\theta_{\beta,\hat{\beta}}} \overset{a.s.}{\sim} \frac{1-\gamma_{\beta,z}^2}{1-(\frac{\gamma_{\beta,z}}{\Psi})^2} = \frac{\sin^2\theta_{\beta,\hat{\beta}_{\bar{\rho}_N}}}{\sin^2\theta_{\beta,\hat{\beta}}}.$$

iii. *For the oracle value ρ_N^*, the tracking error and forecast variance ratio for the minimum variance portfolio for $\widehat{\mathbf{\Sigma}}_{\rho_N^*} = \hat{\sigma}^2\hat{\beta}_{\rho^*}\hat{\beta}_{\rho^*}^T + \hat{\delta}^2\mathbf{I}$ satisfy,*

$$\mathscr{T}_{\widehat{w}}^2 \overset{a.s.}{\sim} \frac{1}{N}\delta^2\frac{\gamma_{\hat{\beta}_{\rho^*},z}^2 - \gamma_{\beta,z}^2}{(1-\gamma_{\hat{\beta}_{\rho^*},z}^2)(1-\gamma_{\beta,z}^2)} \tag{19}$$

$$\mathscr{R}_{\widehat{w}} \overset{a.s.}{\sim} \hat{\delta}^2/\delta^2. \tag{20}$$

That is, after the optimal correction, the forecast variance ratio for the minimum variance portfolio no longer converges to 0 while the tracking error to the true minimum variance portfolio does.

iv. *Define the corrected eigenvalue as,*

$$\hat{\sigma}_{\rho_N^*}^2 = \Phi_{\rho_N^*}^2\hat{\sigma}^2, \quad \Phi_{\rho_N^*} = \frac{\gamma_{\hat{\beta},z}}{\gamma_{\hat{\beta}_{\rho^*},z}}$$

then for the equally weighted portfolio $w = \frac{1}{N}1_N$, the forecast variance ratio also satisfies $\mathscr{R}_w \overset{a.s.}{\sim} \xi$ as $N \to \infty$.

Remark 3.6. *Geometrically, there are two views of $\hat{\beta}_{\rho^*}$. One is that $\hat{\beta}_{\rho^*}$ is the projection of $\hat{\beta}$ onto $\mathcal{S}_\beta$. The other is that $\hat{\beta}_{\rho^*}$ is the projection of β onto the geodesic defined by $\hat{\beta}_\rho$. In either case, our goal is to find the intersection of the geodesic and the space $\mathcal{S}_\beta$.*

Remark 3.7. *While we consider a specific target, z, in principle the target does not actually matter. It is possible that these kind of factor corrections can be applied beyond the first factor, given enough information to create a reasonable prior.*

The first takeaway from this result is that in the high dimensional limit, it is always possible to improve on the PCA estimate by moving along the geodesic between $\hat{\beta}$ and z. As $\gamma_{\beta,z}$ approaches 0 or for a larger T, the optimal correction approaches 0. Conversely, as $\gamma_{\beta,z}$ approaches 0 or for smaller T, the magnitude of the correction is larger. For $\gamma_{\beta,z} = 1$, the proper choice is naturally to choose z since β and z are aligned in that case.

The improvement in the angle as measured by the ratio of squared sines is bounded in the interval $(1 - \gamma^2_{\beta,z}, 1)$. As $\gamma_{\beta,z}$ approaches 0 or for larger T, the improvement diminishes and the ratio approaches 1. Conversely, for large values of c_1, the improvement approaches $1 - \gamma^2_{\beta,z}$, indicating that improvement is naturally constrained by how close β is to z in the first place.

In the application to the minimum variance portfolio, the initial idea is to correct the sample eigenvector so that we reduce the angle to the population eigenvector. However, it is not immediately clear that this should have a dramatic effect. Even more surprising is that underestimation of risk has a large component due to sample eigenvector bias and not any sample eigenvalue bias. While an improved estimate $\hat{\beta}_\rho$ has the potential to greatly improve forecast risk, this represents only a single dimension on which to evaluate the performance of a portfolio. We could be sacrificing small tracking error to the true long-short minimum variance portfolio in exchange for better forecasting. That however is not the case here.

Furthermore, we also offer a correction to the sample eigenvalue that preserves risk estimation performance with respect to the equally weighted portfolio $w = \frac{1}{N} 1_N$. While we have not considered all non-optimized portfolios, our eigenvalue correction provides the desired results for the equally weighted portfolio.

Since Ψ and ξ are unobservable non-degenerate random variables, determining their realized values, even with asymptotic estimates, is an impossible task. Hence perfect corrections to kill off the driving term of underestimation of risk are not possible. However, it is possible to make corrections that materially improve risk forecasts.

3.2 Data-driven estimators

3.2.1 Homogenous specific variance

We introduce procedure for constructing an estimator $\hat{\rho}$ for the asymptotic oracle correction parameter $\bar{\rho}_N$ given in (18). It is based on estimates the specific variance δ^2 and c_1 from Assumption 3.2. From Yata & Aoshima (2012), we have the estimator given by,

$$\hat{\delta}^2 = \frac{\text{Tr}(\mathbf{S}) - \hat{\lambda}_1}{N - 1 - \frac{N}{T}}, \tag{21}$$

where $\mathbf{S}$ is the sample covariance matrix for the data matrix $\mathbf{R}$ and $\hat{\lambda}_1 = \hat{\sigma}^2$ is the first eigenvalue of $\mathbf{S}$. A natural estimate for the true eigenvalue $\lambda_1()$ is $\widehat{\lambda}_1^S = \max\{\hat{\lambda}_1 - \hat{\delta}^2 N/T, 0\}$. For $\hat{\lambda}_1$ sufficiently large, the estimate of c_1 is given by,

$$\hat{c}_1 = \frac{N}{T\hat{\lambda}_1 - N\hat{\delta}^2}. \tag{22}$$

Given the estimates of δ^2 and c_1, we need a precise value for ξ, as well as ξ, in order to have a bona fide estimator. We approximate ξ by its expectation $\mathbb{E}[\xi] = 1$ to obtain a completely data driven correction parameter estimate $\hat{\rho}$,

$$\hat{\rho} = \frac{\widehat{\Psi}\gamma_{\hat{\beta},z}}{1 - (\widehat{\Psi}\gamma_{\hat{\beta},z})^2}\left(\widehat{\Psi} - \widehat{\Psi}^{-1}\right), \tag{23}$$

where $\widehat{\Psi}^2 = 1 + \hat{\delta}^2\hat{c}_1$.

We compute the factor variance as,

$$\hat{\sigma}^2_{\hat{\rho}} = \Phi^2_{\hat{\rho}}\hat{\sigma}^2, \quad \Phi_{\hat{\rho}} = \frac{\gamma_{\hat{\beta},z}}{\gamma_{\hat{\beta}_{\hat{\rho}},z}},$$

where $\hat{\sigma}^2 = \hat{\lambda}_1$ is the first sample eigenvalue from $\mathbf{S}$.

3.2.2 Heterogenous specific variance

Our analysis thus far has rested on the simplifying assumption that security return specific variances have a common value. Empirically, this is not the case, and the numerical experiments discussed below in Section 4 allow for the more complex and realistic case of heterogenous specific variances. To address the issue, we modify both the oracle estimator and the bona fide estimator by rescaling betas by specific variance.

Under heterogeneous specific variance, the oracle value ρ_N^* is given by the formula,

$$\rho_N^* = \frac{\gamma^{\widehat{\mathbf{\Delta}}}_{\beta,z} - \gamma^{\widehat{\mathbf{\Delta}}}_{\beta,\hat{\beta}}\gamma^{\widehat{\mathbf{\Delta}}}_{\hat{\beta},z}}{\gamma^{\widehat{\mathbf{\Delta}}}_{\beta,\hat{\beta}} - \gamma^{\widehat{\mathbf{\Delta}}}_{\beta,z}\gamma^{\widehat{\mathbf{\Delta}}}_{\hat{\beta},z}},$$

where $\gamma^{\widehat{\mathbf{\Delta}}}_{x,y} = x^T\widehat{\mathbf{\Delta}}^{-1}y$ is a weighted inner product. Furthermore, the risk adjusted returns $\mathbf{R}\mathbf{\Delta}^{-1/2}$ have covariance $\widetilde{\mathbf{\Sigma}}$ given by,

$$\widetilde{\mathbf{\Sigma}} = \sigma^2\mathbf{\Delta}^{-1/2}\beta\beta^\top\mathbf{\Delta}^{-1/2} + \mathbf{I}.$$

For the risk adjusted returns, Theorem 3.5 holds. The oracle formula coupled with the risk adjusted returns suggest we use $\widetilde{\mathbf{R}} = \mathbf{R}\widehat{\boldsymbol{\Delta}}^{-1/2}$ as the data matrix where $\widehat{\boldsymbol{\Delta}}$ is the specific risk estimate from the standard PCA method. Why should we expect this to work? The purpose of the scaling $\mathbf{R}\boldsymbol{\Delta}^{-1/2}$ is to make the specific return distribution isotropic and $\widetilde{\mathbf{R}}$ approximates that. Since we are only trying to obtain an estimator $\hat{\rho}$ that is close to ρ_N^*, this approximation ends up fine. And for ellipses specified by $\boldsymbol{\Delta}$ with relatively low eccentricity, the estimator in (23) actually works just fine since the distribution is relatively close to isotropic. So for larger eccentricity, we require the following adjustment just to get the data closer to an isotropic specific return distribution.

The updated formulas for the heterogenous specific risk correction estimators are given below, and we use them in our numerical experiments. For an initial estimate of specific risk $\widehat{\boldsymbol{\Delta}}$, the modified quantities are,

$$\tilde{\hat{\beta}} = \frac{\widehat{\boldsymbol{\Delta}}^{-1/2}\hat{\beta}}{\|\widehat{\boldsymbol{\Delta}}^{-1/2}\hat{\beta}\|_2}, \quad \tilde{z} = \frac{\widehat{\boldsymbol{\Delta}}^{-1/2}z}{\|\widehat{\boldsymbol{\Delta}}^{-1/2}z\|_2}, \tag{24}$$

$$\widetilde{\mathbf{S}} = \widehat{\boldsymbol{\Delta}}^{-1/2}\mathbf{S}\widehat{\boldsymbol{\Delta}}^{-1/2}, \quad \tilde{\hat{\lambda}}_1 = \hat{\lambda}_1\|\widehat{\boldsymbol{\Delta}}^{-1/2}\hat{\beta}\|^2, \tag{25}$$

$$\tilde{\hat{\delta}}^2 = \frac{\mathrm{Tr}(\widetilde{\mathbf{S}}) - \tilde{\hat{\lambda}}_1}{N - 1 - \frac{N}{T}}, \quad \hat{c}_1 = \frac{N}{T\tilde{\hat{\lambda}}_1 - N\tilde{\hat{\delta}}^2}, \tag{26}$$

$$\hat{\rho} = \frac{\widehat{\Psi}\gamma_{\tilde{\hat{\beta}},\tilde{z}}}{1 - (\widehat{\Psi}\gamma_{\tilde{\hat{\beta}},\tilde{z}})^2}\left(\widehat{\Psi} - \widehat{\Psi}^{-1}\right). \tag{27}$$

We use the PCA estimate of the specific risk $\widehat{\boldsymbol{\Delta}} = \mathbf{diag}(\mathbf{S} - \hat{\sigma}^2\hat{\beta}\hat{\beta}^\top)$ as the initial estimator.

Once we have the estimated $\hat{\rho}$, we return to the original data matrix $\mathbf{R}$ and apply the correction as before to $\hat{\beta}$, the first eigenvector of the sample covariance matrix. The method for correcting the sample eigenvalue remains the same and we opt to recompute the specific variances using the corrected factor exposures and variance. A full specification of the algorithm can be found in the appendix.

4 Numerical examples

The theoretical results in Sections 2 and 3 came out of our attempt to solve a practical problem. We were looking for the best PCA-based estimate of covariance matrix for the purposes of portfolio construction and risk forecasting. As documented in Theorems 2.1 and 3.5, estimation error leads to a dispersion bias in the dominant factor, and that bias has an optimal correction along the geodesic connecting the estimated factor to the unique dispersion-free vector on the sphere.

In this section, we provide simulation results quantifying the improvements in accuracy of portfolio weights and risk forecasts due to our bias mitigation. We consider two specific examples. First, we focus on a minimum variance portfolio because of its extreme sensitivity to estimation error in risk forecasts, its complete insensitivity to estimation error in expected return and its importance to investors. In simulation, our bias correction substantially diminishes the tracking error of a minimum variance portfolio and substantially raises its variance forecasting ratio toward 1—both desirable. Second, we focus on an equally weighted portfolio, which is also important to investors. Like many practical portfolios, an equally weighted portfolio is constructed without the use of optimization, and has substantial exposure to the market factor. Since the weights of an equally weighted portfolio are always correct, tracking error is irrelevant. Variance forecast ratio of an equally weighted portfolio, however, is important, and our bias correction has little discernible effect on it.

The first step in our simulation is to calibrate the one-factor model (1) used to generate returns in an empirically sound way, and the study in Clarke et al. (2011) serves as a rough guide. Our single factor is modeled on the market. In practice, the dispersion of market betas varies across different regimes, so we examine values of $\gamma_{\beta,z}$ ranging from 0.5 to 1.0.The parameter $\gamma_{\beta,z}$ controls the dispersion of the factor exposures where $\gamma_{\beta,z} = 0.9$ corresponds approximately to an exposure vector with mean 1 and cross-sectional variance of 0.25. Historically, the dispersion has been greater in turbulent markets than in calm markets, and that prompted us to vary $\gamma_{\beta,z}$ parameter in the experiments.

The factor variance corresponds to an annualized volatility of roughly 16%, which reflects the average behavior of the US equity market over the past half century. Finally, we draw annualized specific volatilities $\{\delta_n^2\}$ from a uniform distribution on $[10\%, 64\%]$.

In each experiment, we use our calibrated factor model to simulate a year's worth of daily returns, $T = 250$, to N securities. From this data set, we construct a sample covariance matrix, $\mathbf{S}$, and extract three estimators of the factor model covariance matrix $\mathbf{\Sigma}$. The first is extracted with standard PCA. Specifically our estimate of the factor exposures is the dominant eigenvector, $\hat{\beta}$, of $\mathbf{S}$, and our estimate, $\hat{\sigma}^2$, of factor variance is the dominant eigenvalue. Our estimates of specific variances, $\widehat{\mathbf{\Delta}}$, are sample variances of residual returns.

The second estimator is extracted in the same way except that we apply oracle shrinkage in (17) to the dominant eigenvector. The third applies data-driven shrinkage in (23) to the dominant eigenvector and eigenvalue. We use these three estimated covariance matrices to construct minimum variance portfolios and to forecast their risk. Finally, we use the covariance matrices to forecast the risk of an equally weighted portfolio.

We run 50 simulations for each model calibration, and we box plot the annualized tracking errors[16] and variance forecast ratios of the estimates.

[16]For ease of interpretation, we report tracking error (rather than its square) at an annual horizon.

Results are shown in Figure 2 for fixed dispersion, $\gamma_{\beta,z} = 0.9$ and N ranging from 500 to 3000 and in Figure 3 with fixed $N = 500$ and dispersion $\gamma_{\beta,z}$ ranging from 0.5 to 1.0. Each figure has three panels. The first and second are annualized tracking error and variance forecast ratio for a minimum variance portfolio, the third is variance forecast ratio for an equally weighted portfolio.

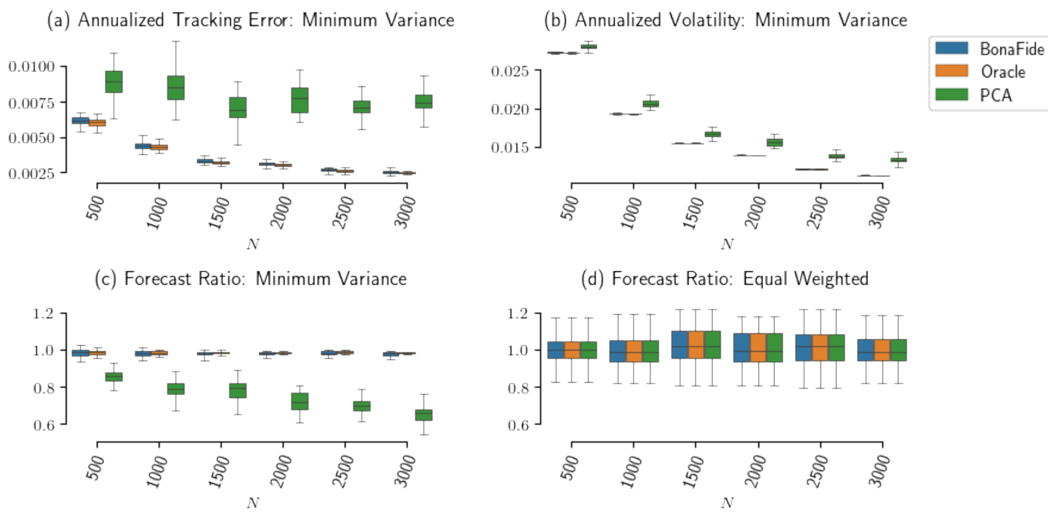

Figure 2: Performance statistic for 50 simulated data sets of $T = 250$ observations as N varies and $\gamma_{\beta,z} = 0.9$. Panel (a): Annualized tracking error for minimum variance portfolios. Panel (b): Annualized volatility for minimum variance portfolios. Panel (c): Variance forecast ratio for minimum variance portfolios. Panel (d): Variance forecast ratio for equally weighted portfolios.

Panel (a) in Figure 2 indicates that when the dispersion of the factor is set by $\gamma_{\beta,z} = 0.9$, the distance between the optimized and optimal minimum variance portfolios tends to decline as N grows in the prescribed range. The decline is more rapid and more consistent for the two corrected estimators. In Panel (b), we see that as N grows in the prescribed range, risk is underforecast with increasing severity for the PCA estimator, but is much more accurate for the two corrected estimators. Panel (c) shows that as N varies, the variance forecast ratio of an equally weighted portfolio is noisy for all three estimators, and there is no clear indication of bias for any of them.

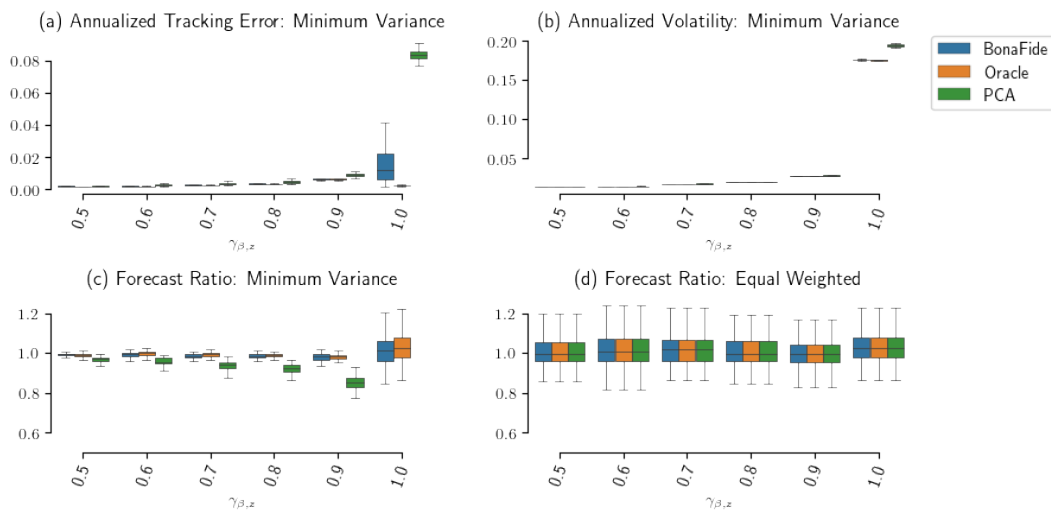

Figure 3: Performance statistic for 50 simulated data sets of $T = 250$ observations as $\gamma_{\beta,z}$ varies and $N = 500$. Panel (a): Annualized tracking error for minimum variance portfolios. Panel (b): Annualized volatility for minimum variance portfolios. Panel (c): Variance forecast ratio for minimum variance portfolios. Panel (d): Variance forecast ratio for equally weighted portfolios.

Panel (a) in Figure 3 indicates that for $N = 500$, the distance between the optimized and optimal minimum variance portfolios tends to increase as the dominant factor becomes more concentrated. The dramatic decline in performance at $\gamma_{\beta,z} = 1$ for both standard PCA and the data-driven correction reflects the sea change in the composition of a minimum variance portfolio that occurs when the dominant eigenvector is sufficiently concentrated. Specifically, the minimum variance portfolio tends to lose its short positions and become long-only. If specific variances are all equal, the minimum variance portfolio becomes equally weighted when the dominant eigenvector is dispersionless. Panel (b) suggests that it is easier to accurately forecast the risk of minimum variance portfolio when the dominant eigenvector is more dispersed. Panel (c) shows that at different levels of dispersion, the variance forecast ratio of an equally weighted portfolio is noisy for all three estimators, and there is no clear indication of bias for any of them.

5 Summary

In this article, we develop a correction for bias in PCA-based covariance matrix estimators. The bias is excess dispersion in a dominant eigenvector, and the form of the correction is suggested by formulas for estimation error metrics applied to minimum variance portfolios.

The intuition behind our result comes from a simple example, where the entries of the vector of exposures to the market are constant. In this situation, an estimate of the exposure vector necessarily has greater dispersion than the true exposure vector. Of course, the vector of exposures to the market are generally not constant. But the tendency for PCA-based estimates of non-constant market exposure vectors to have excess dispersion can be formalized in an idealized setting. For a fixed number of observations, T, with probability that tends to 1 as N tends to ∞, the dominant sample eigenvector will be overly dispersed as long as the mean factor entry is non-zero. We identify an oracle correction that optimally shrinks the sample eigenvector along a spherical geodesic toward the distinguished zero-dispersion vector, and we provide asymptotic guarantees that oracle shrinkage reduces both types of error. These findings are especially relevant to equity return covariance matrices, which feature a dominant factor whose overwhelmingly positive exposures tend to be overly dispersed by PCA estimators.

Our results fit into two streams of academic literature. The first is the large-N-fixed-T branch of random matrix theory, which belongs to statistics. The second is empirical finance, which features results about Bayesian adjustments to estimated betas.

To enable practitioners to use our results, we develop a data-driven estimator of the oracle. Simulation experiments support the practical value of our correction, but much work remains to be done. That includes the development of estimates of the size and likelihood of the exposure bias in finite samples, the identification and correction of biases in other risk factors, and empirical studies. Explicit formulas for error metrics in combination with the geometric perspective in this article provide a way to potentially improve construction and risk forecasts of investable optimized portfolios.

A Algorithm

Our corrected covariance matrix algorithm is given in Algorithm 1 where the input is a data matrix of returns R.

Algorithm 1 1-Factor Bias Corrected PCA Covariance Estimator

Require: $\mathbf{R} = (R_1, \ldots, R_T)$
Require: $z = 1_N/\sqrt{N}$

procedure BIAS CORRECTED PCA COVARIANCE($\mathbf{R}$)
  $\mathbf{S} \leftarrow \frac{1}{T}\mathbf{R}\mathbf{R}^T$
  $\hat{\mathbf{U}} \leftarrow [\hat{u}_1, \ldots, \hat{u}_N], \quad \hat{\ } \leftarrow \mathrm{Diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_N)$ ▷ Eigendecomposition of $\mathbf{S}$
  $\hat{\beta} \leftarrow \mathrm{sign}(\hat{u}_1^T z)\hat{u}_1$ ▷ Orient such that $\hat{\beta}^T z > 0$.
  $\hat{\sigma}^2 \leftarrow \hat{\lambda}_1$
  $\widehat{\mathbf{\Delta}} \leftarrow \mathrm{Diag}(\mathbf{S} - \widehat{\boldsymbol{L}}), \quad \widehat{\boldsymbol{L}} \leftarrow \hat{\sigma}^2 \hat{\beta}\hat{\beta}^T$ ▷ Initial PCA estimate.
  $\widetilde{\mathbf{S}} \leftarrow \widehat{\mathbf{\Delta}}^{-1/2}\mathbf{S}\widehat{\mathbf{\Delta}}^{-1/2}, \quad \tilde{\hat{\lambda}}_1 \leftarrow \hat{\lambda}_1 \|\widehat{\mathbf{\Delta}}^{-1/2}\hat{\beta}\|^2$
  $\tilde{\hat{\beta}} \leftarrow \frac{\widehat{\mathbf{\Delta}}^{-1/2}\hat{\beta}}{\|\widehat{\mathbf{\Delta}}^{-1/2}\hat{\beta}\|_2}, \quad \tilde{z} \leftarrow \frac{\widehat{\mathbf{\Delta}}^{-1/2}z}{\|\widehat{\mathbf{\Delta}}^{-1/2}z\|_2}$
  $\widehat{\Psi} \leftarrow 1 + \tilde{\hat{\delta}}^2 \hat{c}_1, \quad \tilde{\hat{\delta}}^2 \leftarrow \frac{\mathrm{Tr}(\widetilde{\mathbf{S}}) - \tilde{\hat{\lambda}}_1}{N - 1 - \frac{N}{T}}, \quad \hat{c}_1 \leftarrow \frac{N}{T\tilde{\hat{\lambda}}_1 - N\tilde{\hat{\delta}}^2}$
  $\hat{\beta}_{\hat{\rho}} \leftarrow \frac{\hat{\beta} + \hat{\rho} z}{\sqrt{1 + 2\hat{\rho}\gamma_{\hat{\beta},z} + \hat{\rho}^2}}, \quad \hat{\rho} \leftarrow \frac{\widehat{\Psi}\gamma_{\tilde{\hat{\beta}},\tilde{z}}}{1 - (\widehat{\Psi}\gamma_{\tilde{\hat{\beta}},\tilde{z}})^2}\left(\widehat{\Psi} - \widehat{\Psi}^{-1}\right)$
  $\widehat{\mathbf{\Delta}}_{\hat{\rho}} \leftarrow \mathrm{Diag}(\mathbf{S} - \widehat{\boldsymbol{L}}_{\hat{\rho}}), \quad \widehat{\boldsymbol{L}}_{\hat{\rho}} \leftarrow \hat{\sigma}^2_{\hat{\rho}} \hat{\beta}_{\hat{\rho}} \hat{\beta}^T_{\hat{\rho}}, \quad \hat{\sigma}^2_{\hat{\rho}} \leftarrow \frac{\gamma^2_{\hat{\beta},z}}{\gamma^2_{\hat{\beta}_{\hat{\rho}},z}} \hat{\sigma}^2$
  return $\hat{\mathbf{\Sigma}} \leftarrow \widehat{\boldsymbol{L}}_{\hat{\rho}} + \widehat{\mathbf{\Delta}}_{\hat{\rho}}$ ▷ The 1-factor bias corrected PCA estimator.
end procedure

B Proof of main results

We start off with some foundational asymptotic results from the literature. Let $\mathbf{X} \sim \mathcal{N}(0, \mathbf{\Lambda})$ where $\mathbf{\Lambda} = \mathbf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_2)$ is a diagonal matrix satisfying with λ_1 satisying Assumption $\mathcal{C}1$ in Shen et al. (2016). Let $\mathbf{S_X} = \frac{1}{T}\mathbf{X}\mathbf{X}^T$ be the sample covariance for $\mathbf{X}$ with eigendecomposition $\mathbf{S_X} = \hat{\mathbf{V}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}}^T$. Further let $\hat{v}_1$ be the first sample eigenvector given by,

$$\hat{v}_1 = \begin{bmatrix} \hat{v}_{11} \\ \ldots \\ \hat{v}_{N1} \end{bmatrix},$$

and define $\tilde{v} = \begin{bmatrix} \hat{v}_{21} & \dots & \hat{v}_{N1} \end{bmatrix}^T$. By Paul (2007, Theorem 6), $\tilde{v} \sim \text{Unif}(B(N-2))$ where $B(N-2)$ is a unit $N-2$ sphere.

Via a simple scaling by λ_2, by Shen et al. (2016, Theorem 6.3) we have $e_1^T \hat{v}_1 = \hat{v}_{11} \overset{a.s.}{\to} \Psi^{-1}$ where $\Psi^2 = 1 + \lambda_2 c_1/\xi$ and $\xi = \chi_T^2/T$.

We introduce the following lemma and corollary, which we will use in the proof of the two theorems. We leave the proofs until the end.

Lemma B.1. *If $X_n \sim \text{Unif}(B(n-1)) \in \mathbb{R}^n$ and $Y_n \in \mathbb{R}^n$ is a sequence independent of X_n such that $\|Y_n\|_2 = 1$, then $X_n^T Y_n \overset{a.s.}{\to} 0$ as $n \to \infty$.*

Corollary B.2. *For $\mathbf{R}$ where $\mathbf{R}_t \sim \mathcal{N}(0, \mathbf{\Sigma})$, $= \sigma^2 \beta\beta^\top + \delta^2 \mathbf{I}$, and $\hat{\beta}$ the first sample eigenvector of the sample covariance matrix $\mathbf{S}$, then,*

$$\gamma_{\hat{\beta},z} := \hat{\beta}^T z \overset{a.s.}{\sim} \Psi^{-1}\gamma_{\beta,z}, \quad \gamma_{\beta,\hat{\beta}} \overset{a.s.}{\to} \Psi^{-1}.$$

Proof of Theorem 2.1.
The asymptotic behaviors of $\mathcal{T}^2_{\widehat{w}}$, $\mathcal{R}_{\widehat{w}}$, and $\mathcal{R}^{\boldsymbol{L}}_{\widehat{w}}$ for general estimates $\hat{\sigma}^2_N$, $\hat{\beta}$, $\hat{\delta}^2$ are a direct result of Lemmas C.9 and C.11.

For the PCA estimation procedure, from Corollary B.2, we conclude that

$$\widehat{m}_{\hat{\beta}} \overset{a.s.}{\sim} \frac{\Psi^{-1}\gamma_{\beta,z}}{1 - (\Psi^{-1}\gamma_{\beta,z})^2}, \quad r_{\hat{\beta}} \overset{a.s.}{\to} \Psi.$$

Since $\Psi - \Psi^{-1} > 0$, we conclude by substitution above that

$$\mathcal{T}^2_{\widehat{w}} \overset{a.s.}{\sim} \frac{\sigma_N^2}{N} \frac{\Psi^{-2}\gamma^2_{\beta,z}}{\left(1 - \Psi^{-2}\gamma^2_{\beta,z}\right)^2} \left(\Psi - \Psi^{-1}\right)$$
$$\mathcal{R}_{\widehat{w}} \overset{a.s.}{\to} 0,$$

□

Proof of Proposition 2.2. The conclusion is clear from Corollaries C.10 and C.12.

□

Proof of Theorem 3.5.

i. It is easy to verify that ρ_N^* solves $0 = r_{\hat{\beta}_\rho} - \gamma_{\beta,\hat{\beta}_\rho}$ and maximizes $\gamma_{\beta,\hat{\beta}_\rho}$.

ii. Convergence of ρ_N^* to $\bar{\rho}_N$ stems from Corollary B.2. It is also clear that $\bar{\rho}_N$ if $\gamma_{\beta,z} > 0$ since $\Psi - \Psi^{-1} > 0$.

   For the asymptotic improvement due to shrinkage, we rely on Björck & Golub (1973, Theorem 1), which shows that principal angle can be derived from the singular value decomposition of,

$$\beta^T \begin{bmatrix} z & \frac{\hat{\beta}-\gamma_{\hat{\beta},z}z}{\sqrt{1-\gamma^2_{\hat{\beta},z}}} \end{bmatrix} = \begin{bmatrix} \gamma_{\beta,z} & \frac{\gamma_{\beta,\hat{\beta}}-\gamma_{\hat{\beta},z}\gamma_{\beta,z}}{\sqrt{1-\gamma^2_{\hat{\beta},z}}} \end{bmatrix}.$$

By maximizing $\gamma_{\beta,\hat{\beta}_\rho}$ (or equivalently minimizing the angle between β and β_ρ), we are directly choosing ρ_N^* such that $\hat{\beta}_{\rho_N^*}$ is the principal vector with corresponding principal angle to $\hat{\beta}$. Finding the vector in terms of correction quantity is easier through direct maximization of $\gamma_{\beta,\hat{\beta}_\rho}$, and finding the improvement is easier through the principal angle computation, despite the results being equivalent.

From the above product, the squared cosine of the principal angle and its asymptotic value is,

$$\begin{aligned}
\gamma^2_{\beta,\hat{\beta}_{\rho_N^*}} &= \left\| \begin{bmatrix} \gamma_{\beta,z} & \frac{\gamma_{\beta,\hat{\beta}}-\gamma_{\hat{\beta},z}\gamma_{\beta,z}}{\sqrt{1-\gamma^2_{\hat{\beta},z}}} \end{bmatrix} \right\|_2^2 \\
&= \gamma^2_{\beta,z} + \frac{(\gamma_{\beta,\hat{\beta}} - \gamma_{\hat{\beta},z}\gamma_{\beta,z})^2}{1-\gamma^2_{\hat{\beta},z}} \\
&\overset{a.s.}{\sim} \gamma^2_{\beta,z} + \frac{\Psi^{-2}(1-\gamma^2_{\beta,z})^2}{1-(\frac{\gamma_{\beta,z}}{\Psi})^2} \\
&= \gamma^2_{\beta,\hat{\beta}_{\bar{\rho}_N}}
\end{aligned}$$

Since $\sin^2\theta_{\beta,\hat{\beta}} = 1-\gamma^2_{\beta,\hat{\beta}} \overset{a.s.}{\to} 1-\Psi^{-2}$, we see clearly that,

$$\frac{\sin^2\theta_{\beta,\hat{\beta}_{\rho_N^*}}}{\sin^2\theta_{\beta,\hat{\beta}}} \overset{a.s.}{\sim} \frac{1-\gamma^2_{\beta,z}}{1-(\frac{\gamma_{\beta,z}}{\Psi})^2} = \frac{\sin^2\theta_{\beta,\hat{\beta}_{\bar{\rho}_N}}}{\sin^2\theta_{\beta,\hat{\beta}}}$$

iii. For the oracle value ρ_N^*, $\hat{\beta}_{\rho_N^*} \in \mathcal{S}_\beta$ so the result follows from Proposition 2.2.

iv. For the equally weighted portfolio $w = \frac{1}{N}1_N$ using $\hat{\beta}_{\rho_N^*}$,

$$\begin{aligned}
\mathscr{R}_w &= \frac{\hat{\sigma}^2_{\rho_N^*} N\gamma^2_{\hat{\beta}_{\rho_N^*},z} + \hat{\delta}^2\frac{1}{N}}{\sigma^2 N\gamma^2_{\beta,z} + \delta^2\frac{1}{N}} \\
&\overset{a.s.}{\sim} \frac{\hat{\sigma}^2_{\rho_N^*}\gamma^2_{\hat{\beta}_{\rho_N^*},z}}{\sigma^2\gamma^2_{\beta,z}},
\end{aligned}$$

where $\hat{\sigma}^2_{\rho_N^*} = \left(\frac{\gamma_{\hat{\beta},z}}{\gamma_{\hat{\beta}_{\rho_N^*},z}}\right)^2 \hat{\sigma}^2$ is the corrected factor variance. Using Corollary B.2, we get,

$$\mathscr{R}_w \overset{a.s.}{\sim} \frac{\sigma^2 \xi \Psi^2 \Psi^{-2} \gamma^2_{\beta,z}}{\sigma^2 \gamma^2_{\beta,z}} \overset{a.s.}{\sim} \xi.$$

□

PROOF OF LEMMA B.1. By orthogonal invariance of X_n and the independence of Y_n,

$$X_n^T Y_n = X_n^T Q_n^T Q_n Y_n = (Q_n X_n)_1 \overset{\mathcal{D}}{=} X_{n1}$$

where Q_n is an orthongal matrix such that $Q_n Y_n = e_1$, e_1 is the first canonical vector, and X_{n1} is the first entry of X_n. We know from Muller (1959) for X_{n1},

$$X_{n1} \overset{\mathcal{D}}{=} \frac{Z_1}{\sqrt{Z_1^2 + \chi^2_{n-1}}},$$

where $Z_1 \sim \mathcal{N}(0, 1)$ and is independent of χ^2_{n-1}. We have,

$$\mathbb{P}(|X_{n1}| \geq \epsilon) \leq \frac{\mathbb{E} X_{n1}^4}{\epsilon^4} \leq \frac{1}{\epsilon^4 n^2} \mathbb{E}\left[Z_1^4\right] \mathbb{E}\left[\left(\frac{n}{\chi^2_n}\right)^2\right] \leq \frac{C}{n^2},$$

where the inverse chi-squared distribution has finite moment for n large enough and C is some constant related to the moments of the standard normal distribution and the inverse chi-squared distribution. By the Borel-Cantelli lemma, we conclude the result.

□

PROOF OF COROLLARY B.2. Let $\mathbf{X} = \mathbf{U}\mathbf{R}$ where $\mathbf{U}$ is the matrix of eigenvectors $(\beta, u_2, \ldots u_N)$ of $\mathbf{\Sigma}$ so that $\text{Cov}(\mathbf{X}) =$ as introduced in the beginning of this section. Also as before let $\mathbf{S}_{\mathbf{X}} = \frac{1}{T}\mathbf{X}\mathbf{X}^T = \hat{\mathbf{V}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}}^T$ be the sample covariance of $\mathbf{X}$ and its eigendecomposition. Then $\mathbf{S} = \frac{1}{T}\mathbf{R}\mathbf{R}^T = \mathbf{U}\hat{\mathbf{V}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}}^T\mathbf{U}^T$ so that the first sample eigenvector of $\mathbf{S}$, $\hat{\beta}$, is given by,

$$\hat{\beta} = \hat{v}_{11}\beta + \sum_{j=2}^{N} \hat{v}_{j1} u_j.$$

Then we have $\gamma_{\hat{\beta},\beta} = \hat{v}_{11}$ and,

$$\gamma_{\hat{\beta},z} = \hat{v}_{11}\gamma_{\beta,z} + \tilde{v}^T \omega_N,$$

where $\omega_N = \begin{bmatrix} u_2^T z & \cdots & u_N^T z \end{bmatrix}$. As noted before, by Shen et al. (2016, Theorem 6.3), $\hat{v}_{11} \overset{a.s.}{\to} \Psi^{-1}$. We know that both $\|\tilde{v}\|_2$ and $\|\omega_N\|_2$ are bounded as,

$$\|\tilde{v}\|_2 \le 1, \quad 1 = \|U^T z\|_2^2 = (\beta^T z)^2 + \sum (u_j^T z)^2 = \gamma_{\beta,z}^2 + \|\omega_N\|_2^2.$$

Therefore, from Lemma B.1 for $X_N = \frac{\tilde{v}}{\|\tilde{v}\|_2}$ and $Y_N = \frac{\omega_N}{\|\omega_N\|_2}$ we have $\tilde{v}^T \omega_N$ converges almost surely to 0 so we conclude the result.

□

C Asymptotic estimates

Throughout this section we will assume a homogeneous risk model $\delta^2 \mathbf{I}$ such that

$$\mathbf{\Sigma} = \sigma_N^2 \beta\beta^T + \delta^2 \mathbf{I}.$$

Also all estimated quantities, $\hat{\sigma}_N^2$, $\hat{\beta}$, $\hat{\delta}^2$ are general and not necessarily the PCA estimates.

As in Clarke et al. (2011), define the minimum variance long-short threshold as

$$\beta_{\text{MV}} = \frac{\sigma^2 + \delta^2}{\sqrt{N}\gamma_{\beta,z}\sigma^2},$$

where an analogous term $\widehat{\beta}_{\text{MV}}$ is defined using estimated quantities. Further, as in Clarke et al. (2011), let w and $\bar{w}$ stand for the normalized and unnormalized weights of the minimum variance portfolio given by,

$$w = \frac{\beta_{\text{MV}} - \beta}{N\beta_{\text{MV}} - \sqrt{N}\gamma_{\beta,z}} = \frac{\frac{\sigma^2+\delta^2}{\sqrt{N}\gamma_{\beta,z}\sigma^2} - \beta}{\sqrt{N}\left(\frac{\sigma^2+\delta^2}{\gamma_{\beta,z}\sigma^2} - \gamma_{\beta,z}\right)}, \tag{28}$$

$$\bar{w} = \beta_{\text{MV}} - \beta = \frac{\sigma^2 + \delta^2}{\sqrt{N}\gamma_{\beta,z}\sigma^2} - \beta \tag{29}$$

As before, analogous quantities $\widehat{w}$ and $\widehat{\bar{w}}$ are defined using estimated quantities.

The following lemmas establish fundamental building blocks for analyzing tracking error and forecast variance ratio. As above, a formula in terms of true quantities has an analog in terms of estimated quantities. Further, our factor model specification implies $\sigma_N^2 = \mathcal{O}(N)$. The lemmas are presented without proofs since they are straightforward to verify from the formulas given above and the asymptotic nature of σ_N^2.

Lemma C.1.

$$\begin{aligned}
w^T w &= \frac{\left(\frac{\frac{\delta^2}{\sigma_N^2}+1}{\gamma_{\beta,z}}\right)^2 - 2\left(\frac{\delta^2}{\sigma_N^2}+1\right)+1}{N\left(\frac{\frac{\delta^2}{\sigma_N^2}+1}{\gamma_{\beta,z}} - \gamma_{\beta,z}\right)^2}, \\
&\sim \frac{1}{1-\gamma_{\beta,z}^2}, \quad N \to \infty \\
\bar{w}^T \bar{w} &= \left(\frac{\frac{\delta^2}{\sigma_N^2}+1}{\gamma_{\beta,z}}\right)^2 - 2\left(\frac{\delta^2}{\sigma_N^2}+1\right)+1 \\
&\sim \frac{1-\gamma_{\beta,z}^2}{\gamma_{\beta,z}^2}, \quad N \to \infty, \\
\widehat{w}^T w &= \frac{\frac{1}{\gamma_{\beta,z}\gamma_{\hat{\beta},z}}\left(\frac{\delta^2}{\sigma_N^2}+1\right)\left(\frac{\hat{\delta}^2}{\hat{\sigma}_N^2}+1\right)}{N\left(\frac{\frac{\delta^2}{\sigma_N^2}+1}{\gamma_{\beta,z}} - \gamma_{\beta,z}\right)\left(\frac{\frac{\hat{\delta}^2}{\hat{\sigma}_N^2}+1}{\gamma_{\hat{\beta},z}} - \gamma_{\hat{\beta},z}\right)} \\
&\quad + \frac{\gamma_{\beta,\hat{\beta}} - \frac{\gamma_{\beta,z}}{\gamma_{\hat{\beta},z}}\left(\frac{\hat{\delta}^2}{\hat{\sigma}_N^2}+1\right) - \frac{\gamma_{\hat{\beta},z}}{\gamma_{\beta,z}}\left(\frac{\delta^2}{\sigma_N^2}+1\right)}{N\left(\frac{\frac{\delta^2}{\sigma_N^2}+1}{\gamma_{\beta,z}} - \gamma_{\beta,z}\right)\left(\frac{\frac{\hat{\delta}^2}{\hat{\sigma}_N^2}+1}{\gamma_{\hat{\beta},z}} - \gamma_{\hat{\beta},z}\right)} \\
&\sim \frac{1-\gamma_{\hat{\beta},z}^2 - \gamma_{\beta,z}^2 + \gamma_{\beta,z}\gamma_{\hat{\beta},z}\gamma_{\beta,\hat{\beta}}}{N(1-\gamma_{\hat{\beta},z}^2)(1-\gamma_{\beta,z}^2)}
\end{aligned}$$

Lemma C.2.

$$\begin{aligned}
w^T \beta &= \frac{\delta^2/\sigma_N^2}{\sqrt{N}\left(\frac{\frac{\delta^2}{\sigma_N^2}+1}{\gamma_{\beta,z}} - \gamma_{\beta,z}\right)} \\
&\sim \frac{\delta^2}{N\sigma_N^2}\frac{\gamma_{\beta,z}}{1-\gamma_{\beta,z}}, \quad N \to \infty \\
\bar{w}^T \beta &= \frac{\delta^2}{\sigma_N^2}
\end{aligned}$$

Lemma C.3.

$$
\begin{aligned}
\widehat{w}^T \beta &= \frac{\frac{\hat{\delta}^2}{\hat{\sigma}_N^2} r_{\hat{\beta}} + (r_{\hat{\beta}} - \gamma_{\beta,\hat{\beta}})}{\sqrt{N}\left(\frac{\frac{\hat{\delta}^2}{\hat{\sigma}_N^2}+1}{\gamma_{\hat{\beta},z}} - \gamma_{\hat{\beta},z}\right)} \\
&\sim \frac{1}{\sqrt{N}} \frac{\gamma_{\hat{\beta},z}}{1-\gamma_{\hat{\beta},z}^2}\left(r_{\hat{\beta}} - \gamma_{\beta,\hat{\beta}}\right), \quad N \to \infty \\
\widehat{\bar{w}}^T \beta &= \frac{\hat{\delta}^2}{\hat{\sigma}_N^2} r_{\hat{\beta}} + (r_{\hat{\beta}} - \gamma_{\beta,\hat{\beta}}) \\
&\sim r_{\hat{\beta}} - \gamma_{\beta,\hat{\beta}}, \quad N \to \infty
\end{aligned}
$$

Lemma C.4 (In-Sample Factor Risk)**.**

$$
\begin{aligned}
\hat{\sigma}_N^2 \widehat{w}^T \hat{\beta}\hat{\beta}^T \widehat{w} &= \frac{\frac{\hat{\delta}^4}{\hat{\sigma}_N^2}}{N\left(\frac{\frac{\hat{\delta}^2}{\hat{\sigma}_N^2}+1}{\gamma_{\beta,z}} - \gamma_{\beta,z}\right)^2} \\
&\sim \frac{\hat{\delta}^4}{N\hat{\sigma}_N^2}\frac{\gamma_{\beta,z}^2}{(1-\gamma_{\beta,z}^2)^2}, \quad N \to \infty \\
\hat{\sigma}_N^2 \widehat{\bar{w}}^T \hat{\beta}\hat{\beta}^T \widehat{\bar{w}} &= \frac{\hat{\delta}^4}{\hat{\sigma}_N^2}
\end{aligned}
$$

Lemma C.5 (Out-of-Sample Factor Risk)**.**

$$
\begin{aligned}
\sigma_N^2 \widehat{w}^T \beta\beta^T \widehat{w} &= \sigma_N^2 \left(\frac{\frac{\hat{\delta}^2}{\hat{\sigma}_N^2} r_{\hat{\beta}} + (r_{\hat{\beta}} - \gamma_{\beta,\hat{\beta}})}{\sqrt{N}\left(\frac{\frac{\hat{\delta}^2}{\hat{\sigma}_N^2}+1}{\gamma_{\hat{\beta},z}} - \gamma_{\hat{\beta},z}\right)} \right)^2 \\
&\sim \frac{\sigma_N^2}{N} \frac{\gamma_{\hat{\beta},z}^2}{(1-\gamma_{\hat{\beta},z}^2)^2} \left(r_{\hat{\beta}} - \gamma_{\beta,\hat{\beta}} \right)^2, \quad N \to \infty \\
\sigma_N^2 \widehat{\overline{w}}^T \beta\beta^T \widehat{\overline{w}} &= \sigma_N^2 \left(\frac{\hat{\delta}^2}{\hat{\sigma}_N^2} r_{\hat{\beta}} + (r_{\hat{\beta}} - \gamma_{\beta,\hat{\beta}}) \right)^2 \\
&\sim \sigma_N^2 \left(r_{\hat{\beta}} - \gamma_{\beta,\hat{\beta}} \right)^2, \quad N \to \infty
\end{aligned}
$$

Corollary C.6 (Out-of-Sample Factor Risk)**.** *Let* $\hat{\beta} \in \mathcal{S}_\beta$.

$$
\begin{aligned}
\sigma_N^2 \widehat{w}^T \beta\beta^T \widehat{w} &= \sigma_N^2 \left(\frac{\frac{\hat{\delta}^2}{\hat{\sigma}_N^2} r_{\hat{\beta}}}{\sqrt{N}\left(\frac{\frac{\hat{\delta}^2}{\hat{\sigma}_N^2}+1}{\gamma_{\hat{\beta},z}} - \gamma_{\hat{\beta},z}\right)} \right)^2 \\
&\sim \frac{\sigma_N^2}{N} \frac{\gamma_{\hat{\beta},z}^2}{(1-\gamma_{\hat{\beta},z}^2)^2} \frac{\hat{\delta}^4}{\hat{\sigma}_N^4} r_{\hat{\beta}}^2, \quad N \to \infty \\
\sigma_N^2 \widehat{\overline{w}}^T \beta\beta^T \widehat{\overline{w}} &= \sigma_N^2 \left(\frac{\hat{\delta}^2}{\hat{\sigma}_N^2} r_{\hat{\beta}} \right)^2 \\
&\sim \sigma_N^2 \left(r_{\hat{\beta}} - \gamma_{\beta,\hat{\beta}} \right)^2, \quad N \to \infty
\end{aligned}
$$

Lemma C.7 (In-Sample Specific Risk)**.**

$$\widehat{w}^T \widehat{\mathbf{\Delta}} \widehat{w} = \hat{\delta}^2 \frac{\left(\frac{\frac{\delta^2}{\sigma_N^2}+1}{\gamma_{\beta,z}}\right)^2 - 2\left(\frac{\delta^2}{\sigma_N^2}+1\right)+1}{N\left(\frac{\frac{\delta^2}{\sigma_N^2}+1}{\gamma_{\beta,z}} - \gamma_{\beta,z}\right)^2},$$

$$\sim \frac{\hat{\delta}^2}{1-\gamma_{\beta,z}^2}, \quad N \to \infty$$

$$\widehat{\bar{w}}^T \widehat{\mathbf{\Delta}} \widehat{\bar{w}} = \hat{\delta}^2 \left(\left(\frac{\frac{\delta^2}{\sigma_N^2}+1}{\gamma_{\beta,z}}\right)^2 - 2\left(\frac{\delta^2}{\sigma_N^2}+1\right)+1\right)$$

$$\sim \hat{\delta}^2 \frac{1-\gamma_{\beta,z}^2}{\gamma_{\beta,z}^2}, \quad N \to \infty$$

Lemma C.8 (Out-of-Sample Specific Risk)**.**

$$\widehat{w}^T \mathbf{\Delta} \widehat{w} = \delta^2 \frac{\left(\frac{\frac{\delta^2}{\sigma_N^2}+1}{\gamma_{\beta,z}}\right)^2 - 2\left(\frac{\delta^2}{\sigma_N^2}+1\right)+1}{N\left(\frac{\frac{\delta^2}{\sigma_N^2}+1}{\gamma_{\beta,z}} - \gamma_{\beta,z}\right)^2},$$

$$\sim \frac{\delta^2}{1-\gamma_{\beta,z}^2}, \quad N \to \infty$$

$$\widehat{\bar{w}}^T \mathbf{\Delta} \widehat{\bar{w}} = \delta^2 \left(\left(\frac{\frac{\delta^2}{\sigma_N^2}+1}{\gamma_{\beta,z}}\right)^2 - 2\left(\frac{\delta^2}{\sigma_N^2}+1\right)+1\right)$$

$$\sim \delta^2 \frac{1-\gamma_{\beta,z}^2}{\gamma_{\beta,z}^2}, \quad N \to \infty$$

Lemma C.9 (Squared Tracking Error)**.**

$$\begin{aligned}
\mathscr{T}_{\widehat{w}}^2 &= (\widehat{w}-w)^T \mathbf{\Sigma} (\widehat{w}-w) \\
&\sim \frac{\sigma^2}{N} \frac{\gamma_{\hat{\beta}}^2}{(1-\gamma_{\hat{\beta}}^2)^2} \left(\frac{\gamma_\beta}{\gamma_{\hat{\beta}}} - \gamma_{\beta,\hat{\beta}} \right)^2 + \delta^2 \frac{1}{N(1-\gamma_{\hat{\beta}}^2)} \\
&\quad + \frac{\delta^4}{N\sigma^2} \frac{\gamma_\beta^2}{(1-\gamma_\beta^2)^2} + \delta^2 \frac{1}{N(1-\gamma_\beta^2)} \\
&\quad - 2\frac{\sigma^2}{N} \left(\frac{\gamma_{\hat{\beta}}}{1-\gamma_{\hat{\beta}}^2} \left(\frac{\gamma_\beta}{\gamma_{\hat{\beta}}} - \gamma_{\beta,\hat{\beta}} \right) \right) \left(\frac{\delta^2}{\sigma^2} \frac{\gamma_\beta}{1-\gamma_\beta^2} \right) \\
&\quad - 2\delta^2 \frac{\left(1-\gamma_{\hat{\beta}}^2 - \gamma_\beta^2 + \gamma_\beta \gamma_{\hat{\beta}} \gamma_{\beta,\hat{\beta}}\right)}{N(1-\gamma_\beta^2)(1-\gamma_{\hat{\beta}}^2)} \\
&\sim \frac{\sigma^2}{N} \frac{\gamma_{\hat{\beta}}^2}{(1-\gamma_{\hat{\beta}}^2)^2} \left(\frac{\gamma_\beta}{\gamma_{\hat{\beta}}} - \gamma_{\beta,\hat{\beta}} \right)^2 + \delta^2 \frac{\gamma_{\hat{\beta}}^2 - \gamma_\beta^2}{N(1-\gamma_\beta^2)(1-\gamma_{\hat{\beta}}^2)}, \quad N \to \infty
\end{aligned}$$

Corollary C.10 (Squared Tracking Error)**.** *Let* $\hat{\beta} \in \mathcal{S}_\beta$.

$$\mathscr{T}_{\widehat{w}}^2 \sim \delta^2 \frac{\gamma_{\hat{\beta}}^2 - \gamma_\beta^2}{N(1-\gamma_\beta^2)(1-\gamma_{\hat{\beta}}^2)}, \quad N \to \infty$$

Lemma C.11 (Forecast Variance Ratio)**.**

$$\begin{aligned}
\mathscr{R}_{\widehat{w}} &= \frac{\hat{\sigma}_N^2 \,\widehat{\overline{w}}^T \hat{\beta}\hat{\beta}^T \widehat{\overline{w}} + \widehat{\overline{w}}^T \widehat{\mathbf{\Delta}} \widehat{\overline{w}}}{\sigma_N^2 \,\widehat{\overline{w}}^T \beta\beta^T \widehat{\overline{w}} + \widehat{\overline{w}}^T \mathbf{\Delta} \widehat{\overline{w}}} \\
&= \frac{\frac{\hat{\delta}^4}{\hat{\sigma}_N^2} + \hat{\delta}^2 \left(\left(\frac{\frac{\delta^2}{\sigma_N^2}+1}{\gamma_{\hat{\beta},z}} \right)^2 - 2\left(\frac{\delta^2}{\sigma_N^2} + 1 \right) + 1 \right)}{\sigma_N^2 \left(\frac{\hat{\delta}^2}{\hat{\sigma}_N^2} r_{\hat{\beta}} + (r_{\hat{\beta}} - \gamma_{\beta,\hat{\beta}}) \right)^2 + \delta^2 \left(\left(\frac{\frac{\delta^2}{\sigma_N^2}+1}{\gamma_{\hat{\beta},z}} \right)^2 - 2\left(\frac{\delta^2}{\sigma_N^2} + 1 \right) + 1 \right)} \\
&\sim \frac{\hat{\delta}^2}{\sigma_N^2 \frac{\gamma_{\hat{\beta},z}^2}{(1-\gamma_{\hat{\beta},z}^2)^2} \left(r_{\hat{\beta}} - \gamma_{\beta,\hat{\beta}} \right)^2 + \delta^2}, \quad N \to \infty
\end{aligned}$$

Corollary C.12 (Forecast Variance Ratio)**.** *Let* $\hat{\beta} \in \mathbb{S}_{\beta}$.

$$\mathscr{R}_{\widehat{w}} = \frac{\frac{\hat{\delta}^4}{\hat{\sigma}_N^2} + \hat{\delta}^2 \left(\left(\frac{\frac{\delta^2}{\sigma_N^2}+1}{\gamma_{\hat{\beta},z}} \right)^2 - 2\left(\frac{\delta^2}{\sigma_N^2} + 1 \right) + 1 \right)}{\sigma_N^2 \left(\frac{\hat{\delta}^2}{\hat{\sigma}_N^2} r_{\hat{\beta}} \right)^2 + \delta^2 \left(\left(\frac{\frac{\delta^2}{\sigma_N^2}+1}{\gamma_{\hat{\beta},z}} \right)^2 - 2\left(\frac{\delta^2}{\sigma_N^2} + 1 \right) + 1 \right)}$$

$$\sim \frac{\hat{\delta}^2}{\delta^2}, \quad N \to \infty$$

References

Bai, J. & Ng, S. (2008), 'Large dimensional factor analysis', *Foundations and Trends® in Econometrics* **3**(2), 89–163.

Bianchi, S. W., Goldberg, L. R. & Rosenberg, A. (2017), 'The impact of estimation error on latent factor models of portfolio risk', *The Journal of Portfolio Management* **43**(5), 145–156.

Bickel, P. J. & Levina, E. (2008), 'Covariance regularization by thresholding', *The Annals of Statistics* pp. 2577–2604.

Björck, Å. & Golub, G. H. (1973), 'Numerical methods for computing angles between linear subspaces', *Mathematics of computation* **27**(123), 579–594.

Blume, M. E. (1975), 'Betas and their regression tendencies', *The Journal of Finance* **30**(3), 785–795.

Britten-Jones, M. (1999), 'The sampling error in estimates of mean-variance efficient portfolio weights', *The Journal of Finance* **54**(2), 655–671.

Bun, J., Bouchaud, J. & Potters, M. (2016), 'Cleaning correlation matrices', *Risk Magazine* .

Clarke, R., De Silva, R. & Thorley, S. (2011), 'Minimum-variance portfolio composition', *Journal of Portfolio Management* **2**(37), 31–45.

Connor, G. & Korajczyk, R. A. (1986), 'Performance measurement with the arbitrage pricing theory: A new framework for analysis', *Journal of financial economics* **15**, 373–394.

Connor, G. & Korajczyk, R. A. (1988), 'Risk and return in equilibrium apt: Application of a new test methodology', *Journal of financial economics* **21**, 255–289.

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

DeMiguel, V., Garlappi, L. & Uppal, R. (2007), 'Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy?', *The review of Financial studies* **22**(5), 1915–1953.

El Karoui, N. (2013), 'On the realized risk of high-dimensional markowitz portfolios', *SIAM Journal on Financial Mathematics* **4**(1), 737–783.

El Karoui, N. et al. (2010), 'High-dimensionality effects in the markowitz problem and other quadratic programs with linear constraints: risk underestimation', *The Annals of Statistics* **38**(6), 3487–3566.

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

Frost, P. A. & Savarino, J. E. (1986), 'An empirical bayes approach to efficient portfolio selection', *Journal of Financial and Quantitative Analysis* **21**(3), 293–305.

Gillen, B. J. (2014), 'An empirical bayesian approach to stein-optimal covariance matrix estimation', *Journal of Empirical Finance* **29**, 402–420.

Goldfarb, D. & Iyengar, G. (2003), 'Robust portfolio selection problems', *Mathematics of operations research* **28**(1), 1–38.

Green, R. C. & Hollifield, B. (1992), 'When will mean-variance efficient portfolios be well diversified?', *The Journal of Finance* **47**(5), 1785–1809.

Jagannathan, R. & Ma, T. (2003), 'Risk reduction in large portfolios: Why imposing the wrong constraints helps', *The Journal of Finance* **58**(4), 1651–1683.

Jobson, J. D. & Korkie, B. (1980), 'Estimation for markowitz efficient portfolios', *Journal of the American Statistical Association* **75**(371), 544–554.

Lai, T. L. & Xing, H. (2008), *Statistical models and methods for financial markets*, Springer.

Lai, T. L., Xing, H. & Chen, Z. (2011), 'Mean-variance portfolio optimization when means and covariances are unknown', *The Annals of Applied Statistics* pp. 798–823.

Ledoit, O. & Péché, S. (2011), 'Eigenvectors of some large sample covariance matix ensembles', *Probability Theory and Related Fields* **151**(1–2), 233–264.

Ledoit, O. & Wolf, M. (2003), 'Improved estimation of the covariance matrix of stock returns with an application to portfolio selection', *Journal of empirical finance* **10**(5), 603–621.

Ledoit, O. & Wolf, M. (2004), 'Honey, i shrunk the sample covariance matrix', *The Journal of Portfolio Management* **30**, 110–119.

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

Michaud, R. O. & Michaud, R. O. (2008), *Efficient asset management: a practical guide to stock portfolio optimization and asset allocation*, Oxford University Press.

Muller, M. E. (1959), 'A note on a method for generating points uniformly on n-dimensional spheres', *Communications of the ACM* **2**(4), 19–20.

Onatski, A. (2012), 'Asymptotics of the principal components estimator of large factor models with weakly influential factors', *Journal of Econometrics* **168**(2), 244–258.

Pástor, L. (2000), 'Portfolio selection and asset pricing models', *The Journal of Finance* **55**(1), 179–223.

Paul, D. (2007), 'Asymptotics of sample eigenstructure for a large dimensional spiked covariance model', *Statistica Sinica* pp. 1617–1642.

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

Sharpe, W. F. (1964), 'Capital asset prices: A theory of market equilibrium under conditions of risk', *The Journal of Finance* **19**(3), 425–442.

Shen, D., Shen, H., Zhu, H. & Marron, S. (2016), 'The statistics and mathematics of high dimensional low sample size asympotics', *Statistica Sinica* **26**(4), 1747–1770.

Treynor, J. L. (1962), Toward a theory of market value of risky assets. Presented to the MIT Finance Faculty Seminar.

Vasicek, O. A. (1973), 'A note on using cross-sectional information in bayesian estimation of security betas', *The Journal of Finance* **28**(5), 1233–1239.

Wang, W. & Fan, J. (2017), 'Asymptotics of empirical eigenstructure for high dimensional spiked covariance', *The Annals of Statistics* **45**(3), 1342–1374.

Yata, K. & Aoshima, M. (2012), 'Effective pca for high-dimension, low-sample-size data with noise reduction via geometric representations', *Journal of multivariate analysis* **105**(1), 193–215.