James-Stein for the Leading Eigenvector

Lisa R. Goldberg and Alec N. Kercheval

August 13, 2022

Abstract

Recent research identifies and corrects bias, such as excess dispersion, in the leading sample eigenvector of a factor-based covariance matrix estimated from a high dimension low sample size (HL) data set. We show that eigenvector bias can have a substantial impact on variance-minimizing optimization in the HL regime, while bias in estimated eigenvalues may have little effect. We describe a data-driven eigenvector shrinkage estimator in the HL regime called "James-Stein for Eigenvectors" (JSE) and its close relation to the James-Stein (JS) estimator for a collection of averages. We show, both theoretically and with numerical experiments, that, for certain variance-minimizing problems of practical importance, efforts to correct eigenvalues have little value in comparison to the JSE correction of the leading eigenvector. When certain extra information is present, JSE is a consistent estimator of the leading eigenvector.

Significance Statement

Eigenvectors are used throughout the physical and social sciences to reduce the dimension of complex problems to manageable levels, and to distinguish signal from noise. Our research identifies and mitigates bias in the leading eigenvector of a sample factor-based covariance matrix estimated in the high dimension low sample size (HL) regime. The analysis illuminates how estimation error in a covariance matrix can affect quadratic optimization. Eigenvector estimation in the HL regime may be useful for disciplines, such as finance, machine learning, or genomics, in which high dimensional variables need to be analyzed from a limited number of observations.

Averaging is the most important tool for distilling information from data. To name just two of countless examples, batting average is a standard measure of the likelihood that a baseball player will get on base, and an average of squared security returns is commonly used to estimate the variance of a portfolio of stocks.

The average can be the best estimator of a mean in the sense of having the smallest mean squared error. But a strange thing happens when considering a collection of many averages simultaneously. The aggregate sum of mean squared errors is no longer minimized by the collection of averages. Instead, the error can be reduced by shrinking the averages toward a common target, even if, paradoxically, there is no underlying relation among the quantities.

For baseball players, since an individual batting average incorporates both the true mean and estimation error from sampling, the largest observed batting average is prone to be over-estimated and the smallest under-estimated. That is why the aggregate mean squared error is reduced when the collection of observed averages are all moved toward their center.

This line of thinking has been available at least since Sir Francis Galton introduced "regression towards mediocrity" in 1886. Still, Charles Stein surprised the community of statisticians with a sequence of papers about this phenomenon beginning in the 1950s. Stein showed that it is always possible to lower the aggregate squared error of a collection of three or more averages by explicitly shrinking them toward their collective average. In 1961, Stein improved and simplified the analysis in collaboration with Willard

James. The resulting empirical James-Stein shrinkage estimator (JS) launched a new era of statistics.

This article describes "James-Stein for Eigenvectors" (JSE), a recently discovered shrinkage estimator for the leading eigenvector of an unknown covariance matrix. A leading eigenvector is a direction in a multi-dimensional data set that maximizes explained variance. The variance explained by the leading eigenvector is the leading eigenvalue.

Like a collection of averages, a sample eigenvector is a collection of values that may be overly dispersed. This happens in the high dimension low sample size (HL) regime, when the number of variables is much greater than the number of observations. In this situation, the JSE estimator reduces excess dispersion in the entries of the leading sample eigenvector. The HL regime arises when a relatively small number of observations is used to explain or predict complex high-dimensional phenomena, and it falls outside the realm of classical statistics. Examples of such settings include genome-wide association studies, such as [1] and [2], in which characteristics of a relatively small number of individuals might be explained by millions of single nucleotide polymorphisms (SNPs); machine learning in domains with a limited number of high dimensional observations, such as in [3]; and finance, in which the number of assets in a portfolio can greatly exceed the number of useful observations.

We work in the context of factor models and principal component analysis, which are used throughout the physical and social sciences to reduce dimension and identify the most important drivers of complex outcomes. Principal component analysis (PCA) is a statistical technique that uses eigenvectors as factors. The results in this article are set in the context of a one-factor model that generates a covariance matrix with a single spike. This means that the leading eigenvalue is substantially larger than the others. We do not provide a recipe for practitioners working in higher rank contexts; our goal is to describe these ideas in a setting in which we can report the current state of the theory. However, similar results are reported experimentally for multifactor models by Goldberg, Papanicolaou, Shkolnik, and Ulucam [4], and continuing theoretical work indicates that the success of this approach is not limited to the one-factor case.

We begin this article by describing the JS and JSE shrinkage estimators side by side, in order to highlight their close relationship. We then describe three asymptotic regimes, low dimension high sample size (LH), high dimension high sample size (HH), and high dimension low sample size (HL), in order to clarify the relationship between our work and the literature. Subsequently we describe an optimization-based context in which a high dimensional covariance matrix estimated with the JSE estimator performs substantially better than eigenvalue correction estimators coming from the HH literature. We describe both theoretical and numerical supporting results for performance metrics relevant to minimum variance optimization.

The novelty of the work described in this article lies in an an explicit focus on high-dimensional covariance matrix estimation via shrinkage of eigenvectors, rather than eigenvalues or the entire covariance matrix; the reliance on results from the less-studied HL regime; and the use of optimization-based performance metrics. The bulk of the existing high-dimensional covariance estimation literature concerns correction of biased eigenvalues, or provides results only in the HH regime, or focuses on metrics that do not take account of the use of covariance matrices in optimization.

James-Stein for averages

Suppose there are $p > 3$ unknown means $\mu = (\mu_1, \mu_2, \ldots, \mu_p)$ to be estimated. We observe a fixed number of samples, and compute the corresponding sample averages $z = (z_1, z_2, \ldots, z_p)$.

It is common practice to use z_i as an estimate for the unobserved mean value μ_i, and this may be the best one can do if estimating only a single mean. The discovery of Stein [5] and James & Stein [6] is that a better estimate is obtained by shrinking the sample averages toward their collective average.

Let $m(z) = \sum_{i=1}^{p} z_i/p$ denote the collective average, and $\mathbf{1} = (1, 1, \ldots, 1)$, the p-dimensional vector of 1s. With certain normality assumptions, James and

Stein define:

$$\hat{\mu}^{\mathrm{JS}} = m(z)\mathbf{1} + c^{\mathrm{JS}}(z - m(z)\mathbf{1}). \tag{1}$$

The shrinkage constant c^{JS} is given by

$$c^{\mathrm{JS}} = 1 - \frac{\nu^2}{s^2(z)}, \tag{2}$$

where

$$s^2(z) = \frac{1}{p-3}\sum_{i=1}^{p}(z_i - m(z))^2 \tag{3}$$

is a measure of the variation of the sample averages z_i around their collective average $m(z)$, and ν^2 is an estimate of the conditional variance of each sample average around its unknown mean. The value of ν^2, a measure of the noise affecting each observed average, must be either assumed or estimated independently of $s^2(z)$, and is sometimes tacitly taken to be 1.

The observable quantity $s^2(z)$ incorporates both the unobserved variation of the means and the noise ν^2. The term $\nu^2/s^2(z)$ in equation (2) can be thought of as an estimated ratio of noise to the sum of signal and noise. Equation (1) calls for a lot of shrinkage when the noise dominates the variation of the sample averages around their collective average, and only a little shrinkage when the reverse is true. Readers may consult Efron and Morris [7], [8], and Efron [9] for more complete discussion and motivation behind formula (1) as an empirical Bayes estimator.

James and Stein showed that the JS estimator $\hat{\mu}^{\mathrm{JS}}$ is superior to z in the sense of expected mean squared error,

$$E_{\mu,\nu}\left[|\hat{\mu}^{\mathrm{JS}} - \mu|^2\right] < E_{\mu,\nu}\left[|z - \mu|^2\right]. \tag{4}$$

For any fixed μ and ν, the conditional expected mean squared error is improved when using $\hat{\mu}^{\mathrm{JS}}$ instead of z. This result comes with an unavoidable caveat: z remains the optimal estimate when $p = 1$ and $p = 2$, and sometimes when $p = 3$.

Suppose we have $p > 3$ baseball players, and, for $i = 1, 2, \ldots, p$, player i has true batting average μ_i, meaning that in any at-bat the player has a probability μ_i of getting a hit. This probability is not observable, but we do observe, say over the first 50 at-bats of the season, the realized proportion z_i of hits. Assuming we know ν^2 or have an independent way to estimate it, equation (1) improves on the z_i as estimates of the true means μ_i.

This example lends intuition to the role of the noise to signal-plus-noise ratio $\nu^2/s^2(z)$ in the JS shrinkage constant. If the true batting averages differ widely, but the sample averages tend to be close to the true values, then equation (1) calls for little shrinkage, as appropriate. Alternatively, if the true averages are close together but the sampling error is large, a lot of shrinkage makes sense. The JS estimator properly quantifies the shrinkage and interpolates between these extremes.

James-Stein for eigenvectors

Consider a sequence of n independent observations of a variable of dimension p, drawn from a population with unobserved covariance matrix Σ. The $p \times p$ sample covariance matrix S has the spectral decomposition:

$$S = \lambda^2 h h^\top + \lambda_2^2 v_2 v_2^\top + \lambda_3^2 v_3 v_3^\top \cdots + \lambda_p^2 v_p v_p^\top \tag{5}$$

in terms of the non-negative eigenvalues $\lambda^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_p^2 \geq 0$ and orthonormal eigenvectors $\{h, v_2, \ldots, v_p\}$ of S. Our interest is primarily in the leading eigenvalue λ^2 and its corresponding eigenvector h. In what follows, the sample eigenvector h will play the role of the collection of sample averages z from the previous discussion.

In classical statistics with fixed p, the sample eigenvalues and eigenvectors are consistent estimators of their population counterparts, at least when the population eigenvalues are distinct. This means that the sample estimates converge to the population values as n tends to infinity. However, this may fail when the dimension tends to infinity. The purpose of JSE is to provide an empirical estimator improving on the sample eigenvector h in our HL setting.

JSE is a shrinkage estimator, analogous to JS, that improves on h by having lower squared error with high probability, and leading to better estimates of covariance matrices for use in quadratic optimization. Goldberg, Papanicolaou, and Shkolnik introduced and analyzed the JSE estimator in [10] as a

means to improve the output of quadratic optimization. It is further developed and generalized by Goldberg, et. al. [4], Shkolnik [11], and Gurdogan and Kercheval [12].

We suppose that $p >> n$ and the entries of b have a non-zero average, $m(b) = \sum_{i=1}^{p} b_i/p$, which we are free to assume is positive by change of sign if needed. We also assume the leading sample eigenvector has positive average entry $m(h) > 0$.

The JSE estimator h^{JSE} is defined by shrinking the entries of h toward their average, just as in equation (1):

$$h^{\text{JSE}} = m(h)\mathbf{1} + c^{\text{JSE}}(h - m(h)\mathbf{1}). \tag{6}$$

The shrinkage constant c^{JSE} is

$$c^{\text{JSE}} = 1 - \frac{\nu^2}{s^2(h)}, \tag{7}$$

where

$$s^2(h) = \frac{1}{p}\sum_{i=1}^{p} (\lambda h_i - \lambda m(h))^2 \tag{8}$$

is a measure of the variation of the entries of λh around their average $\lambda m(h)$, and ν^2 is equal to the average of the non-zero smaller eigenvalues of S, scaled by $1/p$,

$$\nu^2 = \frac{tr(S) - \lambda^2}{p \cdot (n-1)}. \tag{9}$$

As with JS, JSE calls for a lot of shrinkage when the average of the non-zero smaller eigenvalues dominates the variation of the entries of λh around their average, and only a little shrinkage when the reverse is true.

The estimator h^{JSE} improves on the sample leading eigenvector h of S, as we describe below, by reducing excess dispersion. Here, dispersion $d(h)$ is a scale-invariant quantity defined as the square root of the average squared deviation of the entries of h from their mean, divided by the mean:

$$d^2(h) = \frac{1}{p}\sum_{i=1}^{p} \left(\frac{h_i - m(h)}{m(h)}\right)^2. \tag{10}$$

To state a precise result, we need to introduce the factor model framework in which we are applying JSE, as initiated in [10] and further elaborated in [11] and [12]. Factor models are widely used to reduce dimension in settings where, among many factors, a few dominate. The prototype is a one-factor model:

$$r = \beta f + \epsilon, \tag{11}$$

where r is an observable p-vector, β is a p-vector of factor loadings, f is a common factor through which the observable variables are correlated, and ϵ is a p-vector of variable-specific effects that are not necessarily small, but are uncorrelated with f and each other. Setting the factor variance to be σ^2 and the specific variance to be δ^2, the population covariance matrix takes the form:

$$\Sigma = \sigma^2 \beta\beta^\top + \delta^2 I. \tag{12}$$

Our theoretical results are asymptotic in the number of variables p, so we introduce a fixed sequence of scalars $\{\beta_i\}_{i=1}^{\infty}$ from which we draw factor loadings. Suppressing dependence on dimension in our notation, let β be the p-vector whose entries are the first p elements of the fixed sequence. The leading eigenvector of the population covariance matrix (12) is $b = \beta/|\beta|$. Let h denote the unit leading eigenvector of the sample covariance matrix, as before.

Theorem 0.1 ([10]). *Assume, as p tends to infinity, that $m(\beta)$ and $d(\beta)$ have positive and finite limits.*

Then, in the limit as $p \to \infty$,

$$\left| \frac{h^{\text{JSE}}}{|h^{\text{JSE}}|} - b \right|^2 < |h - b|^2 \tag{13}$$

almost surely.

We illustrate (13) in Figure 1. The left panel shows JSE shrinkage as defined by equation (6). The right panel shows an equivalent formulation of JSE shrinkage in terms vectors on the unit sphere obtained by normalization.

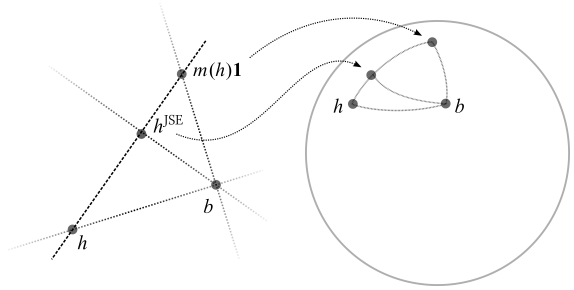


Figure 1: Shrinkage of the sample eigenvector h along the line connecting h and $m(h)\mathbf{1}$ in Euclidean space (left panel) and projected on the unit sphere (right panel.

A more general shrinkage target

In equations (1) and (6), JS and JSE reduce excess dispersion in an estimated vector of interest relative to a "shrinkage target," $\tau = c\mathbf{1}$, with constant entries. Efron and Morris [7] describe the JS estimator for a more general shrinkage target, where $m(h)\mathbf{1}$ is replaced by an arbitrary "initial guess" τ for the unknown μ. In that case the JS estimator becomes

$$\hat{\mu}^{\mathrm{JS}} = \tau + c^{\mathrm{JS}}(h - \tau) \tag{14}$$

where c^{JS} is defined as before but with a modified $s^2(h)$:

$$s^2(h) = \sum_{i=1}^{p} (h_i - \tau_i)^2/(p-2). \tag{15}$$

There is a parallel extension for eigenvector shrinkage. The multiple anchor point shrinkage (MAPS) estimator introduced in [12] is a version of JSE in which the shrinkage target $\tau = m(h)\mathbf{1}$ is generalized to incorporate additional information.

To describe this, first define an "anchor point" to be a vector in R^p, such as $\mathbf{1}$ or perhaps an independent estimate of b, toward which the population eigenvector b may be biased. If $\{q_1, \ldots, q_k\}$ is a collection of anchor points in R^p, $1 \le k < p$, define a shrinkage target τ for h as follows. Let L denote the linear subspace spanned by those anchor points. Then the MAPS shrinkage target $\tau_L(h)$ is defined as the orthogonal projection of h onto L. The generalized JSE estimator shrinks h toward $\tau_L(h)$:

$$h^{\mathrm{JSE}} = \tau_L(h) + c^{MAPS}(h - \tau_L(h)), \tag{16}$$

where the shrinkage constant c^{MAPS} is given by equation (7), with ν^2 as in equation (9) but with

$$s^2(h) = (\lambda^2/p)(1 - ||\tau_L(h)||^2). \tag{17}$$

This is a direct generalization of JSE with a single anchor point $q_1 = \mathbf{1}$, in which this case $\tau_L(h)$ reduces to $m(h)\mathbf{1}$ and formula (17) reduces to (8).

It is shown in [12], under the prevailing assumptions on β, that adding new anchor points at random does no harm asymptotically, but that additional anchor points can improve the estimator if they contain extra information. For example, using a PCA eigenvector estimated from a prior observation period as an additional anchor point will improve the estimate when the betas are slowly varying over time.

For another example, suppose we know the rank ordering of the betas $\beta_1, \beta_2, \ldots, \beta_p$, but not their actual values. A refined JSE estimator can be constructed in the following way. Order the betas by size and group them into k ordered quantiles, where k is approximately $\sqrt{p}$. For $i = 1, \ldots, k$, define the anchor point $q_i = (a_1, \ldots, a_p)$ where $a_j = 1$ if β_j belongs to group i, and zero otherwise, and let L^* be the subspace spanned by $\{q_1, \ldots, q_k\}$.

Theorem 0.2 ([12])**.** *In the setting of Theorem 0.1, the JSE estimator defined by equation (16) with the shrinkage target $\tau_{L^*}(h)$ is a consistent estimator of b in the sense that*

$$\lim_{p\to\infty} \left| \frac{h^{\mathrm{JSE}}}{|h^{\mathrm{JSE}}|} - b \right| = 0 \tag{18}$$

almost surely, with n fixed.

In [12] it is shown that the full rank ordering is not needed, only the ordered groupings are used.

Three regimes

The two James-Stein estimators, for averages and for the leading eigenvector, are structurally parallel, but

the current state of theory guarantees their performance in different settings. The dominance of JS over the sample mean expressed in inequality (4) holds in expectation, typically under normality assumptions, for finite $p > 3$. In contrast, the JSE theory of Theorems 0.1 and 0.2 is asymptotic in the HL regime, and is non-parametric, courtesy of the strong law of large numbers.

The relevance of the HL regime to the analysis of scientific data was recognized as early as 2005, by Hall, Marron and Neeman [13]. The 2018 article by Aoshima et al. [14] surveys results on the HL regime.

The HL regime stands in contrast to the low dimension high sample size (LH) regime of classical statistics, where the number of variables p is fixed and the number of observations n tends to infinity. In the LH regime, a sample covariance matrix based on identically distributed, independent observations is a consistent estimator of the population covariance matrix, converging in expectation as n tends to infinity. Different effects emerge in the high-dimension high-sample-size (HH) regime, in which both p and n tend to infinity. The HH regime is part of random matrix theory, dating back to the 1967 work of Marčenko and Pastur [15]. This three-regime classification of data analysis is discussed by Jung and Marron in their 2009 article [16].

Placing any particular finite problem into an asymptotic context, whether LH, HL, HH or something in between, requires specifying how the model is to be extended asymptotically. For LH this means letting the number of independent observations grow, but the HH and HL regimes require defining a sequence of models of increasing dimension. This extension was natural in early works from random matrix theory that character the limiting spectra of standard Gaussian variables in the HH regime. Johnstone [17] looks at the HH spectrum of eigenvalues in a spiked model, where the eigenvalues of a fixed-dimensional set of eigenvectors are substantially larger than the remaining eigenvalues. The covariance matrix corresponding to the factor model (11) is spiked. In some settings, it can be beneficial to estimate the spiked covariance model guided by Theorems 0.1 and 0.2 from the HL regime.

A schematic diagram of the three regimes is in Figure 2. Duality enables us to use classical statistics to obtain results in the HL regime. This has been observed by various researchers, including Shen and co-authors in 2016 [18] and Wang and Fan in 2017 [19], and used in [10]. For example, if Y is our $p \times n$ data matrix with $p > n$, the $p \times p$ sample covariance matrix $YY^\top/n$ has rank at most n. If we consider the $n \times n$ dual matrix $S^D = Y^\top Y/p$, it has a fixed dimension in the HL regime. The non-zero eigenvalues of S^D and S are related by the multiplicative factor p/n, and the eigenvectors are related by left multiplication by Y or $Y^\top$. Since, for S^D, the roles of p and n are reversed, methods from classical statistics apply.

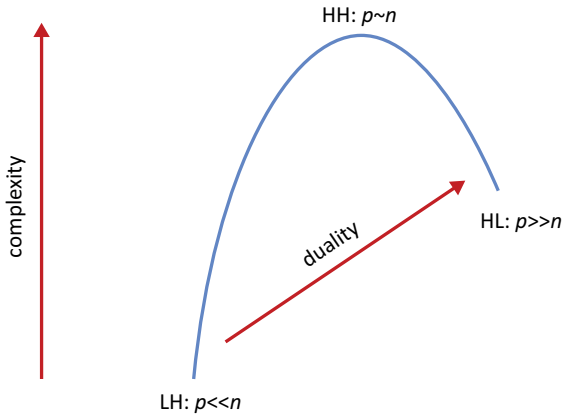


Figure 2: Three asymptotic regimes for data analyis. LH is the low dimension high sample size regime of classical statistics. HH is the high dimension high sample size regime of classical random matrix theory. HL is the high dimension low sample size regime of alternative random matrix theory. HH tends to be more complex than HL because duality arguments allow some features of classical statistics to emerge in the HL regime.

High dimensional covariance matrix estimation

Eigenvalue adjustment to improve covariance performance metrics, or loss functions, goes back at least

to Stein's 1956 and 1986 articles [20] and [21]. In this section we selectively discuss aspects of the literature.

In their 2018 article [22], Donoho, Gavish and Johnstone emphasize the dependence of the optimal estimator on the choice of performance metric. Like Stein [21], they consider estimators obtained by varying the eigenvalues while keeping the sample eigenvectors fixed. In describing a benchmark oracle optimal estimator for their spiked covariance model in the HH regime, they write:

> The oracle procedure does not attain zero loss since it is "doomed" to use the eigenbasis of the empirical covariance, which is a random basis corrupted by noise, to estimate the population covariance.

This situation is reasonable in the context they consider in which there is no prior information, other than data, about the eigenvectors. As indicated in [11] and [12], prior information can allow for the correction of a wide range of eigenvector biases in the HL regime.

Similar themes emerges from a a series of articles [23], [24], [25], [26], [27] and [28], by Ledoit and Wolf. Beginning in 2003, these papers explore high-dimensional covariance matrix estimation with applications to financial portfolio construction and other disciplines. As in the paper by Donoho, Gavish and Johnstone [22], Ledoit and Wolf, in their 2017 article, [28], consider "the class of rotation-equivariant estimators".

Ledoit and Wolf write:

> Rotation equivariance is appropriate in the general case where the statistician has no a priori information about the orientation of the eigenvectors of the covariance matrix....
>
> The fact that we keep the sample eigenvectors does not mean that we assume they are close to the population eigenvectors. It only means that we do not know how to improve upon them.

In earlier papers Ledoit and Wolf consider estimators that shrink a sample covariance matrix toward a target. Some of these estimators modify the sample eigenvectors. By implementing a spiked shrinkage target in [25], Ledoit and Wolf provide prior structural information to the estimator. For the JSE estimator, that structural information is in the form of a factor model and the positive mean assumption on the leading population eigenvector.

In their 2017 article, Wang and Fan [19] develop the S-POET eigenvalue shrinkage estimator that can be applied to the spiked covariance model in the HH and certain HL regimes. They evaluate S-POET with performance metrics based on the relative spectral norm, the relative Frobenius norm, the spectral norm and the max norm. Their candidate estimators, again, use the sample eigenvectors. In the absence of structural information, they also remark that "correction for the biases of estimating eigenvectors is almost impossible".

Despite the challenges of characterizing or correcting sample eigenvectors in high dimensions, there are streams of literature on the subject in both the HH and HL regimes. Much of the literature concerns consistency of sample eigenvectors under different modeling assumptions. HH references include Paul [29], Nadler [30], Mestre [31], and Johnstone and Lu [32]. A 2018 survey by Johnstone and Paul [33] has an extensive reference list. In a 2022 article [34], Zhong, Su and Fan develop an iterative method to shrink eigenvectors in the HH regime. Techniques from the HH regime have been applied to improve optimized portfolios; see, for example, the 2012 paper by Menchero and Orr [35], and the 2013 publication by El Karoui [36].

For the HL regime, asymptotics and estimation of eigenvectors has been studied in work previously cited and, among others, Ahn, Marron, Muller, and Chi [37], Jung, Sen, and Marron [38], Lee, Zou, and Wright [39], and Jung [40].

In the next section, we introduce a focus on optimization error and relevant performance metrics. We show that JSE eigenvector shrinkage, perhaps surprisingly, can substantially dominate the gains due to eigenvalue correction in optimization-based performance metrics.

JSE corrects an optimization bias

Estimated covariance matrices are used in quadratic optimization, which chooses coefficients to minimize the variance of a linear combination of random variables subject to constraints. In what follows, we evaluate estimators of high-dimensional spiked covariance matrices with performance metrics that measure the accuracy of optimized quantities.

We present simulations of practical situations where JSE materially improves optimization-based performance metrics while eigenvalue corrections can have little effect. Our simulations illustrate results from [10] and [12] showing the dependence of optimization-based performance metrics on the optimization bias as the number of variables p tends to infinity, and the lack of dependence of these metrics on errors in eigenvalues. Our context and examples are taken from financial economics, but our results apply in any discipline where spiked covariance models are used as inputs to quadratic optimization.

Quantitative portfolio construction

From a universe of p financial securities, there are countless ways to construct a portfolio. We focus on quantitative portfolio construction, which has relied on mean-variance optimization since Markowitz [41]. In this framework, a portfolio is represented by a vector whose ith entry is the fraction or *weight* of the portfolio invested in security i. A portfolio is *efficient* if it has minimum forecast variance subject to constraints, and the search for efficient portfolios is central to quantitative finance. The simplest efficient portfolio is minimum variance.

A fully invested but otherwise unconstrained minimum variance portfolio is the solution $\hat{w}^*$ to the mean-variance optimization problem

$$\begin{aligned} &\min_{w\in\mathbb{R}^p} w^\top \widehat{\Sigma} w \\ &\text{subject to:} \\ &\quad w^\top \mathbf{1} = 1, \end{aligned} \tag{19}$$

where the $p \times p$ matrix $\widehat{\Sigma}$ is a non-singular estimate of the unknown true security covariance matrix Σ. If the estimate $\widehat{\Sigma}$ is derived from observed data, then $\hat{w}^*$ is a data-driven approximation of the true optimum w^*, defined as the solution to (19) with $\widehat{\Sigma}$ replaced by Σ.

Performance metrics and optimization

We review three performance metrics that are sensitive to different aspects of the impact of covariance matrix estimation error on optimization.

The *variance forecast ratio* (VFR) is the quotient of estimated by true variance of a linear combination of random variables. Considered in 1956 by Stein [20] for arbitrary combinations, the VFR can be substantially less than the maximum value 1 when it is applied to an optimized quantity like a minimum variance portfolio:

$$\mathrm{VFR}(\hat{w}^*) = \frac{\hat{w^*}^\top \widehat{\Sigma} \hat{w}^*}{\hat{w^*}^\top \Sigma \hat{w}^*}. \tag{20}$$

This is because a variance-minimizing optimization tends to place excess weight on securities whose variances and correlations with other securities are under-forecast. In the words of Richard Michaud [42], mean-variance optimizers are "estimation error maximizers". Bianchi, Goldberg and Rosenberg [43] use the VFR to assess risk underforecasting in optimized portfolios. By considering the additional metrics described next, we are able to gauge the accuracy of optimized portfolios themselves, not merely the accuracy of their risk forecasts.

Unlike the VFR the *true variance ratio* TVR makes sense only for optimized combinations of random variables. TVR is the quotient of true variance of the true quantity by true variance of the optimized quantity, and it measures excess variance in the latter:

$$\mathrm{TVR}(\hat{w}^*) = \frac{w^* \Sigma w^*}{\hat{w^*}^\top \Sigma \hat{w}^*}. \tag{21}$$

A more direct measure of the accuracy of an optimized quantity is *tracking error*, which we define as:

$$\mathrm{TE}^2(\hat{w}^*) = (\hat{w}^* - w^*)^\top \Sigma (\hat{w}^* - w^*) \tag{22}$$

for the minimum variance portfolio. Tracking error is widely used by portfolio managers to measure the width of the distribution of the difference in return of two portfolios, and it is commonly applied to measure the distance between a portfolio and its benchmark.

In simulation, these performance metrics illuminate different aspects of the impact of error in $\widehat{\Sigma}$ on $\hat{w}^*$. Since they require knowledge of the true covariance matrix Σ, they cannot be used directly in an empirical study. The numerator of VFR, the true variance of the optimized quantity, can be approximated in out-of-sample empirical tests.

Factor models, eigenvalues, and eigenvectors

When $p > n$, the sample covariance matrix S is singular, and so is not a candidate for $\widehat{\Sigma}$. Factor models are used throughout the financial services industry and the academic literature to generate full-rank estimates of security return covariance matrices. In the discussion below, we rely on the one-factor model specified in (11). However, similar results are obtained numerically in the case of multiple factors and non-homogeneous specific risk in [4], and are supported by theoretical work currently in development.

Writing the factor loadings β as a product $|\beta|b$ of a scale factor and a unit vector, the population covariance matrix (12) takes the form

$$\Sigma = (\sigma^2|\beta|^2)bb^\top + \delta^2 I. \tag{23}$$

The quantities σ^2 and $|\beta|^2$ are not identifiable from data, but we can estimate their product $\eta = \sigma^2|\beta|^2$. Estimating Σ reduces to finding estimators $\hat{b} \in R^p$, and $\hat{\eta}, \hat{\delta}^2 \in R$ so that

$$\widehat{\Sigma} = \hat{\eta}\hat{b}\hat{b}^\top + \hat{\delta}^2 I. \tag{24}$$

In what follows, we construct the estimate (24) in the HL regime. The number of variables p is increased by means of the assumptions that $m(\beta)$ and $d(\beta)$ tend to positive limits as $p \to \infty$. This reflects our intention that the factor loading β_i for the ith variable not tend to zero or infinity with i, but remains in a possibly large neighborhood of a common positive mean value. The number of observations n stays fixed. These considerations are empirically consistent with return data from equity markets.

In the HL regime, the factor and specific variances can be expressed in terms of sample eigenvalues. Recall λ^2 is the leading eigenvalue of the sample covariance matrix S, assumed to be of rank n, and let ℓ^2 denote the average of the remaining non-zero eigenvalues of S,

$$\ell^2 = \frac{tr(S) - \lambda^2}{n-1}. \tag{25}$$

Under the assumptions of Theorem 0.1, Lemma A.2 of [10] provides the asymptotic relationships between eigenvalues of S and factor model parameters. For large p,

$$\lambda^2 \approx \frac{|\beta|^2|f|^2}{n} + \frac{p}{n}\delta^2, \tag{26}$$

where $f = (f_1, \ldots, f_n)$ is the vector of realizations of the common factor return corresponding to the n observations, and

$$\ell^2 \approx \frac{p}{n}\delta^2. \tag{27}$$

The asymptotic equality symbol $\approx$ means equality, after division by p, in the limit as $p \to \infty$. An immediate consequence is approximation for the trace of S in terms of the elements of the factor model:

$$tr(S) \approx \frac{|\beta|^2|f|^2}{n} + p\delta^2. \tag{28}$$

Although we don't have access to $|f|^2/n$, it is an unbiased estimator of the true factor variance σ^2. Relabelling $|f|^2/n$ by $\hat{\sigma}^2$ and applying formulas (26)–(27) gives us:

$$\hat{\sigma}^2|\beta|^2 \quad \approx \quad \lambda^2 - \ell^2. \tag{29}$$

This means we have good asymptotic estimators $\hat{\eta} = \lambda^2 - \ell^2$ and $\hat{\delta}^2 = (n/p)\ell^2$ that determine, for any choice of eigenvector estimator $\hat{b}$, the covariance estimator

$$\hat{\Sigma}(\hat{b}) = (\lambda^2 - \ell^2)\hat{b}\hat{b}^\top + (n/p)\ell^2 I \tag{30}$$

with leading eigenvalue $\lambda^2 - \ell^2 + (n/p)\ell^2$ and trace $\lambda^2 + (n-1)\ell^2$. The leading eigenvalue is approximately equal to the population eigenvalue $\sigma^2|\beta|^2 + \delta^2$.

It also agrees, for $p >> n$, with the S-POET leading eigenvalue estimate of Wang and Fan [19], developed in a regime that includes our spiked HL setting.

It remains to estimate b, the leading population eigenvector. To help quantify the effect of estimation error on our performance metrics, we use the following two quantities defined for any non-zero eigenvector estimate $\hat{b}$ of b. The "optimization bias" $\mathcal{E}(\hat{b})$, introduced in [10], is

$$\mathcal{E}^2(\hat{b}) = \frac{(b,q) - (b,\hat{b})(\hat{b},q)}{1-(\hat{b},q)^2}. \tag{31}$$

and the "eigenvector bias" $\mathcal{D}(\hat{b})$, introduced in [12], is

$$\mathcal{D}(\hat{b}) = \frac{(\hat{b},q)^2(1-(\hat{b},b)^2)}{(1-(\hat{b},q)^2)(1-(b,q)^2)} \tag{32}$$

where q is the unit vector $\mathbf{1}/\sqrt{p}$ and $(\cdot,\cdot)$ denotes the Euclidean inner product. Note $\mathcal{E}^2(b) = 0$, meaning the population eigenvector has zero bias, as desired.

As shown in [10], [12], and discussed below, these bias measures are substantial contributors to the optimization-based performance metrics VFR, TVR and TE. A primary lesson of [10] is that eigenvalue estimates can be less important, for the purpose of optimization in the HL regime, than estimating the leading eigenvector. This is especially true when considering the true variance $(\hat{w}^*)^\top \Sigma \hat{w}^*$ of an estimated minimum risk portfolio $\hat{w}^*$ defined by equation (19) using estimated covariance matrix.

Correcting the optimization bias

In a factor model in the HL regime, JSE can correct the optimization bias (31), leading to greater accuracy in optimized quantities. Theoretical guarantees of this assertion are expressed in terms of $\eta = \sigma^2|\beta|^2$ and its estimator $\hat{\eta}$ from (24), for $\hat{w}^*$, the minimum variance portfolio using the estimated covariance matrix (24). As a consequence of our assumptions on β, η is of order p asymptotically, so the covariance matrix of data generated by our factor model is spiked.

Theorem 0.3 ([10], [12])**.** *Assume the population covariance matrix is given by (23) and that the estimates $\hat{\eta}/p$ and $\hat{\delta}$ have positive limits as $p \to \infty$.*

1. *Asymptotically, the true variance of the estimated portfolio is*

$$(\hat{w}^*)^\top \Sigma \hat{w}^* = (\eta/p)\mathcal{E}^2(\hat{b}) + o(p). \tag{33}$$

In particular, the true variance of the estimated minimum variance portfolio is asymptotically independent of eigenvalue estimates, but depends only on the eigenvector estimate $\hat{b}$ and the true covariance matrix Σ.

2. *$\lim_{p\to\infty} \mathcal{E}(h^{\mathrm{JSE}}) = 0$ and $\lim_{p\to\infty} \mathcal{E}(h) > 0$ almost surely, where h is the leading eigenvector of S.*

3. *Asymptotically, the tracking error of the optimal portfolio $\hat{w}^*$ is*

$$TE^2(\hat{w}) = \frac{\eta}{p}\mathcal{E}^2(\hat{b}) + \frac{\delta^2}{p}\mathcal{D}(\hat{b}) + \frac{C}{p}\mathcal{E}(\hat{b}) + o(p), \tag{34}$$

where C is a constant depending on the population covariance matrix, the data, $\hat{\eta}$, and $\hat{\delta}$, but not on $\hat{b}$ (see [12]).

If we denote by w_{PCA} the minimum variance portfolio constructed using the sample eigenvector h in (30), and w_{JSE} using h^{JSE}, parts 1 and 2 of Theorem 0.3 imply that $\mathrm{TVR}(w_{\mathrm{PCA}})$ tends to zero as the dimension p tends to infinity, but $\mathrm{TVR}(w_{\mathrm{JSE}})$ does not. From parts 2 and 3, it follows that $TE^2(w_{\mathrm{PCA}})$ is bounded below, and $TE^2(w_{\mathrm{JSE}})$ tends to zero.

Simulations calibrated to financial markets in [4], [10] and [12] illustrate that these asymptotic properties are already present for values of p and n that emerge in financial markets. In addition, we observe the variance forecast ratio is drastically improved by the JSE estimator.

Numerical illustration

Consider the problem of estimating a covariance matrix of stock returns from a year's worth of daily observations for an index like the S&P 500. The observation frequency and size of the data window are limited by empirical considerations independent of the dimension: stocks enter and exit the index, markets

undergo changes in volatility, and intra-day sampling magnifies serial correlation.

In the case at hand, we have approximately $n = 252$ days to estimate a covariance matrix for perhaps $p = 500$ variables. Since $p > n$, this problem falls outside the realm of classical statistics. Whether this falls under the HH or HL regime, and which performance metrics should be used, depend on application details.

We examine a hypothetical market driven by the one-factor model (11) with covariance matrix (23). Because the diagonal elements of S are unbiased estimators of the population variances, the trace $tr(S)$ is an unbiased estimator of the sum $tr(\Sigma)$ of the population variances. As a consequence, we preserve $tr(S)$ in our covariance matrix estimators. We consider the following three data-driven, trace-preserving estimators:

$$\Sigma_{\text{raw}} = (\lambda^2 - \frac{n-1}{p-1}\ell^2)hh^\top + \frac{n-1}{p-1}\ell^2 I \tag{35}$$

$$\Sigma_{\text{PCA}} = (\lambda^2 - \ell^2)hh^\top + (n/p)\ell^2 I \tag{36}$$

$$\Sigma_{\text{JSE}} = (\lambda^2 - \ell^2)h^{\text{JSE}}(h^{\text{JSE}})^\top + (n/p)\ell^2 I. \tag{37}$$

Here, Σ_{raw} matches the leading eigenvalue and eigenvector of S without correction. Σ_{PCA} has the corrected leading eigenvalue, but still uses the leading eigenvector h to estimate b; Σ_{JSE} improves further by substituting h^{JSE} of (6) for h.

Our factor model parameters are taken approximately from [10] and [4], which contain detailed information about calibration to financial markets. We draw factor and specific returns f and ϵ independently with mean 0 and standard deviations 16% and 60%, respectively. In the simulation, factor returns are normal, and specific returns are drawn from a t-distribution with 5 degrees of freedom. We use this fat-tailed t-distribution to illustrate that the results do not require Gaussian assumptions; repeating the experiment with several different distributions including the normal gives similar results.

The entries of β, or factor loadings, are inspired by market betas. We draw entries of β independently from a normal distribution with mean 1 and variance 0.25, and hold them fixed across time and simulations.

We compare the effect of eigenvalue vs eigenvector correction on our portfolio performance metrics. In the experiment summarized in Figure 3, we fix $p = 500$, $n = 252$, and examine the tracking error, variance forecast ratio, and true variance ratio for each of the three estimators Σ_{raw}, Σ_{PCA}, and Σ_{JSE}, with box plots summarizing the values for 400 simulations.

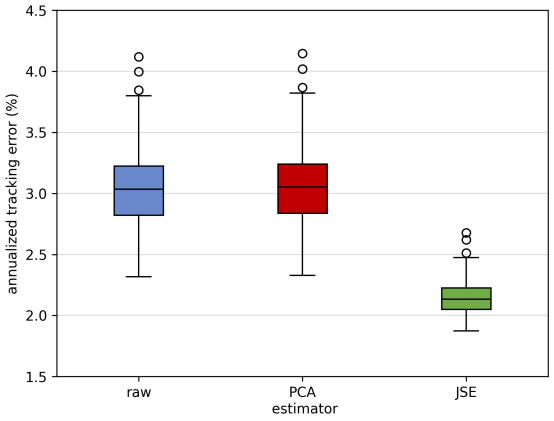
Correcting the leading eigenvalue, from λ^2 to the asymptotically correct $\lambda^2 - (1 - n/p)\ell^2$, has little effect compared to the JSE eigenvector correction. Related experiments described in [10] and [4] confirm that improving the accuracy of optimized quantities has negligible dependence on the eigenvalue estimator and almost entirely on the choice of eigenvector. All else equal, the magnitude of the improvement in accuracy increases as the dispersion of beta decreases.

Comparing our experiment to the numerical study in [19] illustrates a conclusion from [22]: the choice of performance metric materially affects the optimal covariance matrix estimator.

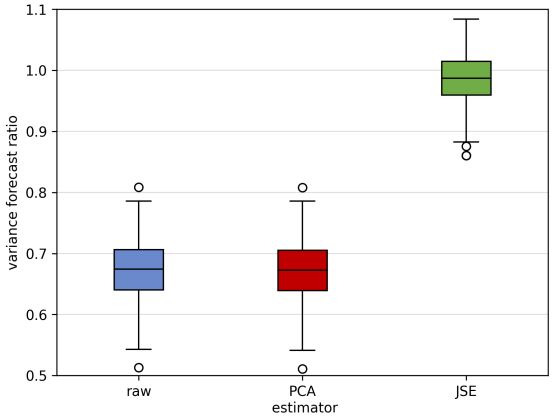
Outlook

This article concerns James Stein for eigenvectors, a shrinkage method that is structurally identical to classical James Stein. JSE has asymptotic guarantees to improve optimization-based performance metrics in the high dimension low sample size HL regime. In the context of an empirically motivated one-factor model with a spiked covariance matrix, we show theoretically and illustrate numerically that optimization error is materially reduced by the JSE estimator, while relatively unaffected by eigenvalue correction.

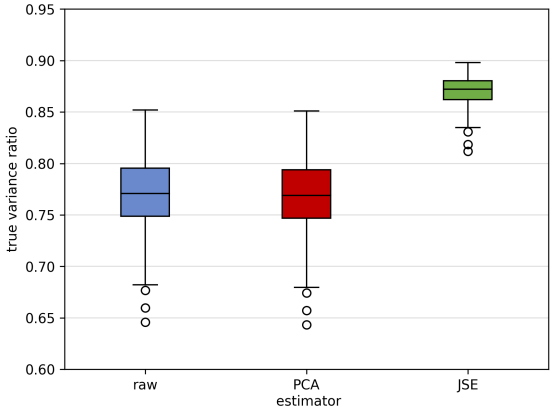
Next steps are to extend the theoretical results to multi-factor models and further develop the link between constrained optimization and eigenvector bias. Open problems include an empirical Bayes formulation of JSE for finite p and n, and a more comprehensive understanding of the relationship between performance metrics and errors in eigenvectors and eigenvalues. The notion of "three regimes" is a simplified framework that allows us to organize results, but, in reality, the three regimes belong to a family of largely uninvestigated possibilities. Applications of



(a) Tracking error



(b) Variance forecast ratio



(c) True variance ratio

Figure 3: Portfolio-level accuracy metrics for simulated minimum variance portfolios optimized with Σ_{raw}, Σ_{PCA}, and Σ_{JSE}. A perfect tracking error is equal to zero, and perfect variance forecast ratios and true variance ratios are equal to one. The estimated covariance matrix is based on $n = 252$ observations of $p = 500$ securities. Each boxplot summarizes 400 simulations. The experiments show that eigenvalue correction (PCA) makes no improvement, but the eigenvector correction (JSE) is substantial.

JSE to GWAS studies, machine learning, and other high dimension low sample size empirical problems await exploration.

Supplementary Materials

Python simulation code used to create the boxplots in Figure 3 is available at https://github.com/kercheval-a/JSE.

Acknowledgements

The results presented here were developed in collaboration with Hubeyb Gurdogan, Alex Papanicolaou, Alex Shkolnik and Simge Ulucam. The authors are grateful to Jeongyoun Ahn, Sungkyu Jung, Youhong Lee, Ola Mahmoud, Caroline Ribet, Ken Ribet, Stephanie Ribet and Michelle Shkedi for comments on early drafts. This article has benefited from review by two anonymous referees and an anonymous editor. We thank Stephanie Ribet and Alex Shkolnik for graphics support.

References

[1] Gen Li and Sungkyu Jung. Incorporating covariates into integrated factor analysis of multi-view data. *Biometrics*, 73:1433–1442, 2017.

[2] The 1000 genomes project consortium: A global reference for human genetic variation. *Nature*, 526:68–74, 2015.

[3] A Vabalas, E Gowen, E Poliakoff, and AJ Casson. Machine learning algorithm validation with a limited sample size. *PLoS ONE*, 14(11), 2019.

[4] Lisa R. Goldberg, Alex Papacinicolau, Alex Shkolnik, and Simge Ulucam. Better betas. *The Journal of Portfolio Management*, 47(1):119–136, 2020.

[5] Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate distribution. In *Proc. Third Berkeley Symp. Math. Stat. Prob.*, pages 197–206, 1956.

[6] W. James and Charles Stein. Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Stat. Prob.*, pages 361–397, 1961.

[7] Bradley Efron and Carl Morris. Data analysis using stein's estimator and its generalizations. *J. of the American Statistical Assoc.*, 70(350):311–319, 1975.

[8] Bradley Efron and Carl Morris. Stein's paradox in statistics. *Scientific American*, 236(5):119–127, 1977.

[9] Bradley Efron. *Large Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge Univ Press, 2010.

[10] Lisa R. Goldberg, Alex Papanicolau, and Alex Shkolnik. The dispersion bias. *SIAM Journal of Financial Mathematics, forthcoming*, 2022.

[11] Alex Shkolnik. James-stein shrinkage for principal components. *Stat*, 2021.

[12] Hubeyb Gurdogan and Alec Kercheval. Multi anchor point shrinkage for the sample covariance matrix. *The SIAM Journal on Financial Mathematics, forthcoming*, 2022.

[13] Peter Hall, J.S. Marron, and Amnon Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444, 2005.

[14] Makoto Aoshima, Dan Shen, Haipeng Shen, Kazuyoshi Yata, Yi-Hui Zhou, , and J. S. Marron. A survey of high dimension low sample size asymptotics. *Australia & New Zealand Jorurnal of Statistics*, 60(1):4–19, 2018.

[15] Vladimir Marčcenko and Leonid Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1867.

[16] Sungkyu Jung and J.S. Marron. PCA consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B):4104–4130, 2009.

[17] Iain Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

[18] Dan Shen, Haipeng Shen, Hongtu Zhu, and J. S. Marron. The statistics and mathematics of high dimension low sample size asymptotics. *Statistica Sinica*, 26(4):1747–1770, 2016.

[19] W. Wang and J. Fan. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics*, 45(3):1342–1374, 2017.

[20] Charles Stein. Some problems in multivariate analysis part i. Technical report no. 6, Office of Naval Research, 1956.

[21] Charles Stein. Lectures on the theory of estimation of many parameters. *Journal of Soviet Mathematics*, 34:1973–1403, 1986.

[22] David Donoho, Matan Gavish, and Iain Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *The Annals of Statistics*, 46(4):1742–1778, 2018.

[23] Oliver Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock retursn with an application to portfolio selection. *J. of Empirical Finance*, 10:603–621, 2003.

[24] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411, 2004.

[25] Olivier Ledoit and Michael Wolf. Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119, 2004.

[26] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, 2012.

[27] Olivier Ledoit and Michael Wolf. Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions. *Journal of Multivariate Analysis*, 139:360–384, 2015.

[28] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *Review of Financial Studies*, 30(12):4349–4388, 2017.

[29] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17:1617–1642, 2007.

[30] Boaz Nadler. Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6):2791–2817, 2008.

[31] Xavier Mestre. of covariance matrices using their sample estimates. *IEEE TRANSACTIONS ON INFORMATION THEORY*, 54(4):5113–5129, 2008.

[32] Iain M. Johnstone and Arthur Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 06 2009.

[33] Iain Johnstone and Debashis Paul. Pca in high dimensions: An orientation. *Proc IEEE Inst Electr Electron Eng.*, 106(8):1277–1292, 2018.

[34] Xinyi Zhong, Chang Su, and Zhou Fan. Empirical Bayes PCA in high dimensions. *J R Stat Soc Series B*, 2022.

[35] Jose Menchero, Jun Wang, and D.J. Orr. Improving risk forecasts for optimized portfolios. *Financial Analysts Journal*, 68(3):40–50, 2012.

[36] Noureddin El Karoui. On the realized risk of high-dimensional markowitz portfolios. *SIAM Journal on Financial Mathematics*, 4:737–783, 2013.

[37] Jeongyoun Ahn, J. S. Marron, Keith M. Muller, and Yuen-Yuh Chi. The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika*, 94(3):76—766, 2007.

[38] Sungkyu Jung, Arusharka Sen, and J. S. Marron. Boundary behavior in high dimension, low sample size asymptotics of PCA. *Journal of Multivariate Analysis*, 109:190–203, 2012.

[39] Seunggeun Lee, Fei Zou, and Fred A. Wright. Convergence of sample eigenvalues, eigenvectors, and principal component scores for ultra-high dimensional data. *Biometrika*, 101(2):484–490, 2014.

[40] Sungkyu Jung. Adjusting systematic bias in high dimensional principal component scores. *Statistica Sinica*, 32:939–959, 2022.

[41] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–92, 1952.

[42] Richard O. Michaud. The Markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal*, 45(1):31–43, 1989.

[43] Stephen W. Bianchi, Lisa R. Goldberg, and Allan Rosenberg. The impact of estimation error on latent factor model forecasts of portfolio risk. *The Journal of Portfolio Management*, 43(5):147–156, 2017.