Multiple Anchor Point Shrinkage for the Sample Covariance Matrix*

Hubeyb Gurdogan† and Alec Kercheval‡

This version: February 19, 2022

Abstract.
Estimation of the covariance of a high-dimensional returns vector is well-known to be impeded by the lack of long data history. We extend the work of Goldberg, Papanicolaou, and Shkolnik (GPS) [14] on shrinkage estimates for the leading eigenvector of the covariance matrix in the high dimensional, low sample-size regime, which has immediate application to estimating minimum variance portfolios. We introduce a more general framework of shrinkage targets – multiple anchor point shrinkage – that allows the practitioner to incorporate additional information – such as sector separation of equity betas, or prior beta estimates from the recent past – to the estimation. We prove some asymptotic statements and illustrate our results with some numerical experiments.

Key words. Covariance matrix estimation, shrinkage, minimum variance portfolio

AMS subject classifications. 91G60, 91G70, 62H25

1. Introduction.

This paper is about the problem of estimating covariance matrices for large random vectors, when the data for estimation is a relatively small sample. We discuss a shrinkage approach to reducing the estimation error asymptotically in the high dimensional, bounded sample size regime, denoted HL. We note at the outset that this context differs from that of the more well-known random matrix theory of the asymptotic "HH regime" in which the sample size grows in proportion to the dimension (e.g. [8]). See [19] for earlier discussion of the HL regime, and [9] for a discussion of the estimation problem for factor models in high dimension.

Our interest in the HL asymptotic regime comes from the problem of portfolio optimization in financial markets. There, a portfolio manager is likely to confront a large number of assets, like stocks, in a universe of hundreds or thousands of individual issues. However, typical return periods of days, weeks, or months, combined with the irrelevance of the distant past, mean that the useful length of data time series is usually much shorter than the dimension of the returns vectors being estimated.

In this paper we extend the successful shrinkage approach introduced in [14] (GPS) to a framework that allows the user to incorporate additional information into the shrinkage target and improve results. Our "multiple anchor point shrinkage" (MAPS) approach includes the GPS method as a special case.

The problem of sampling error for portfolio optimization has been widely studied ever since Markowitz [25] introduced the approach of mean-variance optimization. That paper

*Submitted to the editors (date). The authors thank Lisa Goldberg and Alex Shkolnik for many helpful conversations. Any errors are our own.

†Consortium for Data Analytics and Risk, University of California, Berkeley, CA (hgurdogan@berkeley.edu)

‡Department of Mathematics, Florida State University, Tallahassee, FL (akercheval@fsu.edu, http://www.math.fsu.edu/~kercheva/)

Funding: Thanks to the Simons Foundation for partial support of this work.

immediately gave rise to the importance of estimating the covariance matrix Σ of asset returns, as the risk, measured by variance of returns, is given by $w^T \Sigma w$, where w is the vector of weights defining the portfolio.

For a survey of various approaches over the years, see [14] and references therein. Reducing the number of parameters via factor models has long been standard; see for example [26] and [27]. The applicability of factor models in a very general HL setting is justified by [3]. Discussion of consistent estimation of factors in the HL and HH regimes is contained in [5] and [6]. There, the HH regime in which both p and n tend to infinity is required for exact consistency. In comparison, Theorem 2.3 below attains a consistent estimator of a single factor in the HL setting for a bounded number of observations.

[30] and [12] initiated a Bayesian approach to portfolio estimation and the efficient frontier. Practitioners are frequently interested in estimating the sensitivity (called "beta") of asset returns to the overall market return. Vasicek used a prior cross-sectional distribution for betas to produce an empirical Bayes estimator for beta that amounts to shrinking the least-squares estimator toward the prior in an optimal way. This is one of a number of "shrinkage" approaches in which initial sample estimates of the covariance matrix are "shrunk" toward a prior e.g. [21], [2], [22], [23], [10]. [24] describes a nonlinear shrinkage estimator of the covariance matrix focused on correcting the eigenvalues, set in the HH asymptotic regime. A number of results in the HL and HH regimes related to correcting biases in the spiked covariance setting of factor models are described in [31].

The key insight of [14] was to identify the PCA leading eigenvector of the sample covariance matrix as the primary culprit contributing to sampling error for the minimum variance portfolio problem in the HL asymptotic regime. Their approach to *eigenvector* shrinkage is not explicitly Bayesian, but can be viewed in that spirit (see section 2.5). This is the starting point for the present work.

It is worth pointing out that shrinkage approaches to estimation are far broader than estimating covariance matrices. The books [11] and [16] discusses an array of shrinkage estimators, mainly centered on the famous James-Stein (JS) estimator [20], [7]. The JS estimator as a prototype is not merely incidental to this work: it turns out that there are close structural parallels between JS and GPS/MAPS, as described in the recent works [29] and [13].

1.1. Mathematical setting and background. Next we describe the mathematical setting, motivation, and results in more detail. We restrict attention to a familiar and well-studied (e.g. [28]) baseline model for financial returns: the one-factor, "single-index" or "market", model

$$\mathbf{r} = \beta x + \mathbf{z}, \tag{1.1}$$

where $\mathbf{r} \in \mathbb{R}^p$ is a p-dimensional random vector of asset (excess) returns in a universe of p assets, $\beta \in \mathbb{R}^p$ is an unobserved non-zero vector of parameters to be estimated, $x \in \mathbb{R}$ is an unobserved random variable representing the common factor return, and $\mathbf{z} \in \mathbb{R}^p$ is an unobserved random vector of residual returns specific to the individual assets.

With the assumption that the components of $\mathbf{z}$ are uncorrelated with x and each other, the returns of different assets are correlated only through β, and therefore the covariance matrix of $\mathbf{r}$ is

$$\Sigma = \sigma^2 \beta\beta^T + \Delta,$$

where σ^2 denotes the variance of x, and Δ is the diagonal covariance matrix of $\mathbf{z}$. Typical models in practice use multiple drivers of correlation, so this model represents a base case in which to set our results. However, to the extent that we will measure success below by the performance of the estimated minimum variance portfolio, to a good approximation only a single market factor is relevant ([4], [15]).

Under the further simplifying model assumption[1] that each component of $\mathbf{z}$ has a common variance δ^2 (also not observed), we obtain the covariance matrix of returns

$$\Sigma = \sigma^2 \beta\beta^T + \delta^2 \mathbf{I}, \tag{1.2}$$

where $\mathbf{I}$ denotes the $p \times p$ identity matrix.

This means that β, or its normalization $b = \beta/||\beta||$, is the leading eigenvector of Σ, corresponding to the largest eigenvalue $\sigma^2||\beta||^2 + \delta^2$. As estimating b becomes the most significant part of the estimation problem for Σ, a natural approach is to take as an estimate the first principal component (leading unit eigenvector) h_{PCA} of the sample covariance of returns data generated by the model. This principal component analysis (PCA) estimate is our starting point.

Consider the optimization problem

$$\min_{w \in \mathbb{R}^p} w^T \Sigma w$$
$$e^T w = 1$$

where $e = (1, 1, \ldots, 1)$, the vector of all ones.

The solution, the "minimum variance portfolio", is the unique fully invested portfolio minimizing the variance of returns. Of course the true covariance matrix Σ is not observable and must be estimated from data. Denote an estimate by

$$\hat{\Sigma} = \hat{\sigma}^2 \hat{\beta}\hat{\beta}^T + \hat{\delta}^2 \mathbf{I} \tag{1.3}$$

corresponding to estimated parameters $\hat{\sigma}$, $\hat{\beta}$, and $\hat{\delta}$.

Let $\hat{w}$ denote the solution of the optimization problem

$$\min_{w \in \mathbb{R}^p} w^T \hat{\Sigma} w$$
$$e^T w = 1.$$

It is interesting to compare the estimated minimum variance

$$\hat{V}^2 = \hat{w}^T \hat{\Sigma} \hat{w}$$

[1] The assumption of homogeneous residual variance δ^2 is a mathematical convenience. If the diagonal covariance matrix Δ of residual returns can be reasonably estimated, then the problem can be rescaled as $\Delta^{-1/2}\mathbf{r} = \Delta^{-1/2}\beta x + \Delta^{-1/2}\mathbf{z}$, which has covariance matrix $\sigma^2 \beta_\Delta \beta_\Delta^T + I$, where $\beta_\Delta = \Delta^{-1/2}\beta$.

with the actual variance of $\hat{w}$:

$$V^2 = \hat{w}^T \Sigma \hat{w},$$

and consider the variance forecast ratio $V^2/\hat{V}^2$ as one measure of the error made in the estimation of minimum variance, hence of the covariance matrix Σ.

The remarkable fact proved in [14] is that, asymptotically as p tends to infinity with n fixed, the true variance of the estimated portfolio doesn't depend on $\hat{\sigma}$, $\hat{\delta}$, or $||\hat{\beta}||$, but only on the unit eigenvector $\hat{\beta}/||\hat{\beta}||$. Under some mild assumptions stated later, they show the following.

Definition 1.1. *For a p-vector $\beta = (\beta(1), \ldots, \beta(p))$, define the mean $\mu(\beta)$ and dispersion $d^2(\beta)$ of β by*

$$\mu(\beta) = \frac{1}{p}\sum_{i=1}^{p} \beta(i) \text{ and } d^2(\beta) = \frac{1}{p}\sum_{i=1}^{p} \big(\frac{\beta(i)}{\mu(\beta)} - 1\big)^2. \tag{1.4}$$

We use the notation for normalized vectors

$$b = \frac{\beta}{||\beta||},\ q = \frac{e}{\sqrt{p}},\ \text{and } h = \frac{\hat{\beta}}{||\hat{\beta}||}.$$

Proposition 1.1 ([14]). *The true variance of the estimated portfolio $\hat{w}$ is given by*

$$V^2 = \hat{w}^T \Sigma \hat{w} = \sigma^2 \mu^2(\beta)(1 + d^2(\beta))\mathcal{E}^2(h) + o_p$$

where $\mathcal{E}(h)$ is defined by

$$\mathcal{E}(h) = \frac{(b,q) - (b,h)(h,q)}{1 - (h,q)^2},$$

and where the remainder o_p is such that for some constants c, C, $c/p \le o_p \le C/p$ for all p.

In addition, the variance forecast ratio $V^2/\hat{V}^2$ is asymptotically equal to $p\mathcal{E}^2(h)$.

Goldberg, Papanicolaou and Shkolnik call the quantity $\mathcal{E}(h)$ the *optimization bias* associated to an estimate h of the true vector b. They note that the optimization bias $\mathcal{E}(h_{PCA})$ is asymptotically bounded above zero almost surely, and hence the variance forecast ratio explodes as $p \to \infty$.

With this background, the estimation problem becomes focused on finding a better estimate h of b from an observed time series of returns. GPS [14] introduces a shrinkage estimate for b – the GPS estimator h_{GPS} – obtained by "shrinking" the PCA eigenvector h_{PCA} along the unit sphere toward q, to reduce excess dispersion. That is, h_{GPS} is obtained by moving a specified distance (computed only from observed data) toward q along the spherical geodesic connecting h_{PCA} and q. "Shrinkage" refers to the reduced geodesic distance to the "shrinkage target" q.

The GPS estimator h_{GPS} is a significant improvement on h_{PCA}. First, $\mathcal{E}(h_{GPS})$ tends to zero with p, and in fact $p\mathcal{E}^2(h_{GPS})/\log\log(p)$ is bounded (proved in [17]). In [14] it is conjectured, with numerical support, that $E[p\mathcal{E}^2(h_{GPS})]$ is bounded in p, and hence the expected variance forecast ratio remains bounded. Moreover, asymptotically h_{GPS} is closer than h_{PCA} to the true value b in the ℓ_2 norm, and it yields a portfolio with better tracking error against the true minimum variance portfolio.

1.2. Our contributions. The purpose of this paper is to generalize the GPS estimator by introducing a way to use additional information about beta to adjust the shrinkage target q in order to improve the estimate.

We can consider the space of all possible shrinkage targets τ as determined by the family of all nontrivial proper linear subspaces L of $\mathbb{R}^p$ as follows. Given L (assumed not orthogonal to h), let the unit vector $\tau(L)$ be the normalized orthogonal projection of h onto L. $\tau(L)$ is then a shrinkage target for h determined by L (and h). We will describe such a subspace L as the linear span of a set of unit vectors called "anchor points". In the case of a single anchor point q, note that $\tau(\text{span}\{q\}) = q$, so this case corresponds to the GPS shrinkage target.

The "MAPS" estimator is a shrinkage estimator with a shrinkage target defined by an arbitrary collection of anchor points, usually including q. When q is the only anchor point, the MAPS estimator reduces to the GPS estimator. We can therefore think of the MAPS approach as allowing for the incorporation of additional anchor points when this provides additional information.

In Theorem 2.2, we show that expanding $\text{span}\{q\}$ by adding additional anchor points at random asymptotically does no harm, but makes no improvement.

In Theorem 2.3, we show that if the user has certain mild *a priori* rank ordering information about groups of components of β, even with no information about magnitudes, an appropriately constructed MAPS estimator is a consistent estimator in the sense that it converges exactly to the true vector b in the asymptotic limit, even though the sample size is held fixed.

Theorem 2.4 shows that if the betas have positive serial correlation over recent history, then adding the prior PCA estimator h as an anchor point improves the ℓ_2 error in comparison with the GPS estimator, even if the GPS estimator is computed with the same total data history.

The benefit of improving the ℓ_2 error in addition to the optimization bias is that it also allows us to reduce the tracking error of the estimated minimum variance fully invested portfolio, discussed in Section 3 and Theorem 3.1.

In the next sections we present the main results. The framework, assumptions, and statements of the main theorems are presented in Sections 2 and 3. Some simulation experiments are presented in Section 4 to illustrate the impact of the main results for some specific situations. Proofs of the theorems of Section 2 are organized in Section 5, followed by Section 6 describing some open questions for further work.

To limit the length of this article, the proofs of some of the needed technical propositions and lemmas appear in a separate document [18], available online. Additional details and computations may be found in [17].

2. Main Theorems.

2.1. Assumptions and Definitions. We consider a simple random sample history generated from the basic model (1.1). The sample data can be summarized as

$$R = \beta X^T + Z \tag{2.1}$$

where $R \in \mathbb{R}^{p\times n}$ holds the observed individual (excess) returns of p assets for a time window that is set by $n \geq 2$ consecutive observations. We may consider the observables R to be generated by non-observable random variables $\beta \in \mathbb{R}^p$, $X \in \mathbb{R}^n$ and $Z \in \mathbb{R}^{p\times n}$.

The entries of X are the market factor returns for each observation time; the entries of Z are the specific returns for each asset at each time; the entries of β are the exposure of each asset to the market factor, and we interpret β as random but fixed at the start of the observation window of times $1, 2, 3, ..., n$ and remaining constant throughout the window. Only R is observable.

In this paper we are interested in asymptotic results as p tends to infinity with n fixed. Therefore we consider equation (2.1) as defining an infinite sequence of models, one for each p.

To specify the relationship between models with different values of p, we need a more precise notation. We'll let β refer to an infinite sequence $(\beta(1), \beta(2), \dots) \in \mathbb{R}^\infty$, and $\beta^p = (\beta(1), \dots, \beta(p)) \in \mathbb{R}^p$ the vector obtained by truncation after p entries. When the value p is understood or implied, we will frequently drop the superscript and write β for β^p.

Similarly, $Z \in \mathbb{R}^{\infty\times n}$ is a vector of n sequences (the columns), and $Z^p \in \mathbb{R}^{p\times n}$ is obtained by truncating the sequences at p.

With this setup, passing from p to $p + 1$ amounts to simply adding an additional asset to the model without changing the existing p assets. The pth model is denoted

$$R^p = \beta^p X^T + Z^p,$$

but for convenience we will often drop the superscript p in our notation when there is no ambiguity, in favor of equation (2.1).

Let $\mu_p(\beta)$ and $d_p(\beta) \geq 0$ denote the mean and dispersion of β^p, given by

$$\mu_p(\beta) = \frac{1}{p}\sum_{i=1}^{p} \beta(i) \quad \text{and} \quad d_p(\beta)^2 = \frac{1}{p}\sum_{i=1}^{p} (\frac{\beta(i) - \mu_p(\beta)}{\mu_p(\beta)})^2. \tag{2.2}$$

We make the following assumptions regarding β, X and Z:

A1. (Regularity of beta) The entries $\beta(i)$ of β are uniformly bounded, independent random variables, fixed prior to time 1. The mean $\mu_p(\beta)$ and dispersion $d_p(\beta)$ converge to limits $\mu_\infty(\beta) \in (0, \infty)$ and $d_\infty(\beta) \in (0, \infty)$.

A2. (Independence of beta, X, Z) β, X and Z are jointly independent.

A3. (Regularity of X) The entries X_i of X are iid random variables with mean zero, variance σ^2 .

A4. (Regularity of Z) The entries Z_{ij} of Z have mean zero, finite variance δ^2, and uniformly bounded fourth moment. In addition, the n-dimensional rows of Z are mutually independent, and within each row the entries are pairwise uncorrelated.[2]

The assumptions above are for the sake of convenience and to simplify the statements of results, but in practice are non-binding or can be partly relaxed. In assumption A1, boundedness is automatic in a finite market, and the betas can be viewed as constants as a special case if desired (until section 2.4). Once β is determined, it is held fixed during the observation window of length n. In contrast, X and the columns of Z are drawn independently

[2]Note we do not assume β, X, or Z are Normal or belong to any specific family of distributions.

at each of the n observations times. The existence of the limits $\mu_\infty(\beta)$ and $d_\infty(\beta)$ could be relaxed by considering the limit superior and inferior of the sequence at the cost of more complicated theorem statements, so long as $\liminf \mu_p(\beta) \neq 0$, with a change of sign if needed to make it positive.

Assumptions A2 and A3 are conveniences that simplify the analysis and statements of results. In [14] X and Z are only assumed uncorrelated, so the stronger independence assumption, used in our proofs, is not necessary in all cases. Assumption A4 is one of a few alternatives that serve the proofs. The fourth moment condition can be dropped in favor of the additional assumption that the rows of Z are identically distributed, but we prefer boundedness conditions as they are always satisfied in finite markets.

With the given assumptions the covariance matrix Σ_β of R, conditional on β, is

$$\Sigma_\beta = \sigma^2 \beta\beta^T + \delta^2 I. \tag{2.3}$$

Since β stays constant over the n observations, the sample covariance matrix $\frac{1}{n}RR^T$ converges to Σ_β almost surely if n is taken to ∞, and is the maximum likelihood estimator of Σ_β.

We will work with normalized vectors on the unit sphere $\mathbb{S}^{p-1} \subset \mathbb{R}^p$. To that end we define

$$b = \frac{\beta}{||\beta||}\ ,\ q = \frac{e}{\sqrt{p}}, \tag{2.4}$$

where $e = e^p = (1, 1, \ldots, 1) \in \mathbb{R}^p$, and $||.||$ denotes the usual Euclidean norm.

The vector b is the leading eigenvector of Σ_β (corresponding to the largest eigenvalue). We denote by h the PCA estimator of b, i.e. h is the first principal component, or the unit leading eigenvector, of the sample covariance matrix $\frac{1}{n}RR^T$. For convenience we always select the sign of the unit eigenvector h such that the inner product $(h, q) > 0$, ignoring the probability zero case $(h, q) = 0$.

Since β and X appear in the model $R = \beta X + Z$ only as a product, there is a scale ambiguity that we can resolve by combining their scales into a single parameter η:

$$\eta^p = \frac{1}{p}|\beta^p|^2\sigma^2.$$

It is easy to verify that

$$\eta^p = \mu_p(\beta)^2(d_p(\beta)^2 + 1)\sigma^2,$$

and therefore by our assumptions η^p tends to a positive, finite limit η^∞ as $p \to \infty$.

Our covariance matrix becomes

$$\Sigma_\beta \equiv \Sigma_b = p\eta bb^T + \delta^2 I, \tag{2.5}$$

where we drop the superscript p when convenient. The scalars η, δ and the unit vector b are to be estimated by $\hat{\eta}$, $\hat{\delta}$, and h. As described above, asymptotically only the estimate h of b will be significant. Improving this estimate is the main technical goal of this paper.

In [14] the PCA estimate h is replaced by an estimate h_{GPS} that is “data driven”, meaning that it is computable solely from the observed data R. We henceforth use the notation

$h_{GPS} = \hat{h}_q$, for a reason that will be clear shortly. As an intermediate step we also consider a non-observable "oracle" version h_q, defined as the point on the short $\mathbb{S}^{p-1}$-geodesic joining h to q that is closest to b. (Recall that both b and h are chosen to lie in the half-sphere centered at q.) The oracle version is not data driven because it requires knowledge of the unobserved vector b that we are trying to estimate, but it is a useful concept in the definition and analysis of the data driven version. Both the data driven estimate $\hat{h}_q$ and the oracle estimate h_q can be thought of as obtained from the eigenvector h via "shrinkage" along the geodesic connecting h to the anchor point, q.

The GPS data-driven estimator $\hat{h}_q$ is successful in improving the variance forecast ratio, and in arriving at a better estimate of the true variance of the minimum variance portfolio. In this paper we have the additional goal of reducing the ℓ_2 error of the estimator, which, for example, is helpful in reducing tracking error. To that end, we introduce the following new data driven estimator, denoted $\hat{h}_L$.

Let $L = L_p \subset \mathbb{R}^p$ denote a nontrivial proper linear subspace of $\mathbb{R}^p$. If v is any vector in $\mathbb{R}^p$, we write

$$\underset{L}{\mathrm{proj}}(v)$$

for the Euclidean orthogonal projection of v onto L. Denote by k_p the dimension of L_p, with $1 \leq k_p \leq p - 1$.

Let $h = h^p$ denote our normalized leading eigenvector of $\frac{1}{n} R^p (R^p)^T$, s_p^2 its largest eigenvalue, and l_p^2 the average of the remaining non-zero eigenvalues. Then we define the data driven "MAPS" (Multiple Anchor Point Shrinkage) estimator by

$$\hat{h}_L = \frac{\tau_p h + \underset{L}{\mathrm{proj}}(h)}{||\tau_p h + \underset{L}{\mathrm{proj}}(h)||} \tag{2.6}$$

where

$$\tau_p = \frac{\psi_p^2 - ||\underset{L}{\mathrm{proj}}(h)||^2}{1 - \psi_p^2} \text{ and } \psi_p = \sqrt{\frac{s_p^2 - l_p^2}{s_p^2}}. \tag{2.7}$$

Here ψ_p measures the relative gap between s_p^2 and l_p^2. The MAPS estimator can be viewed as obtained by "shrinking" the PCA estimator h toward the target $\underset{L}{\mathrm{proj}}(h)$ along the sphere $\mathbb{S}^{p-1}$ by a specified amount.

Recall that we sometimes use a superscript to emphasize the dimension of a vector, and the notation $(\cdot, \cdot)$ for the Euclidean inner product of two vectors. The next lemma from [14] describes the asymptotic limit of ψ_p and inner products (h^p, b^p), (h^p, q^p), and (b^p, q^p) as the dimension p tends to infinity.

Lemma 2.1 ([14]). *The limits* $\psi_\infty = \lim_{p\to\infty} \psi_p, (h, b)_\infty = \lim_{p\to\infty}(h^p, b^p), (h, q)_\infty = \lim_{p\to\infty}(h^p, q^p)$, *and* $(b, q)_\infty = \lim_{p\to\infty}(b^p, q^p)$ *exist almost surely. Moreover,*

$$\psi_\infty = (h, b)_\infty \in (0, 1),$$

and

$$(h, q)_\infty = (h, b)_\infty (b, q)_\infty \in (0, 1).$$

When L is the one-dimensional subspace spanned by the vector q, then $\hat{h}_L$ is precisely the GPS estimator $\hat{h}_q$, located along the short spherical geodesic connecting h to q. The phrase "multiple anchor point" comes from thinking of q as an "anchor point" shrinkage target in the GPS paper, and L as a subspace spanned one or more anchor points. The new shrinkage target determined by L is the normalized orthogonal projection of h onto L. When L is the one-dimensional subspace spanned by q, the normalized projection of h onto L is just q itself. In the event that L is orthogonal to h, the MAPS estimator $\hat{h}_L$ reverts to h itself.

2.2. The MAPS estimator with random extra anchor points. Does adding anchor points to create a MAPS estimator from a higher-dimensional subspace improve the estimation? The answer depends on whether there is any relevant information about b in the added anchor points. In the case where there is no added information and we simply add new anchor points at random, the next theorem says this doesn't help.

First some terminology. We say that L_p is a *random linear subspace* of $\mathbb{R}^p$ if it is non-trivial, proper, and the span of a collection of random, linearly independent unit vectors. The random linear subspace H_p is a *uniform random subspace* of $\mathbb{R}^p$ if, in addition, it has spanning vectors are uniformly distributed on the sphere $\mathbb{S}^{p-1}$.[3] We say L_p is *independent of a random variable* Ψ if it has spanning vectors that are independent of Ψ.

Definition 2.1. *A non-decreasing sequence* $\{k_p\}$ *of positive integers is* square root dominated *if*

$$\sum_{p=1}^{\infty} \frac{k_p^2}{p^2} < \infty.$$

For example, any non-decreasing sequence satisfying $k_p \leq Cp^{\alpha}$ for some $C > 0$ and $\alpha < 1/2$ is square root dominated. Roughly speaking, a square-root dominated sequence is one that grows more slowly than $\sqrt{p}$. In particular, any constant sequence qualifies.

Theorem 2.2. *Let the assumptions **1**,**2**,**3** and **4** hold. Suppose, for each p,* L_p *is a random linear subspace and* H_p *is a uniform random subspace of* $\mathbb{R}^p$*. Suppose also that* L_p *is independent of* Z*, and* H_p *is independent of both* Z *and* β*. Assume also the sequences* $\dim L_p$ *and* $\dim H_p$ *are square root dominated.*

Let $L'_p = span\{L_p, q^p\}$ *and* $H'_p = span\{H_p, q^p\}$.

Then, almost surely,

$$\limsup_{p\to\infty} ||\hat{h}_{L'} - b|| \leq \lim_{p\to\infty} ||\hat{h}_q - b||, \tag{2.8}$$

$$\lim_{p\to\infty} ||\hat{h}_{H'} - b|| = \lim_{p\to\infty} ||\hat{h}_q - b||, \tag{2.9}$$

and

$$\lim_{p\to\infty} ||\hat{h}_H - b|| = \lim_{p\to\infty} ||h - b||. \tag{2.10}$$

[3]Uniform random subspaces are called Haar random subspaces in [18] because they can be defined alternatively in terms of the Haar (uniform) measure on the orthogonal group.

The limits on the right hand sides of (2.8), (2.9), and (2.10) exist by an easy application of Lemma 2.1. The need for some upper bounds, such as square root domination, for the dimensions of L and H can be understood by considering the extreme case of maximum dimension p. In that case, the MAPS estimators all reduce to h itself, so (2.8) and (2.9) fail.

Theorem 2.2 says adding random anchor points to form a MAPS estimator does no harm asymptotically, but also makes no improvement asymptotically. Inequality (2.8) says that adding anchor points to q that are independent of Z creates a MAPS estimator that is asymptotically never worse, in the Euclidean distance, than the GPS estimator $\hat{h}_q$, though it might be better (intuitively, if the MAPS estimator incorporates some addtional information about β).

Equation (2.9) says that the GPS estimator is asymptotically neither improved nor harmed by adding extra anchor points uniformly at random when they are independent of β and Z. Therefore the goal will be to find useful anchor points that take advantage of additional information about β that might be available. Necessarily those anchor points will not be independent of β, but can be thought of as creating choices of L'_p to create a strict inequality in (2.8).

Equation (2.10) confirms that the anchor point q used by the GPS estimator has value: without it, a random selection of anchor points independent of β and Z will define a MAPS estimator that is asymptotically no better than the PCA estimator h. While q is not random, it has an implicit relationship to β coming from Assumption A1, which is motivated by the fact that equity betas are empirically observed to cluster around 1. In this sense, the non-random anchor point q contains baseline information about β. This is one of the central intuitions behind the GPS estimator in [14].

As a final remark, notice that in Theorem 2.2 we do not require L or H to be independent of X (but X, Z, and β are mutually independent by Assumption A2). The asymptotic analysis in the proof requires independence from Z in order to apply a version of the strong law of large numbers as $p \to \infty$. In contrast, X does not depend on p and so its contribution can be controlled *a priori* uniformly in p.

2.3. The MAPS estimator with rank order information about the entries of beta.

We now wish to consider what kind of information about β could be added in the form of anchor points to create an improved MAPS estimator.

In this section we consider rank order information. Use of estimated rank ordering of unknown quantities is not new in finance, but has mostly been applied to estimated ordering of returns rather than betas, such as in [1]. Here we consider order information about betas, used in connection with shrinkage estimation.

It so happens that if a well-informed observer somehow knows the rank-ordering of the components of β^p for each p – that is, which entry is the largest, which second largest, etc., then that information alone, without knowing the actual magnitudes, is sufficient to determine b asymptotically with zero error almost surely, using an appropriate MAPS estimator. The resulting consistent estimator is unexpected because the asymptotics are not with regard to sample size n tending to infinty, but rather dimension $p \to \infty$ with fixed n.

In fact, significantly less information than this is needed to create a consistent MAPS estimator in this sense. It suffices to be able to separate the components of beta into ordered groups, where the rank ordering of the groups is known, but not the ordering within groups. The meaning of ordered groups and the constraints on group sizes are explained below.

Definition 2.2. *For any $p \in \mathbb{N}$, let $\mathcal{P} = \mathcal{P}(p)$ be a partition of the index set $\{1, 2, .., p\}$ (i.e. a collection of pairwise disjoint non-empty subsets, called atoms, whose union is $\{1, 2, .., p\}$). The number of atoms of $\mathcal{P}$ is denoted by $|\mathcal{P}|$.*

We say the sequence of partitions $\mathcal{P}(p)$ is **semi-uniform** *if there exists $M > 0$ such that for all p,*

$$\max_{I \in \mathcal{P}(p)} |I| \leq M \frac{p}{|\mathcal{P}(p)|}. \tag{2.11}$$

In other words, no atom is larger than a fixed multiple M of the average atom size.

*Given $\beta \in \mathbb{R}^p$, we say $\mathcal{P}$ is β-**ordered** if, for each distinct $I, J \in \mathcal{P}$, either $\max_{i \in I} \beta_i \leq \min_{j \in J} \beta_j$ or $\max_{j \in J} \beta_j \leq \min_{j \in I} \beta_i$.*

Intuitively, a semi-uniform β-ordered partition $\mathcal{P}(p)$ defines a way to organize the elements β_i^p of β^p into disjoint groups (atoms) that are of similar size, and such that for each group, no element outside the group lies strictly in between two elements of the group.

It is easy to see that many such semi-uniform β-ordered partitions always exist, and are easily constructed if a rank ordering if the betas is known. For example, for each p, first rank order the elements of β^p, then divide the elements into deciles by taking the largest ten percent, then the next ten percent, etc., rounding as needed. The result is ten atoms, and each atom is approximately $p/10$ in size. If in addition we want the number of atoms to tend to infinity with p, we can replace "ten percent" by a percentage that declines toward zero as $p \to \infty$. If instead of ten percent we choose $0 < \alpha < 1/2$ and let the atoms be of size approximately $p^{1-\alpha}$, there will be approximately p^α atoms in the resulting semi-uniform, β-ordered partition $\mathcal{P}(p)$, and the sequence $|\mathcal{P}(p)|$ will be square root dominated.

Once we have such a partition, each atom $A \subset \{1, 2, \ldots, p\}$ defines an anchor point as follows.

Definition 2.3. *For any $A \subset \{1, 2, ..., p\}$ let $1_A \in \mathbb{R}^p$ denote the vector defined by the indicator function of A: $1_A(i) = 0$ if $i \in A$, and otherwise $1_A(i) = 0$. We may then define, for any partition $\mathcal{P} = \mathcal{P}(p)$, an induced linear subspace $L(\mathcal{P})$ of $\mathbb{R}^p$ by*

$$L(\mathcal{P}) = span_p\{1_A \big| A \in \mathcal{P}\} \equiv < 1_A \big| A \in \mathcal{P} > . \tag{2.12}$$

Theorem 2.3. *Let the assumptions **1**,**2**,**3** and **4** hold. Consider a semi-uniform sequence $\{\mathcal{P}(p) : p = 1, 2, 3, \ldots\}$ of β-ordered partitions such that the sequence $\{|\mathcal{P}(p)|\}$ tends to infinity and is square root dominated. Then*

$$\lim_{p \to \infty} ||\hat{h}_{L(\mathcal{P}(p))} - b|| = 0 \quad \textit{almost surely.} \tag{2.13}$$

Theorem 2.3 says that if we have certain prior information about the ordering of the β elements in the sense of finding an ordered partition (but with no other prior information about the actual magnitudes of the elements or their ordering within partition atoms), then asymptotically we can estimate b exactly.

Having in hand a true β-ordered partition *a priori* will usually not be possible because even the ordering of the betas is not likely to be known in practice. However, Theorem 2.3 suggests the hypothesis that partial grouped order information about the betas can still be helpful in improving our estimate of β.

We test this hypothesis in Section 4.2 by considering industry sectors as a proposed way to form a partition of asset betas. To the extent that betas for equities belonging to the same sector are similar, and separated from those of other sectors, the partition will be approximately β-ordered. The experiments of Section 4.2 illustrate, as least in that case, that these approximations can suffice to create a MAPS estimator that improves on the PCA and GPS versions.

2.4. A data-driven dynamic MAPS estimator. Theorem 2.4 of this section shows that even with no *a priori* information about betas beyond the observed time series of returns, we can still use the MAPS framework to improve the GPS estimator by making more efficient use of the data history.

In the analysis above we have treated β as a constant throughout the sampling period, but in reality we expect β to vary slowly over time. To capture this in a simple way, let's now assume that we have access to returns observations for p assets over a fixed number of $2n$ periods. The first n periods we call the first (or previous) time block, and the second n periods the second (or current) time block. We then have returns matrices $R_1, R_2 \in \mathbb{R}^{p\times n}$ corresponding to the two time blocks, and $R = [R_1 R_2] \in \mathbb{R}^{p\times 2n}$ the full returns matrix over the full set of $2n$ observation times.

Define the sample covariance matrices S, S_1, S_2 as $\frac{1}{2n}RR^T$, $\frac{1}{n}R_1R_1^T$, and $\frac{1}{n}R_2R_2^T$, respectively. Let h, h_1, h_2 denote the respective (normalized) leading eigenvectors (PCA estimators) of S, S_1, S_2. (Of the two choices of eigenvector, we always select the one having non-negative inner product with q.)

Instead of a single β for the entire observation period, we suppose there are random vectors β_1 and β_2 that enter the model during the first and second time blocks, respectively, and are fixed during their respective blocks. We assume both β_1 and β_2 satisfy assumptions (1) and (2) above, and denote by b_1 and b_2 the corresponding normalized vectors. The vectors β_1 and β_2 should not be too dissimilar in the mild sense that $(\beta_1, \beta_2) \geq 0$.

Definition 2.4. *Define the co-dispersion $d_p(\beta_1, \beta_2)$ and pointwise correlation $\rho_p(\beta_1, \beta_2)$ of β_1 and β_2 by*

$$d_p(\beta_1, \beta_2) = \frac{1}{p}\sum_{i=1}^{p}\big(\frac{\beta_1(i)}{\mu_p(\beta_1)} - 1\big)\big(\frac{\beta_2(i)}{\mu_p(\beta_2)} - 1\big)$$

and

$$\rho_p(\beta_1, \beta_2) = \frac{d_p(\beta_1, \beta_2)}{d_p(\beta_1)d_p(\beta_2)}.$$

The Cauchy-Schwartz inequality shows $-1 \leq \rho_p(\beta_1, \beta_2) \leq 1$. Furthermore, it is straightforward to verify that

$$(b_1, b_2) - (b_1, q)(b_2, q) = \frac{d_p(\beta_1, \beta_2)}{\sqrt{1 + d_p(\beta_1)^2}\sqrt{1 + d_p(\beta_2)^2}}. \tag{2.14}$$

and hence $d_p(\beta_1, \beta_2)$, and $\rho_p(\beta_1, \beta_2)$ have limits $d_\infty(\beta_1, \beta_2)$, and $\rho_\infty(\beta_1, \beta_2)$ as $p \to \infty$.

The motivation for this model is our expectation that estimated betas are not fixed, but nevertheless recent betas still provide some useful information about current betas. To make this precise in support of the following theorem, we make the following additional assumptions.

A5. [Relation between β_1 and β_2] Almost surely, $(\beta_1, \beta_2) > 0$, $\mu_\infty(\beta_1) = \mu_\infty(\beta_2)$, $d_\infty(\beta_1) = d_\infty(\beta_2)$, and $\lim_{p\to\infty} d_p(\beta_1, \beta_2) = d_\infty(\beta_1, \beta_2)$ exists.

Theorem 2.4. *Assume β_1, β_2, R, X, Z satisfy assumptions 1-5. Denote by $\hat{h}^s_q$ and $\hat{h}^d_q$ the GPS estimators for R_2 and R, respectively, i.e. the current (single) and previous plus current (double) time blocks. Let h_1 and h_2 be the PCA estimators for R_1 and R_2, respectively.*

Let $L_p =< h_1, q >$ and define a MAPS estimator for the current time block as

$$\hat{h}_L = \frac{\tau_p h_2 + \underset{L}{\text{proj}}(h_2)}{||\tau_p h_2 + \underset{L}{\text{proj}}(h_2)||} \quad \text{where} \quad \tau_p = \frac{\psi_p^2 - ||\underset{L}{\text{proj}}(h_2)||^2}{1 - \psi_p^2}, \tag{2.15}$$

where ψ_p is computed from the eigenvalues of the sample covariance matrix corresponding to the current time block R_2. Then, almost surely,

$$\lim_{p\to\infty} \left(||\hat{h}_L - b_2|| - ||\hat{h}^s_q - b_2||]\right) \le 0 \quad \text{and} \quad \lim_{p\to\infty} \left(||\hat{h}_L - b_2|| - ||\hat{h}^d_q - b_2||]\right) \le 0, \tag{2.16}$$

and, if $0 < |\rho_\infty(\beta_1, \beta_2)| < 1$,

$$\lim_{p\to\infty} \left(||\hat{h}_L - b_2|| - ||\hat{h}^s_q - b_2||\right) < 0 \text{ and } \lim_{p\to\infty} \left(||\hat{h}_L - b_2|| - ||\hat{h}^d_q - b_2||\right) < 0. \tag{2.17}$$

Theorem 2.4 says that the MAPS estimator obtained by adding the PCA estimator h from the previous time block as a second anchor point outperforms the GPS estimator asymptotically, as measured by ℓ_2 error, whether the latter is estimated with the most recent time block R_2 or with the full $2n$ (double) data set. This works when the previous time block carries some information about the current beta (non-zero correlation). In the case of perfect correlation $\rho_\infty(\beta_1, \beta_2) = 1$ the two betas are equal, and we then return to the GPS setting where beta is assumed constant across the entire $2n$ observations, so no improved performance is expected.

The cost of implementing this "dynamic MAPS" estimator is comparable to that of the GPS estimator, so should generally be preferred when no rank order information is available for beta.

In this analysis we have chosen to use two historical time blocks of equal length n for the sake of a definite statement and to illustrate the idea. It is likely that the idea also works when the time blocks have different lengths, or when there are multiple historical time blocks in use. Theoretical or experimental analysis could determine rules for making such choices, but we do not do so here.

2.5. Remarks and connections. The theorems above illustrate a general theme of the MAPS framework: the performance of a shrinkage estimator like GPS can be improved when additional information can be added in the form of additional anchor points. For Theorem 2.3, that means a certain amount of prior ordering information about the betas can be converted to anchor points that are good enough to make a bona fide consistent estimator of b. For Theorem 2.4, the use of a PCA estimator from a prior interval in time as an additional anchor point improves the estimator if betas are correlated across time. The general point is that when there is some prior information about the betas that is independent of the time interval used for the estimation, the investigator should formulate that information as one or more anchor points and use the MAPS technique.

This discussion has close connections to Bayesian decision theory (BDT), which makes use of a prior distribution of a parameter to be estimated. One could view the addition of an anchor point in the MAPS framework as an adjustment to a prior distribution for beta.

We think it likely that the MAPS approach can be reformulated in BDT terms, although our results in the current form don't conform to them. We don't formulate the prior information in terms of a prior distribution of the parameters. And since our setting is asymptotic as $p \to \infty$, our conclusions are almost sure statements, rather than statements about minimizing posterior expected loss. However, the structural connections between GPS/MAPS and the James-Stein estimator mentioned in the introduction provides a link. The JS estimator is a kind of empirical Bayes estimator, for example see [11]. Similarly, the GPS/MAPS estimator is an empirical version of an "oracle" estimator – see Section 5.

Another connection, especially for Theorem 2.4, is to the setting of machine learning. Although Theorem 2.4 itself is not about machine learning because there is no training process, one could imagine the use of prior time intervals as input to a training process that finds optimal anchor points as a function of the prior data. This is likely to improve on our default use of the PCA leading eigenvector as additional anchor point.

3. Tracking Error.

Our task has been to estimate the covariance matrix of returns for a large number p of assets but a short time series of n returns observations.

Recall that for the returns model (1.1), under the given assumptions, we have the true covariance matrix

$$\Sigma_b = p\eta bb^T + \delta^2 I,$$

where η and δ are positive constants and b is a unit p-vector, and we are interested in corresponding estimates $\hat{\eta}$, $\hat{\delta}$, and h that define an estimator

$$\Sigma_h = p\hat{\eta}hh^T + \hat{\delta}^2 I.$$

Our focus on the estimator h and relative neglect of $\hat{\eta}$ and $\hat{\delta}$ is justified by Proposition 1.1, showing that the true variance of the estimated minimum variance portfolio $\hat{w}$, and the variance forecast ratio, are asymptotically controlled by h alone through the optimization bias

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

The preceding theorems have focused on a particular measure of estimation error for h: the ℓ_2 error (Euclidean distance) $||h - b|| = 2(1 - (h,b))$. By comparison, [14, 15] focus on the variance forecast ratio of the minimum variance portfolio. This error measure has the benefit of demonstrating improvement of a quantity of direct interest to practitioners, with the drawback of focusing on a single type of portfolio. The ℓ_2 error is not a familiar financial quantity, but is an ingredient in the optimization bias above, and also in estimating tracking error, as we describe next.

We turn to a third important measure of covariance estimation quality: the tracking error for the minimum variance portfolio, which is controlled in part by the ℓ_2 error of h. Tracking error is a term conventionally used in the finance industry as a measure of the distance between a portfolio and its benchmark. Here, we adopt the same idea to measure the distance between an estimated minimum variance portfolio and the true portfolio, as follows.

Recall that w denotes the true minimum variance portfolio using Σ, and $\hat{w}$ is the minimum variance portfolio using the estimated covariance matrix $\hat{\Sigma}$.

Definition 3.1. *The (true) tracking error $\mathcal{T}(h)$ associated to $\hat{w}$ is defined by*

$$\mathcal{T}^2(h) = (\hat{w} - w)^T \Sigma (\hat{w} - w). \tag{3.1}$$

Definition 3.2. *Given the notation above, define the* eigenvector bias $\mathcal{D}(h)$ *associated to a unit leading eigenvector estimate h as*

$$\mathcal{D}(h) = \frac{(h,q)^2(1-(h,b)^2)}{(1-(h,q)^2)(1-(b,q)^2)} = \frac{(h,q)^2||h-b||^2}{||h-q||^2||b-q||^2}.$$

Theorem 3.1. *Let h be an estimator of b such that $\mathcal{E}(h) \to 0$ as $p \to \infty$ (such as a GPS or MAPS estimator). Then the tracking error of h is asymptotically (neglecting terms of higher order in $1/p$) given by*

$$\mathcal{T}^2(h) = \eta \mathcal{E}^2(h) + \frac{\delta^2}{p}\mathcal{D}(h) + \frac{C}{p}\mathcal{E}(h), \tag{3.2}$$

where

$$C = \frac{2}{\xi(1+d_\infty^2(\beta))}(\delta^2 + \frac{\eta}{\hat{\eta}}\hat{\delta}^2)$$

and $\xi > 0$ is a constant depending only on ψ_∞, $\mu_\infty(\beta)$, and $d_\infty(\beta)$.

We consider what this theorem means for various estimators h. For the PCA estimate, it was already shown in [14] that $\mathcal{E}(h_{PCA})$ is asymptotically bounded below, and hence so is the tracking error.

On the other hand, $\mathcal{E}(h_{GPS})$ tends to zero as $p \to \infty$. In addition [14] shows that

$$\limsup_{p\to\infty} p\, \mathcal{E}^2(h_{GPS}) = \infty$$

almost surely, while [17] shows

$$\limsup_{p\to\infty} \frac{p\, \mathcal{E}^2(h_{GPS})}{\log\log p} < \infty,$$

and we conjecture the same is true for the more general estimator h_{MAPS}.

This implies the leading terms, asymptotically, are

$$\mathcal{T}^2(h_{MAPS}) \leq \eta \mathcal{E}^2(h_{MAPS}) + (\delta^2/p)\mathcal{D}(h_{MAPS})$$

Note here the estimated parameters $\hat{\eta}$ and $\hat{\delta}$ have dropped out, with the tracking error asymptotically controlled by the eigenvector estimate h alone.

Theorem 3.1 helps justify our interest in the ℓ_2 error results of Theorems 2.3 and 2.4. Reducing the ℓ_2 error $||h-b||$ of the h estimate controls the second term $\mathcal{D}(h)$ of the asymptotic estimate for tracking error. We therefore expect to see improved total tracking error when we are able to make an informed choice of additional anchor points in forming the MAPS estimator. This is borne out in our numerical experiments described in Section 4.

Proof of Theorem 3.1

Lemma 3.2. *There exists $\xi > 0$, depending only on ψ_∞, $\mu_\infty(\beta)$, and $d_\infty(\beta)$, such that for any p sufficiently large, and any linear subspace L of $\mathbb{R}^p$ that contains q,*

$$||h_L - q||^2 > \xi > 0,$$

where h_L is the MAPS estimator determined by L.

The Lemma follows from the fact that $(h_L, q) \leq (h_{GPS}, q)$, and is proved for the case h_{GPS} using the definitions and the known limits

$$(h_{PCA}, q)_\infty = (b, q)_\infty (h_{PCA}, b)_\infty \tag{3.3}$$

$$(b, q)^2_\infty = \frac{1}{1 + d^2_\infty(\beta)} \in (0, 1) \tag{3.4}$$

$$(h_{PCA}, b)_\infty = \psi_\infty > 0. \tag{3.5}$$

From the Lemma and equation (3.4), we may assume without loss of generality that $\xi > 0$ is an asymptotic lower bound for both $||h_L - q||^2 = 1 - (h_L, q)^2$ and $||b - q||^2 = 1 - (b, q)^2$.

Next, we recall it is straightforward to find explicit formulas for the minimum variance portfolios w and $\hat{w}$:

$$w = \frac{1}{\sqrt{p}} \frac{\rho q - b}{\rho - (b, q)}, \qquad \text{where } \rho = \frac{1 + k^2}{(b, q)}, \quad k^2 = \frac{\delta^2}{p\eta} \tag{3.6}$$

and

$$\hat{w} = \frac{1}{\sqrt{p}} \frac{\hat{\rho} q - h}{\hat{\rho} - (h, q)}, \qquad \text{where } \hat{\rho} = \frac{1 + \hat{k}^2}{(h, q)}, \quad \hat{k}^2 = \frac{\hat{\delta}^2}{p\hat{\eta}}. \tag{3.7}$$

We may use these expressions to obtain an explicit formula for the tracking error:

$$\begin{aligned}
\mathcal{T}^2(h) &= (\hat{w} - w)^T \Sigma (\hat{w} - w) = (\hat{w} - w)^T (p\eta b b^T + \delta^2 I)(\hat{w} - w) \\
&= p\eta(\hat{w} - w, b)^2 + \delta^2 ||\hat{w} - w||^2.
\end{aligned}$$

We now estimate the two terms on the right hand side separately.

(1) For the first term $p\eta(\hat{w} - w, b)^2$, it is convenient to introduce the notation

$$\Gamma = \frac{k^2}{1 + k^2 - (b, q)^2} \text{ and } \hat{\Gamma} = \frac{\hat{k}^2}{1 + \hat{k}^2 - (h, q)^2},$$

and since

$$\Gamma \le \frac{k^2}{\xi} \text{ and } \hat\Gamma \le \frac{\hat k^2}{\xi}$$

both Γ and $\hat\Gamma$ are of order $1/p$.

A straightforward computation verifies that

$$(w, b) = \frac{1}{\sqrt{p}}\Gamma(b, q) \tag{3.8}$$

$$(\hat w, b) = \frac{1}{\sqrt{p}}\left(\mathcal{E}(h) + \hat\Gamma[(b, q) - \mathcal{E}(h)]\right). \tag{3.9}$$

We then obtain

$$p(\hat w - w, b)^2 = p[(\hat w, b) - (w, b)]^2 \tag{3.10}$$

$$= \mathcal{E}(h)^2 + 2\mathcal{E}(h)G + G^2, \tag{3.11}$$

where $G = \hat\Gamma((b, q) - \mathcal{E}(h)) - \Gamma(b, q)$.

Since asymptotically (b, q) is bounded below and $\mathcal{E}(h) \to 0$, the third term G^2 is of order $1/p^2$ and can be dropped. We thus obtain the asymptotic estimate

$$p(\hat w - w, b)^2 \le \mathcal{E}^2 + 2\mathcal{E}(h)(\hat\Gamma - \Gamma)(b, q).$$

Multiplying by η and using the bounds on $\Gamma, \hat\Gamma$ and the limit of (b, q), we obtain

$$p\eta(\hat w - w, b)^2 \le \mathcal{E}^2 + \frac{C}{p}\mathcal{E}(h),$$

where C is the constant defined in the statement of the theorem.

(2) We now turn to the second term $||\hat w - w||^2 = ||\hat w||^2 + ||w||^2 - 2(\hat w, w)$.

Using the definitions of $\hat w$ and w and the fact that k^2, $\hat k^2$ are of order $1/p$, after a calculation we obtain, to lowest order in $1/p$,

$$p||\hat w - w||^2 = \frac{(h, q)^2[1 - (h, b)^2]}{(1 - (h, q)^2)(1 - (b, q)^2)} + \frac{1 - (h, q)^2}{1 - (b, q)^2}\mathcal{E}^2(h). \tag{3.12}$$

Since $\mathcal{E}(h) \to 0$, we may neglect the second term, and putting (1) and (2) together yields

$$\mathcal{T}^2(h) \le \mathcal{E}^2 + \frac{C}{p}\mathcal{E}(h) + \frac{\delta^2}{p}\mathcal{D}(h).$$

4. Simulation Experiments. To illustrate the previous theorems and test whether the MAPS estimators can be successful for realistic finite values of p, we present the results of two numerical experiments. In section 4.1, we draw two correlated random vectors β_1 and β_2 in $\mathbb{R}^p$, $p = 500$, with a variable correlation that we control. Returns are generated using β_1 for a first block of observations, then using β_2 for a second block of equal length. These are used to test whether the dynamic MAPS estimator of Theorem 2.4 is successful against GPS (which assumes $\beta_1 = \beta_2$). In addition, since we know the exact ordering of the beta components, we can compare results with a MAPS estimator defined with a beta-ordered partition as in Theorem 2.3.

In section 4.2, we turn to the use of historical CAPM betas for stocks in the S&P500, rather than simulated betas. This allows us to test a MAPS estimator defined by a partition determined by the 11 sectors of the familiar Global Industry Classification Standard of MSCI and S&P. Under the hypothesis that betas for stocks in the same industry sector tend to have similar magnitudes, classification by sector represents a potential approximation to a true (but usually not observable) beta-ordered partition. We test this data-driven MAPS estimator against PCA, GPS, and the consistent MAPS estimator defined with a true beta-ordered partition.

These simple experiments are only proof-of-concept examples illustrating the potential for success. We have not attempted the worthwhile project of systematically studying the possible choices of history length or sector divisions in order to optimize outcomes in real markets.

The Python code used to run these experiments and create the figures is available at `https://github.com/hugurdog/MAPS_NumericalExperiments`.

4.1. Simulated betas with correlation. To model the possibility that the true betas may vary slowly during the time window used for estimation, and as a test for Theorems 2.3 and 2.4, we create a simple two-block simulation model with $p = 500$ stocks in which the true betas are held constant with value $\beta_1 \in \mathbb{R}^p$ during one block of time, and then shift to a second but correlated value β_2 for a subsequent block of time.

Each block has $n = 25$ observations, so the total observation window is of size $2n = 50$ for each of our $p = 500$ stocks. The $p \times n$ returns matrix for the first block is denoted R_1 and for the second R_2, and

$$R_t = \beta_t X_t + Z_t, \quad t = 1, 2, \tag{4.1}$$

where $X_t \in \mathbb{R}^n$ is a vector of the n unobserved common factor returns in block t, and $Z_t \in \mathbb{R}^{p\times n}$ is the matrix of specific returns in block t.

We generate the $p \times n$ matrices R_1 and R_2 from Equation (4.1) by randomly generating β, X, and Z:

- the market returns $X_t(j)$, $j = 1, \ldots, n$, are an iid random sample drawn from a normal distribution with mean 0 and variance $\sigma^2 = 0.16$,
- all components of the asset specific returns $\{Z_t(i, j), i = 1, \ldots, p; j = 1, \ldots, n\}$ are i.i.d. normal with mean 0 and variance $\delta^2 = (.5)^2$, and
- the p-vectors β_1 and β_2 are defined by drawing $\beta, \eta \in \mathbb{R}^p$ independently from a Normal distribution with mean 1 and variance $(.5)^2 I_{p\times p}$, and setting

  $$\beta_1 = \beta \text{ and } \beta_2 = \rho\beta + \sqrt{1-\rho^2}\eta,$$

  where the correlation ρ ranges through values in $\{0, 0.3, 0.6, 1.0\}$.

With this simulated returns data, we compare performance for the following four choices of h:

1. the PCA estimator on the double block $R = [R_1, R_2]$ (PCA)
2. the GPS estimator on the double block $R = [R_1, R_2]$ (GPS)
3. the dynamic MAPS estimator defined on the double block $R = [R_1, R_2]$ by equation (2.15) (Dynamic MAPS)
4. the MAPS estimator on the single block R_2 incorporating knowledge of a beta ordered partition $\mathcal{P}$ as in Theorem 2.3. The partition is constructed by rank ordering the betas and then grouping them into 7 ordered groups of 71, and a small eighth group of the lowest three. (Beta Ordered MAPS)

We report the performance of each of these estimators according to the following two metrics:

- The ℓ_2 error $||b - h||$ between the true normalized beta $b = \frac{\beta}{|\beta|}$ of the current data block R_2 and the estimated unit vector h.
- The tracking error between the true and estimated minimum variance portfolios w and $\hat{w}$:

$$\mathcal{T}^2(\hat{w}) = (\hat{w} - w)^T \Sigma (\hat{w} - w). \tag{4.2}$$

In our double-block context, this tracking error is specified as follows. Σ in (4.2) is the true covariance matrix of the most recent data block R_2:

$$\Sigma = \sigma^2 \beta_2 {\beta_2}^T + \delta^2 I, \tag{4.3}$$

which then also determines the true fully invested minimum variance portfolio w. The estimated minimum variance portfolio $\hat{w}$ is determined by the estimated covariance matrix

$$\hat{\Sigma} = \hat{\sigma}^2 \hat{\beta} \hat{\beta}^T + \hat{\delta}^2 I = (\hat{\sigma}^2 |\hat{\beta}|^2) h h^T + \hat{\delta}^2 I. \tag{4.4}$$

For our comparison, and following the lead of [14], we fix the asymptotically correct values

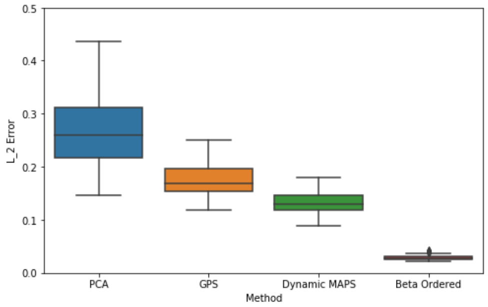
$$\hat{\sigma}^2 |\hat{\beta}|^2 = s_p^2 - l_p^2 \text{ and } \hat{\delta}^2 = \frac{n}{p} l_p^2 \tag{4.5}$$

(notation as in equation 2.7) across each of the four cases, and vary only the estimator $h = \hat{\beta}/|\hat{\beta}|$ as described above. The motivation for this choice is that in our simulation the parameters σ^2 and δ^2 remain constant across the double time window. Hence the best data-driven estimates for $\hat{\sigma}^2$ and $\hat{\delta}^2$ will be obtained by using s_p^2 and l_p^2 computed from the full double block of data R. This puts all the methods compared on the same footing and isolates h as the sole variable in the experiment.

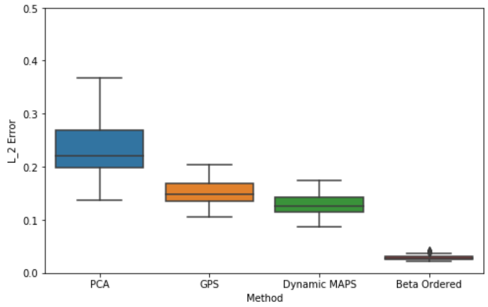
Results of the comparison are displayed below. For each choice of ρ, the experiment was run 100 times, resulting in 100 ℓ_2 error and tracking error values each. These values are summarized using standard box-and-whisker plots generated in Python using the package matplotlib.pyplot.boxplot.

Figure 1 shows the squared ℓ_2 error $||h - b||^2$ for different estimators h (in the same order, left to right, as listed above) for the cases $\rho = 0, 0.3, 0.6, 1.0$. Throughout the range, the dynamical MAPS estimator outperforms the other two data-driven estimators, but the beta-ordered MAPS estimator remains in the lead. The case $\rho = 0$ could be compared to the case of a Bayesian estimator where the additional anchor point is providing information only about the distribution of the components of β. As the correlation ρ tends toward one, the GPS and Dynamic MAPS errors become equal. At $\rho = 1$, $\beta_1 = \beta_2$ and the GPS assumption of constant β over the $2n$ period is satisfied.

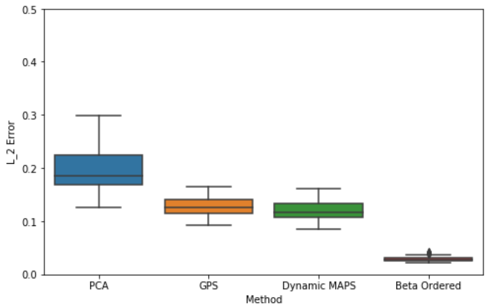
Figure 2 displays the scaled tracking error $p\mathcal{T}^2(h)$ outcomes across a range of correlation values $\rho(\beta_1, \beta_2)$. Dynamic MAPS does best among all data-driven methods, and beta ordered MAPS is significantly better than all others. As before, the Dynamic MAPS lead disappears as ρ tends to 1, when $\beta_1 = \beta_2$.



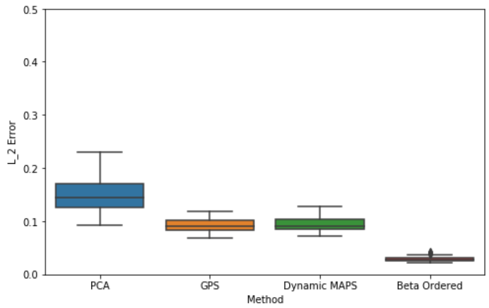
(a) $\rho = 0$



(b) $\rho = 0.3$



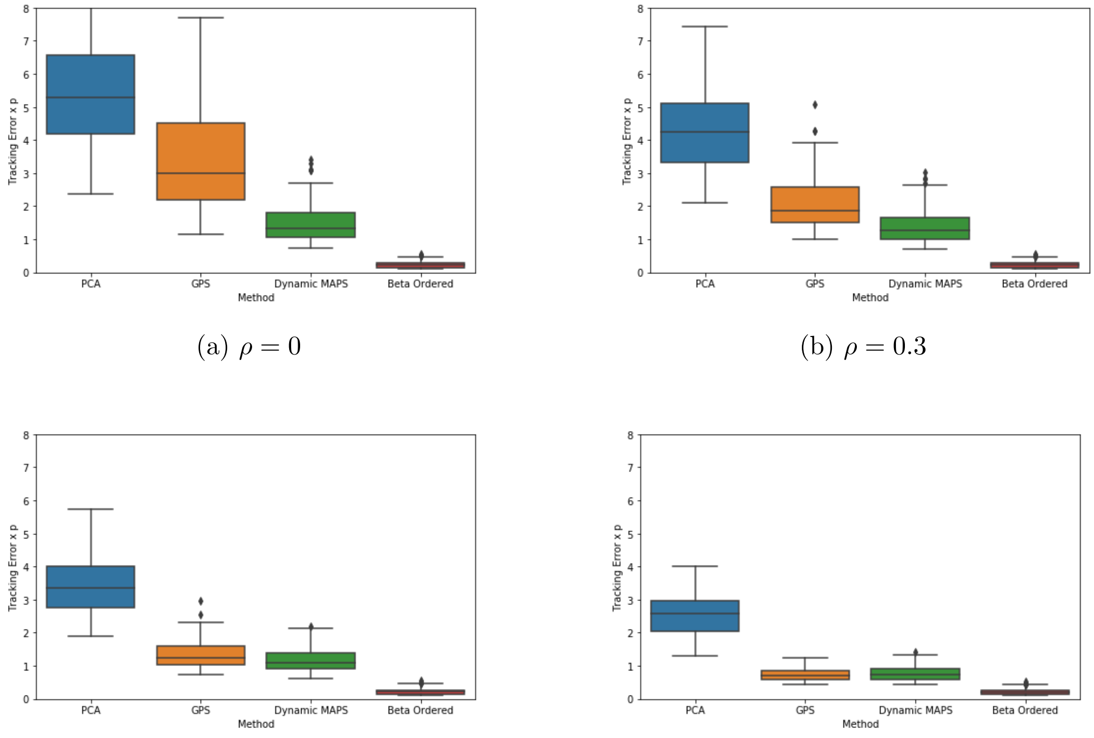
(c) $\rho = 0.6$



(d) $\rho = 1$

Figure 1: Results of simulation experiments measuring ℓ_2 error for different estimators: PCA, GPS, Dynamic MAPS, and Beta Ordered, and varying correlation ρ between betas in the two different time blocks. When beta correlation between time blocks is low, dynamic MAPS outperforms GPS. The non-empirical beta-ordered MAPS outperforms all others.

4.2. Simulations with historical betas. In this section we use historical rather than randomly generated betas to test the quality of MAPS estimators defined using a sector partition and a beta-ordered partition. We use 24 historical monthly CAPM betas for each of the



(a) $\rho = 0$ (b) $\rho = 0.3$

(c) $\rho = 0.6$ (d) $\rho = 1$

Figure 2: Tracking error results of simulation experiments for different estimators PCA, GPS, Dynamic MAPS, and Beta Ordered. The pointwise correlation ρ is the correlation between betas in the two different time blocks. Results are similar to the ℓ_2 error plots.

$p = 488$ S&P500 firms provided by WRDS[4] between the dates 01/01/2018 and 11/30/2020. We denote these betas by $\beta_1, \ldots, \beta_{24} \in \mathbb{R}^p$.

The WRDS beta suite estimates beta each month from the prior 12 monthly returns. Therefore in this experiment we set $n = 12$ months, and using these betas simulate 24 different sets of monthly asset returns $R_t \in \mathbb{R}^{p \times n}$, each for $n = 12$ months.

For each $t = 1, \ldots, 24$, we generate the returns matrix R_t according to

$$R_t = \beta_t X_t + Z_t, \tag{4.6}$$

where the unobserved market return $X_t \in \mathbb{R}^n$ and the asset specific return $Z_t \in \mathbb{R}^{p \times n}$ are generated using the same settings as in the previous section.

For each t we also form partitions $\mathcal{P}_t^{true}$ and $\mathcal{P}_t^{sector}$ of the beta indices $\{1, 2, \ldots, p\}$. $\mathcal{P}_t^{true}$ is a true beta-ordered partition with 11 atoms constructed from the true rank ordering of

[4]Wharton Research Data Services, wrds-www.wharton.upenn.edu

β_t. $\mathcal{P}_t^{sector}$ is a partition defined by the 11 industry sectors[5], which we adopt as a possible data-driven proxy for $\mathcal{P}_t^{true}$.

For each t, we then compute the following four estimators of $b_t = \beta_t/|\beta_t|$:

1. The PCA estimator. (PCA)
2. The GPS estimator. (GPS)
3. The MAPS estimator defined as in Theorem 2.3 using the partition $\mathcal{P}_t^{sector}$. (Sector Separated)
4. The MAPS estimator defined using $\mathcal{P}_t^{true}$. (Beta Ordered)

For each of these four choices of estimator h_t, we examine three different measures of error: the squared ℓ_2 error $||h_t - b_t||^2$, the scaled squared tracking error $p\mathcal{T}^2(h_t)$, and the scaled optimization bias $p\mathcal{E}_p^2(h_t)$.

Since we are interested in expected outcomes, we repeat the above experiment 100 times, and take the average of the errors as a monte carlo estimate of the expectations

$$\mathbb{E}[||h_t - b_t||^2], \quad \mathbb{E}[p\mathcal{T}^2(h_t)], \quad \mathbb{E}[p\mathcal{E}_p^2(h_t)],$$

once for each t. We then display box plots in Figure 3 for the resulting distribution of 24 expected errors of each type, corresponding to the 24 historical betas. Outcomes are similar to the simulated beta experiments, where PCA has the poorest performance, Beta Ordered MAPS the best, and in between are the GPS and empirical MAPS.

Using sectors to partition the stocks evidently has some value, as the sector separated MAPS estimator outperforms GPS by a small but significant amount in both ℓ_2 and tracking error. Its success is owed to the tendency for betas of stocks in a common sector to be closer to each other than to betas in other sectors. The Sector Separated MAPS estimator does not require any information not easily available to the practitioner, and so represents a costless improvement on the GPS estimation method.

We also note that further experiments are reported in [17] and [18], in which a dynamic double-block experiment using the historical betas is also carried out, with similar results.

5. Proofs of the Main Theorems.

The proofs of the main theorems proceed by means of some intermediate results involving an "oracle estimator", defined in terms of the unobservable b but equal to the MAPS estimator in the asymptotic limit (Theorem 5.1 below). Several technical supporting propositions and lemmas are needed; to save space their proofs are collected in a separate document, [18], available online.

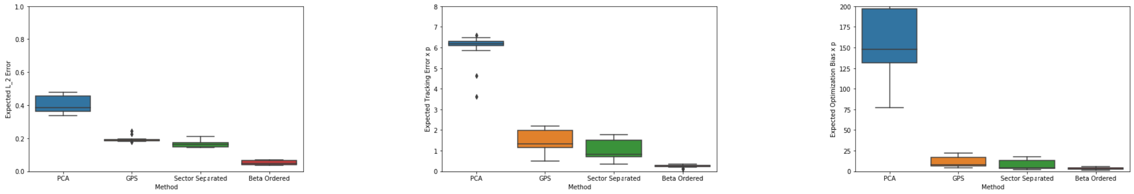
5.1. Oracle Theorems.

A key tool in the proofs is the *oracle estimator* h_L, which is a version of $\hat{h}_L$ but defined in terms of b, our estimation target.

Given a subspace $L = L_p$ of $\mathbb{R}^p$, we define

$$h_L = \frac{\underset{<h,L>}{\text{proj}}\,(b)}{||\underset{<h,L>}{\text{proj}}\,(b)||}. \tag{5.1}$$

[5]The 11 sectors of the Global Industry Classification Standard are: Information Technology, Health Care, Financials, Consumer Discretionary, Communication Services, Industrials, Consumer Staples, Energy, Utilities, Real Estate, and Materials.



(a) ℓ_2 error (b) tracking error (c) optimization bias

Figure 3: Box plots summarizing the distribution of 24 monte carlo-estimated expected errors for the PCA, GPS, Sector Separated, and Beta Ordered estimators (left to right in each figure). The experiment is conducted over 488 S&P 500 companies. This experiment reveals that the Sector Separated estimator is able to capture some of the ordering information and therefore outperforms the GPS estimator. The Beta Ordered estimator performs best.

Here $< h, L >$ denotes the span of h and L, and note that if $L = \{0\}$ we get $h_L = h$, the PCA estimator. A nontrivial example for the selection would be $L_p =< q >$, which generates h_q, the oracle version of the GPS estimator in [14]. The following theorem says that asymptotically the oracle estimator (5.1) converges to the MAPS estimator (2.6).

Theorem 5.1. *Let the assumptions **1**,**2**,**3** and **4** hold. Suppose $\{L_p\}$ be any sequence of random linear subspaces that is independent of the entries of Z, such that $dim(L_p)$ is a square root dominated sequence. Then*

$$\lim_{p\to\infty} ||\hat{h}_L - h_L|| = 0. \tag{5.2}$$

The proof of Theorem 5.1 requires the following proposition, proved in [18].

Proposition 5.2. *Under the assumptions of Theorem 5.1, let $h = h_{PCA}$ be the PCA estimator, equal to the unit leading eigenvector of the sample covariance matrix. Then, almost surely:*

1. $\lim_{p\to\infty} \big((h, \underset{L}{\text{proj}}(h)) - (h, b)^2(b, \underset{L}{\text{proj}}(b))\big) = 0,$
2. $\lim_{p\to\infty} \big((b, \underset{L}{\text{proj}}(h)) - (h, b)(b, \underset{L}{\text{proj}}(b)))\big) = 0,$ *and*
3. $\lim_{p\to\infty} ||\underset{L}{\text{proj}}(h) - (h, b)\underset{L}{\text{proj}}(b)|| = 0.$

In particular, $\frac{\underset{L}{\text{proj}}(h)}{||\underset{L}{\text{proj}}(h)||}$ *converges asymptotically to* $\frac{\underset{L}{\text{proj}}(b)}{||\underset{L}{\text{proj}}(b)||}$.

Proof of the Theorem 5.1:. Recall from (2.6) that,

$$\hat{h}_L = \frac{\tau_p h + \underset{L}{\text{proj}}(h)}{||\tau_p h + \underset{L}{\text{proj}}(h)||} \quad \text{where} \quad \tau_p = \frac{\psi_p^2 - ||\underset{L}{\text{proj}}(h)||^2}{1 - \psi_p^2}.$$

By Lemma 2.1, ψ_p has an almost sure limit $\psi_\infty = (h, b)_\infty \in (0, 1)$, and hence τ_p is bounded in p almost surely.

Let $\Omega_1 \subset \Omega$ be the almost sure set for which the conclusions of Proposition 5.2 hold. Define the notation

$$a_p(\omega) = ||\hat{h}_{L_p} - h_{L_p}||$$

and

$$\gamma_p = \frac{(h,b) - (b, \underset{L}{\text{proj}}(h))}{1 - ||\underset{L}{\text{proj}}(h)||^2}.$$

The proof will follow steps 1-4 below:

1. For every $\omega \in \Omega_1$ and sub-sequence $\{p_k\}_{k=1}^{\infty} \subset \{p\}_1^{\infty}$ satisfying

$$\limsup_{k\to\infty} ||\underset{L_{p_k}}{\text{proj}}(b)||(\omega) < 1$$

we prove

$$0 < \liminf_{k\to\infty} \gamma_{p_k}(\omega) \leq \limsup_{k\to\infty} \gamma_{p_k}(\omega) < \infty$$

and

$$0 < \liminf_{k\to\infty} \tau_{p_k}(\omega) \leq \limsup_{k\to\infty} \tau_{p_k}(\omega) < \infty.$$

2. For every $\omega \in \Omega_1$ and sub-sequence $\{p_k\}_{k=1}^{\infty} \subset \{p\}_1^{\infty}$ satisfying

$$\limsup_{k\to\infty} ||\underset{L_{p_k}}{\text{proj}}(b)||(\omega) < 1$$

we use step 1 to prove $\lim_{k\to\infty} a_{p_k}(w)=0$

3. Set $\Omega_0 = \{\omega \in \Omega \big| \limsup_{p\to\infty} ||\underset{L_p}{\text{proj}}(b)||^2 = 1\}$. Fix $\omega \in \Omega_0 \cap \Omega_1$ and prove using step 2 that $\lim_{p\to\infty} a_p(\omega) = 0$
4. Finish the proof by applying step 2 for all $\omega \in \Omega_0^c \cap \Omega_1$ when $\{p_k\}$ is set to $\{p\}$.

Step 1: Since $\omega \in \Omega_1$ we have the following immediate implications of Proposition 5.2,

$$\limsup_{k\to\infty} ||\underset{L_{p_k}}{\text{proj}}(h)||^2 = (h,b)_\infty^2 \limsup_{k\to\infty} ||\underset{L_{p_k}}{\text{proj}}(b)||^2. \tag{5.3}$$

$$\limsup_{k\to\infty} (b, \underset{L_{p_k}}{\text{proj}}(h)) = (h,b)_\infty \limsup_{k\to\infty} ||\underset{L_{p_k}}{\text{proj}}(b)||^2. \tag{5.4}$$

Using the assumption $\limsup_{k\to\infty} ||\underset{L_{p_k}}{\text{proj}}(b)||^2 < 1$, we update (5.3) and (5.4) as,

$$\limsup_{k\to\infty} ||\underset{L_{p_k}}{\text{proj}}(h)||^2 < (h,b)_\infty^2 < 1 \tag{5.5}$$

(5.6) $$\limsup_{k\to\infty}(b, \operatorname*{proj}_{L_{p_k}}(h)) < (h,b)_\infty$$

for the given $\omega \in \Omega_1$. We can use (5.5) on the numerator of τ_{p_k} to show,

$$\begin{aligned}\liminf_{k\to\infty}\left(\psi^2_{p_k} - ||\operatorname*{proj}_{L_{p_k}}(h)||\right) &\geq \liminf_{k\to\infty}\psi^2_{p_k} - \limsup_{k\to\infty}||\operatorname*{proj}_{L_{p_k}}(h)||^2 \\ &= (h,b)^2_\infty - \limsup_{k\to\infty}||\operatorname*{proj}_{L_{p_k}}(h)||^2 > 0.\end{aligned}$$

That together with the fact that the denominator of τ_{p_k} has a limit in $(0,\infty)$ implies,

(5.7) $$0 < \liminf_{k\to\infty}\tau_{p_k}(\omega) \leq \limsup_{k\to\infty}\tau_{p_k}(\omega) < \infty$$

Similarly we can use (5.6) on the numerator of γ_{p_k} as,

(5.8) $$\liminf_{k\to\infty}\left((h,b) - (b, \operatorname*{proj}_{L_{p_k}}(h))\right) \geq (h,b)_\infty - \limsup_{k\to\infty}(b, \operatorname*{proj}_{L_{p_k}}(h)) > 0.$$

Also (5.5) can be used on the denominator of γ_{p_k} as,

(5.9) $$\liminf_{k\to\infty} 1 - ||\operatorname*{proj}_{L_{p_k}}(h)||^2 > 1 - \limsup_{k\to\infty}||\operatorname*{proj}_{L_{p_k}}(h)||^2 > 0$$

Using (5.8) and (5.9) we get,

(5.10) $$0 < \liminf_{k\to\infty}\gamma_{p_k}(\omega) \leq \limsup_{k\to\infty}\gamma_{p_k}(\omega) < \infty$$

for the given $\omega \in \Omega_1$. This completes the step 1.

Step 2: We have the following initial observation,

(5.11) $$1 \geq ||\operatorname*{proj}_{<h,L_{p_k}>}(b)|| \geq ||\operatorname*{proj}_{<h>}(b)|| = (h,b)$$

and using that we get

$$1 \geq \limsup_{p\to}||\operatorname*{proj}_{<h,L_{p_k}>}(b)|| \geq \liminf_{p\to}||\operatorname*{proj}_{<h,L_{p_k}>}(b)|| \geq (h,b)_\infty > 0.$$

Given that, in order to show $\lim_{k\to\infty} a_{p_k}(\omega) = 0$, it suffices to show $\tau_{p_k}h + \operatorname*{proj}_{L_{p_k}}(h)$ converges to a scalar multiple of $\operatorname*{proj}_{<h,L_{p_k}>}(b)$ since that scalar clears after normalizing the vectors. To motivate that lets re-write $\underset{<h,L_{p_k}>}{\text{proj}}(b)$ as,

$$\begin{aligned}
\underset{<h,L_{p_k}>}{\text{proj}}(b) &= \underset{<h-\underset{L_{p_k}}{\text{proj}}(h),L_{p_k}>}{\text{proj}}(b) \\
&= \underset{L_{p_k}}{\text{proj}}(b) + \left(\frac{h-\underset{L_{p_k}}{\text{proj}}(h)}{||h-\underset{L_{p_k}}{\text{proj}}(h)||}, b\right)\frac{h-\underset{L_{p_k}}{\text{proj}}(h)}{||h-\underset{L_{p_k}}{\text{proj}}(h)||} \\
&= \underset{L_{p_k}}{\text{proj}}(b) + \gamma_{p_k}(h - \underset{L_{p_k}}{\text{proj}}(h)) \qquad (5.12)\\
&= \gamma_{p_k}(h + \frac{1}{\gamma_{p_k}}\underset{L_{p_k}}{\text{proj}}(b) - \underset{L_{p_k}}{\text{proj}}(h)). \qquad (5.13)
\end{aligned}$$

We also have,

$$\tau_{p_k} h + \underset{L_{p_k}}{\text{proj}}(h) = \tau_{p_k}(h + \frac{1}{\tau_{p_k}}\underset{L_{p_k}}{\text{proj}}(h)). \qquad (5.14)$$

Since we have τ_{p_k} and γ_{p_k} satisfying (5.7) and (5.10) respectively, we have the equations (5.13) and (5.14) well defined asymptotically, which is sufficient for our purpose. Hence, from the above argument it is sufficient to show the convergence of $h + \frac{1}{\tau_{p_k}}\underset{L_{p_k}}{\text{proj}}(h)$ to $h + \frac{1}{\gamma_{p_k}}\underset{L_{p_k}}{\text{proj}}(b) - \underset{L_{p_k}}{\text{proj}}(h)$. That is equivalent to showing $\frac{1}{\tau_{p_k}}\underset{L_{p_k}}{\text{proj}}(h)$ converges to $\frac{1}{\gamma_{p_k}}\underset{L_{p_k}}{\text{proj}}(b) - \underset{L_{p_k}}{\text{proj}}(h)$. We can re-write the associated quantity as,

$$\Big|\frac{1}{\tau_{p_k}}\underset{L_{p_k}}{\text{proj}}(h) - \big(\frac{1}{\gamma_{p_k}}\underset{L_{p_k}}{\text{proj}}(b) - \underset{L_{p_k}}{\text{proj}}(h)\big)\Big| = \Big|(1+\frac{1}{\tau_{p_k}})\underset{L_{p_k}}{\text{proj}}(h) - \frac{1}{\gamma_{p_k}}\underset{L_{p_k}}{\text{proj}}(b)\Big| \qquad (5.15)$$

Using Proposition 5.2 part 3 in (5.15), it is equivalent to prove $\Big|(1+\frac{1}{\tau_{p_k}})(h,b) - \frac{1}{\gamma_{p_k}}\Big|$ converges to 0. We re-write it as

$$\begin{aligned}
|(\frac{1}{\tau_{p_k}}+1)(h,b) - \frac{1}{\gamma_{p_k}}| &= \left|\frac{(h,b)(1-||\underset{L_{p_k}}{\text{proj}}(h)||^2}{\psi^2_{p_k} - ||\underset{L_{p_k}}{\text{proj}}(h)||^2} - \frac{1-||\underset{L_{p_k}}{\text{proj}}(h)||^2}{(h,b) - (\underset{L_{p_k}}{\text{proj}}(h),b)}\right| \\
&= |1-||\underset{L_{p_k}}{\text{proj}}(h)||^2|\left|\frac{(h,b)}{\psi^2_{p_k} - ||\underset{L_{p_k}}{\text{proj}}(h)||^2} - \frac{1}{(h,b)-(\underset{L_{p_k}}{\text{proj}}(h),b)}\right| \qquad (5.16)
\end{aligned}$$

Using parts (1) and (2) of Proposition 5.2 and the fact that $\psi^2_{p_k}$ converges to $(h,b)^2_\infty$ shows that (5.16) converges to 0 for the given $\omega \in \Omega_1$. This completes step 2.

Step 3: Fix $\omega \in \Omega_0 \cap \Omega_1$. To show that $\lim_{p\to\infty} a_p(\omega) = 0$, it suffices to show that for any sub-sequence $\{p_k\}_{k=1}^\infty \subset \{p\}_1^\infty$ there exist a further sub-sequence $\{s_t\}_{t=1}^\infty$ such that $\lim_{t\to\infty} a_{s_t}(\omega) = 0$.

Let $\{p_k\}_{k=1}^{\infty}$ be a subsequence. We have one of the following cases,

$$\limsup_{k\to\infty} ||\underset{L_{p_k}}{\text{proj}}(b)||(\omega)^2 < 1$$

or

$$\limsup_{k\to\infty} ||\underset{L_{p_k}}{\text{proj}}(b)||(\omega)^2 = 1$$

If it is strictly less than 1, then we get from the step 2 that $\lim_{k\to\infty} a_{p_k}(\omega) = 0$. In that case we take the further sub-sequence of equal to $\{p_k\}$.

If it is equal to 1, then we get a further sub-sequence $\{s_t\}$ s.t $\lim_{t\to\infty} ||\underset{L_{s_t}}{\text{proj}}(b)||^2 = 1$. Using this and Proposition 5.2 we get the following,

$$\lim_{t\to\infty} ||\underset{L_{s_t}}{\text{proj}}(h)||^2 = (h,b)_\infty^2 \quad \text{and} \quad \lim_{t\to\infty} (b, \underset{L_{s_t}}{\text{proj}}(h)) = (h,b)_\infty$$

which implies $\lim_{t\to\infty} \tau_{s_t}(\omega) = \lim_{t\to\infty} \gamma_{s_t}(\omega) = 0$. Using this on the definition of $\hat{h}_L$ and the equation (5.12) we get,

$$\lim_{t\to\infty} ||\hat{h}_{L_{s_t}} - \frac{\underset{L_{s_t}}{\text{proj}}(h)}{||\underset{L_{s_t}}{\text{proj}}(h)||}|| = 0 \quad \text{and} \quad \lim_{t\to\infty} ||h_{L_{s_t}} - \frac{\underset{L_{s_t}}{\text{proj}}(b)}{||\underset{L_{s_t}}{\text{proj}}(b)||}|| = 0 \tag{5.17}$$

We can now decompose $a_{s_t} = ||\hat{h}_{L_{s_t}} - h_{L_{s_t}}||$ into familiar components via the triangle inequality as follows,

$$a_{s_t} = ||\hat{h}_{L_{s_t}} - h_{L_{s_t}}|| \leq ||\hat{h}_{L_{s_t}} - \frac{\underset{L_{s_t}}{\text{proj}}(h)}{||\underset{L_{s_t}}{\text{proj}}(h)||}|| + ||h_{L_{s_t}} - \frac{\underset{L_{s_t}}{\text{proj}}(b)}{||\underset{L_{s_t}}{\text{proj}}(b)||}||$$
$$+ ||\frac{\underset{L_{s_t}}{\text{proj}}(b)}{||\underset{L_{s_t}}{\text{proj}}(b)||} - \frac{\underset{L_{s_t}}{\text{proj}}(h)}{||\underset{L_{s_t}}{\text{proj}}(h)||}||$$

Using (5.17), we know that the first and the second terms on the right hand side converge to 0 for the given $\omega \in \Omega_0 \cap \Omega_1$. Since we have $\lim_{t\to\infty} ||\underset{L_{s_t}}{\text{proj}}(h)||^2 = (h,b)_\infty^2$ and $\lim_{t\to\infty} ||\underset{L_{s_t}}{\text{proj}}(b)||^2 = 1$, proving the third term on the right hand side converges to 0 is equivalent to proving

$$\lim_{t\to\infty} \left|\left|\underset{L_{s_t}}{\text{proj}}(h) - (h,b)\underset{L_{s_t}}{\text{proj}}(b)\right|\right| = 0,$$

which is true by Proposition 5.2. This completes the step 3.

Step 4: In step 3 we proved the theorem for every $\omega \in \Omega_0 \cap \Omega_1$. Replacing $\{p_k\}$ in step

2 by the whole sequence of indices $\{p\}$, we get the theorem for every $\omega \in \Omega_0^c \cap \Omega_1$. These together shows that we have,

$$\lim_{p\to\infty} a_p(w) = 0 \text{ for all } \omega \in \Omega_1$$

which completes the proof of Theorem 5.1. ■

5.2. Proof of Theorem 2.2. The proof of the first part of Theorem 2.2 is an immediate application of Theorem 5.1.

Proof of the Theorem 2.2(2.8):. From the definitions of h_L and h_q, and as long as $q \in L_p$, we have

$$||h_{L_p} - b|| \leq ||h_q - b||$$

and therefore

$$\begin{aligned} ||\hat{h}_{L_p} - b|| &\leq ||\hat{h}_{L_p} - h_{L_p}|| + ||h_{L_p} - b|| \\ &\leq ||\hat{h}_{L_p} - h_{L_p}|| + ||h_q - b|| \\ &\leq ||\hat{h}_{L_p} - h_{L_p}|| + ||\hat{h}_q - b|| \end{aligned}$$

since $||h_q - b|| \leq ||\hat{h}_q - b||$ for all p. Applying Theorem 5.1 gives ■

$$\limsup ||\hat{h}_{L_p} - b|| \leq \lim_{p\to\infty} ||\hat{h}_q - b||.$$

To prove the remainder of Theorem 2.2 we need the following intermediate result concerning uniform random subspaces, proved in [18].

Proposition 5.3. *Suppose, for each p, z_p is a (possibly random) point in $\mathbb{S}^{p-1}$ and $\mathcal{H}_p$ is a uniform random subspace of $\mathbb{R}^p$ that is independent of z_p. Assume the sequence $\{\dim \mathcal{H}_p\}$ is square root dominated.*

Then

$$\lim_{p\to\infty} ||\operatorname*{proj}_{\mathcal{H}_p}(z_p)||^2 = 0 \textit{ almost surely.}$$

Proof of the Theorem 2.2 (2.9 and 2.10). Theorem 5.1 is applicable. Hence, it suffices to prove the results for the oracle version of the MAPS estimator.

Since the scalars clear after normalization, it suffices to prove the following assertions,

$$\lim_{p\to\infty} ||\operatorname*{proj}_{<h,\mathcal{H}>}(b) - \operatorname*{proj}_{<h>}(b)||_2 = 0 \tag{5.18}$$

and

$$\lim_{p\to\infty} ||\operatorname*{proj}_{<h,q,\mathcal{H}>}(b) - \operatorname*{proj}_{<h,q>}(b)||_2 = 0. \tag{5.19}$$

We first consider (5.18), rewriting the left hand side as

$$\lim_{p\to\infty} ||\operatorname*{proj}_{\mathcal{H}}(b) + \operatorname*{proj}_{h-\operatorname*{proj}_{\mathcal{H}}(h)}(b) - \operatorname*{proj}_{<h>}(b)||_2$$
$$\le ||\operatorname*{proj}_{\mathcal{H}}(b)||_2 + ||\operatorname*{proj}_{h-\operatorname*{proj}_{\mathcal{H}}(h)}(b) - \operatorname*{proj}_{<h>}(b)||_2 \tag{5.20}$$

The first term of (5.20) converges to 0 by setting $z = b$ in Proposition 5.3. Moreover, Propositions 5.3 and 5.2 imply $\operatorname*{proj}_{\mathcal{H}}(h)$ converges to the origin in the ℓ_2 norm. Hence we have $h - \operatorname*{proj}_{\mathcal{H}}(h)$ is converging to h in ℓ_2 norm. That implies the second term in (5.20) converges to 0, which in turn proves (5.18).

Next, rewrite the expression in the assertion (5.19) as,

$$||\operatorname*{proj}_{\mathcal{H}}(b) + \operatorname*{proj}_{<h-\operatorname*{proj}_{\mathcal{H}}(h),q-\operatorname*{proj}_{\mathcal{H}}(q)>}(b) - \operatorname*{proj}_{<h,q>}(b)||$$
$$\le ||\operatorname*{proj}_{\mathcal{H}}(b)|| + ||\operatorname*{proj}_{<h-\operatorname*{proj}_{\mathcal{H}}(h),q-\operatorname*{proj}_{\mathcal{H}}(q)>}(b) - \operatorname*{proj}_{<h,q>}(b)|| \tag{5.21}$$

Similarly the first term of (5.21) converges to 0 by Proposition 5.3. Note that 5.3 also applies when we set $z = q$, and hence $\operatorname*{proj}_{\mathcal{H}}(q)$ converges to the origin in the ℓ_2 norm. Hence the basis elements of $< h - \operatorname*{proj}_{\mathcal{H}}(h), q - \operatorname*{proj}_{\mathcal{H}}(q) >$ converge to the basis elements of $< h, q >$, which implies the second term of (5.21) converges to 0 as well. That completes the proof. ■

5.3. Proof of Theorem 2.3.

We need the following lemma.

Lemma 5.4. *Let $\mathcal{P}(p)$ be a sequence of uniform β-ordered partitions such that $\lim_{p\to\infty} |\mathcal{P}(p)| = \infty$. Then for $L_p = L(\mathcal{P}(p))$ we have,*

$$\lim_{p\to\infty} ||\operatorname*{proj}_{L}(b)|| = 1 \tag{5.22}$$

almost surely.

Proof. To be more precise about $L = L(\mathcal{P})$, set $\mathcal{P}(p) = \{I_1, I_2, ..., I_{k_p}\}$ and denote the defining basis of the corresponding subspace $L_p = L(\mathcal{P})$ by the orthonormal set $\{v_1, v_2, ..., v_{k_p}\}$.

Then

$$1 - ||\underset{L}{\text{proj}}(b)||^2 = 1 - \lim_{p\to\infty} \sum_{i=1}^{k_p} (b, v_i)^2$$

$$= \sum_{i=1}^{p} b_i^2 - \lim_{p\to\infty} \sum_{i=1}^{k_p} (b, v_i)^2$$

$$= \lim_{p\to\infty} \frac{1}{||\beta||^2} \sum_{i=1}^{k_p} (\sum_{j\in I_i} \beta_j^2 - \frac{1}{|I_i|} (\sum_{j\in I_i} \beta_i)^2)$$

$$= \lim_{p\to\infty} \frac{1}{||\beta||^2} \sum_{i=1}^{k_p} (\sum_{j\in I_i} (\beta_j - \frac{1}{|I_i|} (\sum_{j\in I_i} \beta_i))^2 \tag{5.23}$$

Now define the random variables $a_i = \underset{j\in I_i}{max}(\beta_j)$, $c_i = \underset{j\in I_i}{min}(\beta_j)$ for all $1 \le i \le k_p$. Without loss of generality, $c_{k_p} \le a_{k_p} \le ... \le c_1 \le a_1$. Since the sequence $\{\mathcal{P}(p)\}$ is uniform, there exists $M > 0$ such that

$$\max_{I\in\mathcal{P}(p)} |I| \le \frac{Mp}{|\mathcal{P}(p)|}. \tag{5.24}$$

Then

$$\lim_{p\to\infty} \frac{1}{||\beta||^2} \sum_{i=1}^{k_p} (\sum_{j\in I_i} (\beta_j - \frac{1}{|I_i|} (\sum_{j\in I_i} \beta_i))^2 \le \lim_{p\to\infty} \frac{1}{||\beta||^2} \sum_{i=1}^{k_p} |I_i|(a_i - c_i)^2$$

$$\le \lim_{p\to\infty} \frac{\frac{Mp}{k_p}}{||\beta||^2} \sum_{i=1}^{k_p} (a_i - c_i)^2 \tag{5.25}$$

$$= \lim_{p\to\infty} \frac{M}{\frac{||\beta||^2}{p}} \frac{1}{k_p} (a_1 - c_{k_p})^2 \tag{5.26}$$

The term $a_1 - c_{k_p}$ appearing in (5.26) is uniformly bounded since the β's are uniformly bounded. Also, $\frac{||\beta||^2}{p}$ is finite and away from zero asymptotically. Using those together with the fact that $\lim_{p\to\infty} k_p = \infty$ we get the limit in (5.26) equal to 0 for any realization of the random variables β. Note that this is stronger than almost sure convergence. ∎

Proof of the Theorem 2.3:. By an application of Theorem 5.1 it suffices to prove the theorem for the oracle version of the MAPS estimator. Now

$$||b - \underset{<h,L>}{\text{proj}}(b)||^2 \le ||b - \underset{L}{\text{proj}}(b)||^2 = 1 - ||\underset{L}{\text{proj}}(b)||^2 \tag{5.27}$$

and note that application of Lemma 5.4 shows that $||\underset{L}{\text{proj}}(b)||$ converges to 1 as p tends to ∞. ∎

5.4. Proof of Theorem 2.4. The proof of Theorem 2.4 requires the following proposition, from which the first part (2.16) of the theorem easily follows. The proof of the proposition, along with the more difficult proof of the the strict inequality (2.17), appears in [18].

Recall that h_1, h_2 and h are the PCA leading eigenvectors of the sample covariance matrices of the returns R_1, R_2 and R, respectively.

Proposition 5.5. *For each p there is a vector $\tilde{h}$ in the linear subspace $L \subset R^p$ generated by h_1 and h_2 such that $\lim_{p\to\infty} ||\tilde{h} - h|| = 0$ almost surely.*

Proof of (2.16) of Theorem 2.4. Since $dim(L_p) = 2$ and $L_p = span(h_1, q)$ is independent of the asset specific portion Z_2 of the current block, Theorem 2.1 implies that $\hat{h}_L$ converges to h_L almost surely in ℓ_2 norm. Hence it suffices to establish the result for the oracle versions of the MAPS and the GPS estimators.

Note

$$(h_L, b) = ||\operatorname*{proj}_{span(q,h_1,h_2)}(b)|| \tag{5.28}$$

$$(h_q^s, b) = ||\operatorname*{proj}_{span(q,h_2)}(b)|| \tag{5.29}$$

$$(h_q^d, b) = ||\operatorname*{proj}_{span(q,h)}(b)|| \tag{5.30}$$

Using Proposition 5.5 we know there exist $\tilde{h} \in span(h_1, h_2)$ such that $\tilde{h}$ converges to h in l_2 almost surely. Since $span(q, \tilde{h}) \subset span(q, h_1, h_2)$,

$$||\operatorname*{proj}_{span(q,h_1,h_2)}(b)|| \geq ||\operatorname*{proj}_{span(q,\tilde{h})}(b)||.$$

Taking the limits of both sides we get

$$\lim_{p\to\infty}(h_L, b) = \lim_{p\to\infty} ||\operatorname*{proj}_{span(q,h_1,h_2)}(b)|| \geq \lim_{p\to\infty} ||\operatorname*{proj}_{span(q,h)}(b)|| = \lim_{p\to\infty}(h_q^d, b). \tag{5.31}$$

Similarly, since $span(q, h_1) \subset span(q, h_1, h_2)$,

$$\lim_{p\to\infty}(h_L, b) = \lim_{p\to\infty} ||\operatorname*{proj}_{span(q,h_1,h_2)}(b)|| \geq \lim_{p\to\infty} ||\operatorname*{proj}_{span(q,h_1)}(b)|| = \lim_{p\to\infty}(h_q^d, b). \tag{5.32}$$

Inequalities (5.31) and (5.32) complete the proof of Theorem 2.4(a). ■

6. Open Questions. The MAPS approach to estimation of eigenvectors in a factor model setting is flexible because it allows for a general way to inject additional information, in the form of additional anchor points, to improve the estimate. Yet in this paper we have focused on a very simple setting in order to highlight the ideas: a one-factor model with homogeneous specific risk. Moreover, our error measures related to portfolio optimization – tracking error and variance forecast ratio – have focused on the performance of the minimum variance portfolio (motivated by [14]).

Here are a few directions for ongoing and future research.

- How effective can MAPS estimators be in the context of multifactor models, and with variable specific risk? In that setting what are more general connections between ℓ_2 error of betas and tracking error of optimal portfolios?
- What is the general relationship between optimal MAPS shrinkage targets and the linear constraints in a portfolio optimization problem?
- What appropriate systematic empirical tests would be most useful in evaluating MAPS for practical implementation?
- The MAPS approach is general and does not depend on the specific choices of anchor points analyzed here. Are there other useful sets of anchor points, for example possibly excluding the vector q? What other sources of observable information in the market translate into useful anchor points for a successful MAPS estimation of beta? A simple extension of Theorem 2.4 would involve the use of multiple past time blocks to create multiple anchor points, for example.
- The experiments of Section 4.2 involving historical betas and partitions defined by industry sectors had the advantage that sectors define an *a priori* partition that doesn't require unobservable information. This is only one way that a β-ordered partition might be approximated. Another possibility could be to use historical volatilities to form a rank ordering and subsequent partition and anchor points. However, since volatilities are correlated with historical betas, adding volatility anchor points and then computing ℓ_2 error against historical betas would be an unfair test. Instead, a different experiment could be designed using some out-of-sample measure of success in place of the ℓ_2 error.
- The selection of a shrinkage target from observable data may be suited to a machine learning approach to covariance estimation. One or more anchor points could be the output of a trained neural network that could in principle be fed with a much larger universe of observable data than simply the history of returns. This could potentially take the eigenvector shrinkage approach into a much wider realm of applicability.

REFERENCES

[1] R. Almgren and N. Chriss, *Optimal portfolios from ordering information*, J. of Risk, 9 (2006), pp. 1–47.

[2] P. J. Bickel and E. Levina, *Covariance regularization by thresholding*, The Annals of Statistics, (2008), pp. 2577–2604.

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

[4] R. Clarke, H. De Silva, and S. Thorley, *Minimum-variance portfolio composition*, Journal of Portfolio Management, 2 (2011), pp. 31–45.

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

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

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

[8] N. E. El Karoui, *Spectrum estimation for large dimensional covariance matrices using random matrix theory*, The Annals of Statistics, (2008), pp. 2757–2790.

[9] J. Fan, Y. Fan, and J. Lv, *High dimensional covariance matrix estimation using a factor model*, Journal of Econometrics, 147 (2008), pp. 186–197.

[10] J. Fan, Y. Liao, and M. Mincheva, *Large covariance estimation by thresholding principal orthogonal complements*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75 (2013), pp. 603–680.
[11] D. Fourdrinier, W. Strawderman, and M. Wells, *Shrinkage Estimation*, Springer, 2018.
[12] P. A. Frost and J. E. Savarino, *An empirical bayes approach to efficient portfolio selection*, Journal of Financial and Quantitative Analysis, 21 (1986), pp. 293–305.
[13] L. Goldberg and A. Kercheval, *James stein for eigenvectors*, preprint, (2022).
[14] L. Goldberg, A. Papanicolaou, and A. Shkolnik, *The dispersion bias*, SIAM Journal on Financial Mathematics, (2022). to appear.
[15] L. R. Goldberg, A. Papanicolaou, A. Shkolnik, and S. Ulucam, *Better betas*, Journal of Portfolio Management, 47 (2020), pp. 119–136.
[16] M. Gruber, *Improving efficiency by shrinkage*, CRC Press, 1998.
[17] H. Gurdogan, *Eigenvector Shrinkage for Estimating Covariance Matrices*, PhD thesis, Florida State University, 2021.
[18] H. Gurdogan and A. Kercheval, *Multi anchor point shrinkage for the sample covariance matrix (extended version)*, arXiv 2109.00148, (2021).
[19] P. Hall, J. S. Marron, and A. Neeman, *Geometric representation of high dimension, low sample size data*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2005), pp. 427–444.
[20] W. James and C. Stein, *Estimation with quadratic loss*, Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, (1961), pp. 361–379.
[21] T. L. Lai and H. Xing, *Statistical models and methods for financial markets*, Springer, 2008.
[22] O. Ledoit and M. Wolf, *Improved estimation of the covariance matrix of stock returns with an application to portfolio selection*, Journal of empirical finance, 10 (2003), pp. 603–621.
[23] O. Ledoit and M. Wolf, *Honey, I shrunk the sample covariance matrix*, The Journal of Portfolio Management, 30 (2004), pp. 110–119.
[24] O. Ledoit and M. Wolf, *Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks*, The Review of Financial Studies, 30 (2017), pp. 4349–4388.
[25] H. Markowitz, *Portfolio selection*, The Journal of Finance, 7 (1952), pp. 77–91.
[26] B. Rosenberg, *Extra-market components of covariance in security returns*, Journal of Financial and Quantitative Analysis, 9 (1974), pp. 263–274.
[27] S. A. Ross, *The arbitrage theory of capital asset pricing*, Journal of economic theory, 13 (1976), pp. 341–360.
[28] W. Sharpe, *A simplified model for portfolio analysis*, Management Science, 9 (1963), pp. 277–293.
[29] A. Shkolnik, *James-stein estimation of the first principal component*, Stat, Wiley Online Library, https://doi.org/10.1002/sta4.419 (2021).
[30] O. A. Vasicek, *A note on using cross-sectional information in bayesian estimation of security betas*, The Journal of Finance, 28 (1973), pp. 1233–1239.
[31] W. Wang and J. Fan, *Asymptotics of empirical eigenstructure for high dimensional spiked covariance*, The Annals of Statistics, 45 (2017), pp. 1342–1374.