Multi Anchor Point Shrinkage for the Sample Covariance Matrix*

Hubeyb Gurdogan† and Alec Kercheval‡

This version: September 9, 2021

Abstract.
Estimation of the covariance of a high-dimensional returns vector is well-known by practitioners to be impeded by the lack of long data history. We extend the work of Goldberg, Papanicolaou, and Shkolnik (GPS) [6] 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 – multi 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 precise 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 sampling 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. [2]). See [9] for earlier discussion of the HL regime, and [3] 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 [6] (GPS) to a framework that allows the user to incorporate additional information into the shrinkage target and improve results. Our "multi anchor point shrinkage" (MAPS) approach includes the GPS method as a special case, but improves results when some *a priori* order information about the betas is known.

The problem of sampling error for portfolio optimization has been widely studied ever

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

†Department of Mathematics, Florida State University, Tallahassee, FL (hgurdogan@fsu.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.

since [14] introduced the approach of mean-variance optimization. That paper 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 [6] and references therein. Reducing the number of parameters via factor models has long been standard; see for example [15] and [16]. [17] and [5] initiated a Bayesian approach to portfolio estimation and the efficient frontier. 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. [10], [1], [11], [12], [4]. [13] describes a nonlinear shrinkage of the covariance matrix focused on correcting the eigenvalues, set in the HH asymptotic regime.

The key insight of [6] 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. This is the starting point for the present work.

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 baseline model for financial returns: the one-factor, 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 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.

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 $\mathbf{z}$.

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.

[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 $\beta_\Delta = \Delta^{-1/2}\beta$.

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

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

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 [6] is that, asymptotically as p tends to infinity, 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 ([6]). *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, $c/p \leq o_p \leq C/p$ for all p sufficiently large.

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 [6] 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 [7]). In [6] 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 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 converges exactly to the true vector b in the asymptotic limit.

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. To limit the length of this article, the proofs of some of the needed technical propositions and lemmas appear in a separate document [8], available online. Additional details and computations may be found in [7].

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 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 $d_p(\beta)$ 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 at 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 of each other.

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]

We carry out our analysis with the projection of the vectors on the unit sphere $\mathbb{S}^{p-1} \subset \mathbb{R}^p$. To that end we define

$$b = \frac{\beta}{||\beta||}\ , q = \frac{e}{\sqrt{p}}, \tag{2.3}$$

where $e = e^p = (1,1,\ldots,1) \in \mathbb{R}^p$, and $||.||$ denotes the usual Euclidean norm. With the given assumptions the covariance matrix Σ_β of R, conditional on β, is

$$\Sigma_\beta = \sigma^2\beta\beta^T + \delta^2 I. \tag{2.4}$$

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.

Since b is a leading eigenvector of Σ_β (corresponding to the largest eigenvalue), then the PCA estimator h (the unit leading eigenvector h of the sample covariance matrix $\frac{1}{n}RR^T$) is a natural estimator of b. (We always select the choice of unit eigenvector h such that $(h,q) \geq 0$.)

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

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

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

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 [6] 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 orthogonal projection in $\mathbb{S}^{p-1}$ of b onto the geodesic joining h to 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 $L_p \subset \mathbb{R}^p$ denote a nontrivial proper linear subspace of $\mathbb{R}^p$. We will sometimes drop the dimension p from the notation. Denote by k_p the dimension of L_p, with $1 \leq k_p \leq p-1$.

Let h^p denote the 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 eigenvalues. Then we define the data driven "MAPS" (Multi Anchor Point Shrinkage) estimator by

$$\hat{h}_L = \frac{\tau_p h + \underset{L}{\mathrm{proj}}(h)}{||\tau_p h + \underset{L}{\mathrm{proj}}(h)||} \quad \text{where} \quad \tau_p = \frac{\psi_p^2 - ||\underset{L}{\mathrm{proj}}(h)||^2}{1 - \psi_p^2} \tag{2.6}$$

and

$$\psi_p = \sqrt{\frac{s_p^2 - l_p^2}{s_p^2}} \tag{2.7}$$

is the relative gap between s_p^2 and l_p^2.

Lemma 2.1 ([6]). *The limits*

$$\psi_\infty = \lim_{p\to\infty} \psi_p \text{ and } (h,b)_\infty = \lim_{p\to\infty} (h^p, b^p)$$

exist almost surely, and

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

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 spherical geodesic connecting h to q. The phrase "multi 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 in the added anchor points.

We need the concept of a *random linear subspace* of $\mathbb{R}^p$. Let k_p be a positive integer such that $1 \le k_p \le p-1$. Let ξ^p be an $O(p)$-valued random variable, where $O(p)$ denotes the orthogonal group in $\mathbb{R}^p$. Let $\{e_1^p, e_2^p, \ldots, e_p^p\}$ denote the standard Cartesian basis of $\mathbb{R}^p$.

We say that L_p is a *random linear subspace* of $\mathbb{R}^p$ with dimension k_p if, for some ξ^p as above,

$$L_p = \operatorname{span}_p\{\xi^p e_i^p | i = 1, 2, \ldots, k_p\},$$

where span_p denotes the linear span of a set of vectors in $\mathbb{R}^p$.

We say L_p is independent of a random variable X if the generator ξ^p is independent of X. Moreover, we say H_p is a *Haar random subspace* of $\mathbb{R}^p$ if it is a random linear subspace as above, and the random variable ξ^p induces the (uniform) Haar measure on $O(p)$.

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 $k_p \le Cp^\alpha$ for $\alpha < 1/2$ is square root dominated.

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 Haar random subspace of $\mathbb{R}^p$. Suppose also that L_p and H_p are independent of β and Z, and the sequences $\dim L_p$ and $\dim H_p$ are square root dominated.*

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

Then, almost surely,

$$(a) \limsup_{p\to\infty} ||\hat{h}_{L'} - b|| \le ||\hat{h}_q - b||_\infty, \tag{2.8}$$

$$(b) \lim_{p\to\infty} ||\hat{h}_{L''} - b|| = ||\hat{h}_q - b||_\infty, \text{ and} \tag{2.9}$$

$$(c) \lim_{p\to\infty} ||\hat{h}_{H} - b|| = ||h - b||_\infty. \tag{2.10}$$

Theorem 2.2 says adding random anchor points to form a MAPS estimator does no harm, but also makes no improvement asymptotically. Equation (2.9) says that the GPS estimator is neither improved nor harmed by adding extra anchor points uniformly at random. Therefore the goal will be to find useful anchor points that take advantage of additional information that might be available.

2.3. The MAPS estimator with rank order information about the entries of beta. As with stocks grouped by sector, it may be that the betas can be separated into ordered groups, where the rank ordering of the groups is known, but not the ordering within groups. This turns out to be enough information for the MAPS estimator to converge asymptotically to the true value almost surely.

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,*

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

In other words, no atom is larger than a 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 $\max_{i \in I} \beta_i \le \min_{j \in J} \beta_j$ or $\max_{j \in J} \beta_j \le \min_{j \in I} \beta_i$.*

Definition 2.3. *For any $A \subset \{1, 2, ..., p\}$ define a unit vector $v^A \in \mathbb{R}^p$ by*

$$v^A(i) = 1_A(i) \frac{1}{\sqrt{|A|}}, \tag{2.12}$$

where 1_A denotes the indicator function of A. We may then define, for any partition $\mathcal{P} = \mathcal{P}(p)$, an induced linear subspace $L(\mathcal{P})$ of $\mathbb{R}^p$ by

$$L(\mathcal{P}) = span_p\{v^A \big| A \in \mathcal{P}\} \equiv < v^A \big| A \in \mathcal{P} > . \tag{2.13}$$

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

$$\lim_{p \to \infty} ||\hat{h}_{L(\mathcal{P}(p))} - b|| = 0 \quad \textit{almost surely.} \tag{2.14}$$

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 prior information about the magnitudes of the elements or their ordering within partition atoms), then asymptotically we can estimate b exactly.

Having in hand a genuine ordered partition *a priori* is likely only approximately possible in the real world. Theorem 2.3 is suggestive that even partial grouped order information about the betas can be helpful in strictly improving the GPS estimate. This is confirmed empirically in section 4.

The next theorem shows that even with no *a priori* information beyond the observed time series of returns, we can still use MAPS to improve the GPS estimator.

2.4. A data-driven dynamic MAPS estimator. 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 $\frac{1}{n}R_1R_1^T$, 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

$$(b_1, b_2) - (b_1, q)(b_2, q) = \frac{d_p(\beta_1, \beta_2)}{\sqrt{1 + d_p(\beta_1)^2}\sqrt{1 + d_p(\beta_2)^2}}. \tag{2.15}$$

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 $p \to \infty$.

Our motivation for this model is the intuition that the betas for different time periods are noisy representations of a fundamental beta, and that the beta from a recent time block provides some useful information about beta in the current time block. 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

$$\hat{h}_L = \frac{\tau_p h_2 + \underset{L}{\text{proj}}(h_2)}{||\tau_p h_2 + \underset{L}{\text{proj}}(h_2)||} \quad where \quad \tau_p = \frac{\psi_p^2 - ||\underset{L}{\text{proj}}(h_2)||^2}{1 - \psi_p^2}. \tag{2.16}$$

Then, almost surely,

(a)

$$\lim_{p\to\infty}\left(||\hat{h}_L - b_2|| - ||\hat{h}_q^s - b_2||]\right) \leq 0 \quad \text{and} \quad \lim_{p\to\infty}\left(||\hat{h}_L - b_2|| - ||\hat{h}_q^d - b_2||]\right) \leq 0. \tag{2.17}$$

(b) If $0 < |\rho_\infty(\beta_1, \beta_2)| < 1$ *almost surely,*

$$\lim_{p\to\infty}\left(||\hat{h}_L - b_2|| - ||\hat{h}_q^s - b_2||\right) < 0 \ \text{and} \ \lim_{p\to\infty}\left(||\hat{h}_L - b_2|| - ||\hat{h}_q^d - b_2||\right) < 0. \tag{2.18}$$

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, even if the latter estimated 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, 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.

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

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

The theorems above are about finding an estimator h of b that asymptotically controls the ℓ_2 error $||h - b||$. We are ignoring $\hat{\eta}$ and $\hat{\delta}$ because of 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 via the optimization bias

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

We now turn to another important measure of portfolio estimation quality: the tracking error.

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 [6] 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 [6] shows that

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

almost surely, while [7] 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

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

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 $\hat{w}$:

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

and

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

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

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

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

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

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

A straightforward computation verifies that

$$(w, b) = \frac{1}{\sqrt{p}} \Gamma(b, q) \tag{3.6}$$

$$(\hat{w}, b) = \frac{1}{\sqrt{p}} \left(\mathcal{E}(h) + \hat{\Gamma}[(b, q) - \mathcal{E}(h)] \right). \tag{3.7}$$

We then obtain

$$p(\hat{w} - w, b)^2 = p[(\hat{w}, b) - (w, b)]^2 \tag{3.8}$$

$$= \mathcal{E}(h)^2 + 2\mathcal{E}(h)G + G^2, \tag{3.9}$$

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

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

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

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

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 $1/p$,

$$p||\hat{w} - w||^2 = \frac{(h,q)^2[1-(h,b)^2]}{(1-(h,q)^2)(1-(b,q)^2)} + \frac{1-(h,q)^2}{1-(b,q)^2}\mathcal{E}^2(h). \tag{3.10}$$

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

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

4. Simulation Experiments.

To illustrate the previous theorems, we present the results of numerical experiments showing the improvement that MAPS estimators can bring in estimating the covariance matrix. To approximate the asymptotic regime, in these experiments we use $p = 500$ stocks. 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.

First we set up a test bed consisting of a double block of data where we can control the beta correlation. Set $n = 24$. We generate observations $R_i^1, R_i^2 \in \mathbb{R}^p$ for $i = 1, 2, ..., n$, according to the market model of Equation (1.1):

$$R_i^t = \beta_t X_i^t + Z_i^t \ , \ t = 1,2 \ , \ i = 1,2,...,n \tag{4.1}$$

for unobserved market returns $X_i^t \in \mathbb{R}$ and unobserved asset specific returns $Z_i^t \in \mathbb{R}^p$ for each time window of data.

Here the $p \times n$ matrices R^1 and R^2 represent the previous and current block of n consecutive excess returns, respectively, and are obtained from Equation (4.1) by randomly generating β, X, and Z:

- the market returns X_i^t are an iid random sample drawn from a normal distribution with mean 0 and variance $\sigma^2 = 0.16$,
- the asset specific returns $\{Z_i^1\}_1^n$ and $\{Z_i^2\}_1^n$ are i.i.d. normal with mean 0 and variance $\delta^2 I = (.5)^2 I$, and

- the p-vectors β_1 and β_2 are drawn independently of X and Z from a normal distribution with mean 0 and variance $(.5)^2 I$ and with pointwise correlation $\rho_p(\beta^1, \beta^2) \in [0, 1]$ for a range of values of ρ specified below.[3]

The true covariance matrix of the n most recent returns R^2 is

$$\Sigma = \sigma^2 \beta_2 \beta_2{}^T + \delta^2 I, \tag{4.2}$$

which we wish to estimate by

$$\hat{\Sigma} = \hat{\sigma}^2 \hat{\beta} \hat{\beta}^T + \hat{\delta}^2 I. \tag{4.3}$$

Following the lead of [6], we fix

$$\hat{\sigma}^2 |\hat{\beta}|^2 = s_p^2 - l_p^2 \text{ and } \hat{\delta}^2 = \frac{n}{p} l_p^2 \tag{4.4}$$

and vary only the estimator of $\frac{\hat{\beta}}{|\hat{\beta}|} = h$. In our numerical experiments we compare performance for the following choices of h:

1. h^s the PCA estimator on the single block R^2 (PCA1)
2. h^d the PCA estimator on the double block $R = [R^1, R^2]$ (PCA2)
3. $\hat{h}_q^s$, the GPS estimator on the single block R^2 (GPS1)
4. $\hat{h}_{L_D}$ the dynamical MAPS estimator defined on the double block of data $R = [R^1, R^2]$ by the equation **(18)**.(Dynamical MAPS)
5. $\hat{h}_q^d$, the GPS estimator on the double block $R = [R^1, R^2]$ (GPS2)
6. $\hat{h}_{L(\mathcal{P})}^s$ is the MAPS estimator on single block R^2 where $\mathcal{P}$ is a beta ordered uniform partition constructed by using the ordering of the entries of β^2 and where the number of atoms k_p in each partition is set to 8, which is approximately $\sqrt[3]{p}$.[4] (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 $b = \frac{\beta^2}{|\beta^2|}$ of the current data block and the estimated version $h = \frac{\hat{\beta}}{|\hat{\beta}|}$.
- The tracking error between the true and estimated minimum variance portfolios w and $\hat{w}$

$$\mathcal{T}^2(\hat{w}) = (\hat{w} - w)^T \Sigma (\hat{w} - w). \tag{4.5}$$

Results of the comparison are displayed below for values of the pointwise correlation ρ selected from $\{0, 0.2, .4, .6, .8, 1\}$. For each choice of ρ, the experiment was run 100 times, resulting in 100 ℓ_2 and tracking error values each. These values are summarized using standard box-and-whisker plots generated in Python using the package matplotlib.pyplot.boxplot.

[3] exact recipe for β_1, β_2 here.

[4] The largest 479 beta values are partitioned into 7 groups of 71, and the three lowest values form the eighth partition atom.

Figure 1 shows the ℓ_2 error $||h - b||$ for different estimators h (in the same order, left to right, as listed above) for the case $\rho = 0.2$. The worst performer is the single block PCA. (It is independent of ρ since it doesn't see the earlier data at all.) Double block PCA is a little better, but the other estimators are far better. Since GPS effectively assumes that the betas have perfect serial correlation, it's not surprising that the double block GPS does slightly worse than the single block in this case. The best estimator is the MAPS estimator with prior information about group ordering. Assuming no such information is available, the GPS2 and Dynamical MAPS estimators are about tied for best.

Figure 2 shows the results for $\rho = 0, 0.2, 0.4, 0.6, 0.8, 1.0$ in smaller size for visual comparison. Throughout the range, the dynamical MAPS estimator outperforms all the other purely data-driven estimators, but the beta-ordered MAPS estimator remains in the lead.

Figure 3 presents the results for tracking error, reported as $p\mathcal{T}^2$. Results are similar to the ℓ_2 error, but stronger. Again, the dynamical MAPS estimator does best among all methods that don't use order information, and the beta ordered MAPS estimator is significantly better than all others. Figure 4 displays tracking error outcomes for a range of correlation values $\rho(\beta_1, \beta_2)$.

We conclude from these experiments the dynamical MAPS estimator is best when the only the returns are available, and the beta ordered MAPS estimator is preferred when rank order information on the betas is available.

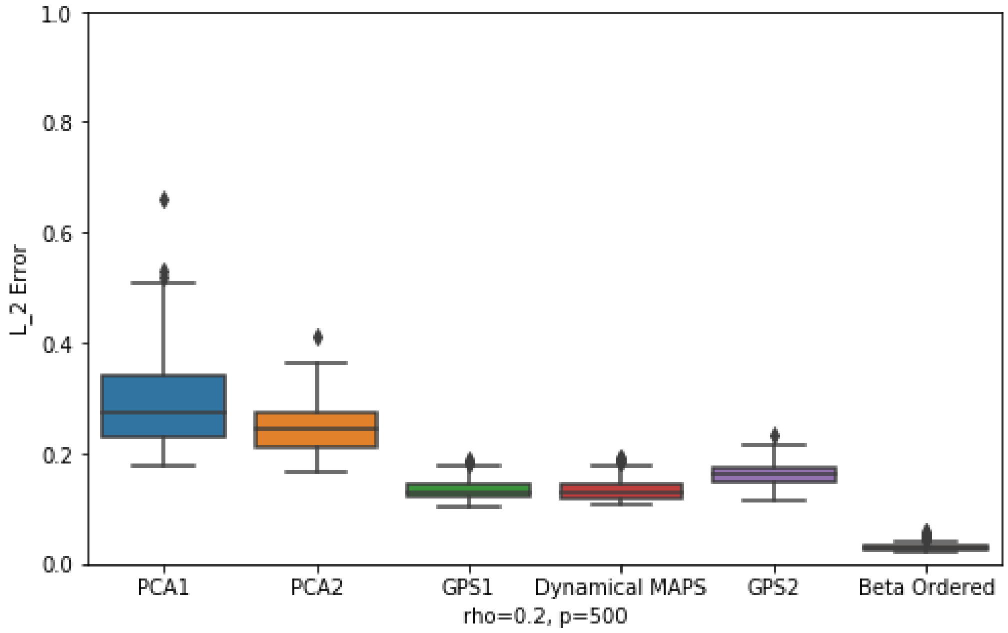


Figure 1: The ℓ_2 error performance of six different estimators as defined in the text. Here the pointwise correlation of betas between the two time blocks is $\rho = 0.2$.

4.2. Simulations with historical betas. In this section we use historical betas rather than randomly generated ones to test the quality of some MAPS estimators. We use 24 historical monthly CAPM betas for each of the S&P 500 firms provided by WRDS[5] between the dates 01/01/2018 and 11/30/2020. We denote these betas as $\beta_1, \ldots, \beta_{24}$. We will have two different mechanism of generating observations of the market model for the single and double data block test beds.

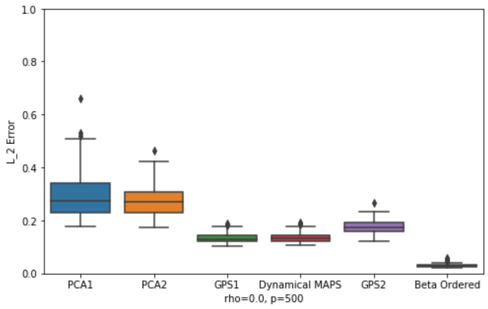
4.2.1. Single Data Block. The WRDS beta suite estimates beta each month from the prior 12 monthly returns. Hence we generate $n = 12$ sequential observations of the market model for each beta separately,

$$R_i^t = \beta_t X_i^t + Z_i^t \quad i = 1, 2, ..., 12, \quad t = 1, 2, \ldots, 24, \tag{4.6}$$

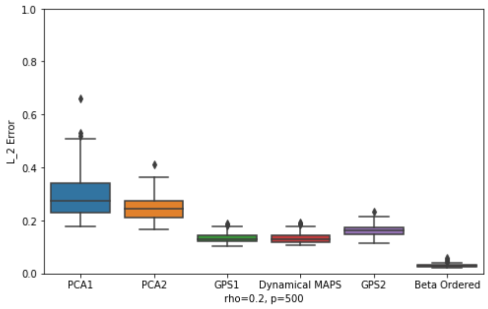
with the unobserved market return X^t and the asset specific return Z^t generated using the same settings as in the previous section.

For each β_t this produces a $p \times n$ returns matrix R^t from which we can derive the following four estimators h^t of β_t:

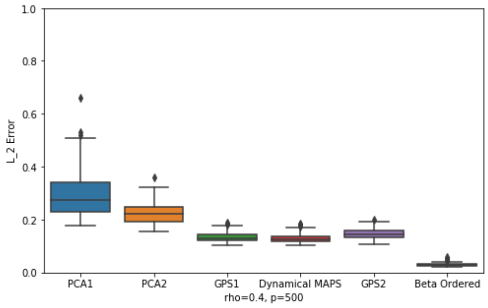
[5] Wharton Research Data Services, wrds-www.wharton.upenn.edu



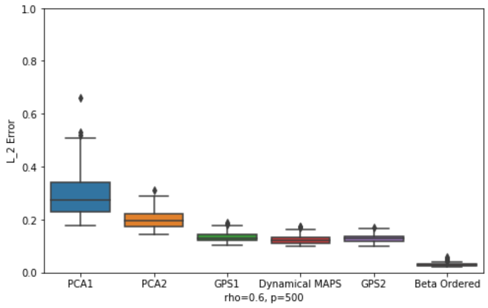
(a) $\rho = 0$



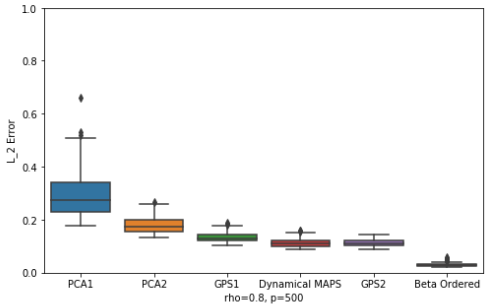
(b) $\rho = 0.2$



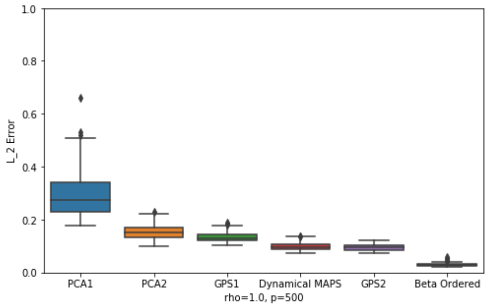
(c) $\rho = 0.4$



(d) $\rho = 0.6$



(e) $\rho = 0.8$



(f) $\rho = 1.0$

Figure 2: Results of simulation experiments for different estimators PCA1, PCA2, GPS1, Random Partition, Dynamical Maps, GPS2, and Beta Ordered. The pointwise correlation ρ is the correlation between betas in the two different time blocks. Figure 2b is the same as Figure 1.

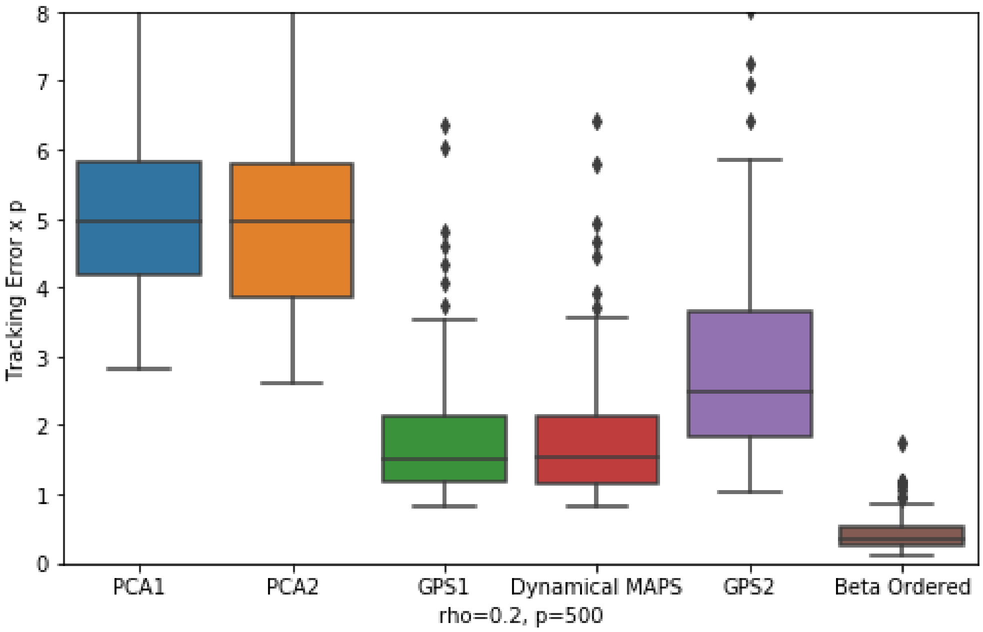


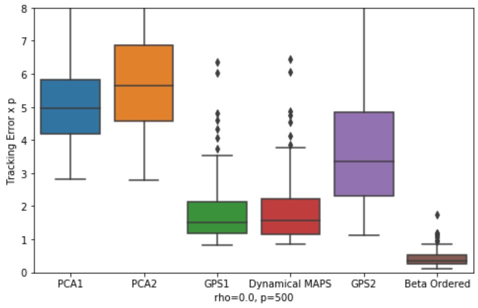
Figure 3: The tracking error error performance of different estimators. Here the pointwise correlation of betas between the two time blocks is $\rho = 0.2$.

1. h^t_{PCA}, the PCA estimator of R^t. (PCA)
2. $\hat{h}^t_q$ the GPS estimator of R^t. (GPS)
3. $\hat{h}^t_{L(\mathcal{P}_s)}$, the MAPS estimator of R^t, where $\mathcal{P}_s$ is a sector partitioning of the indices $\{1, 2, ..., p\}$ in which each atom in the partition contains the indices of one of the 11 sectors in the market[6]. This is one possible data-driven proxy for the beta-ordered uniform partition. (Sector Separated)
4. $\hat{h}^t_{L(\mathcal{P})}$, the MAPS estimator of R^t where $\mathcal{P}$ is a beta ordered uniform partition with 11 atoms constructed by using the true ordering of the entries of β_t. (Beta Ordered)

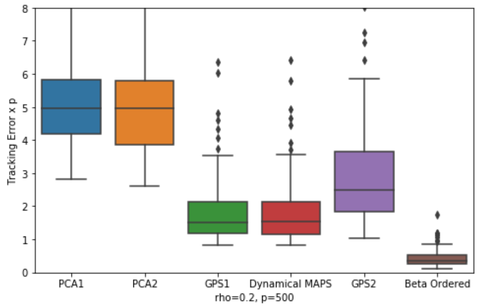
For each of these four choices of estimator h^t, we examine three different measures of error: the squared ℓ_2 error $||h^t - b_t||^2$, the scaled squared tracking error $p\mathcal{T}^2(h^t)$, and the scaled optimization bias $p\mathcal{E}_p^2(h^t)$.

Since we are interested in expected outcomes, we repeat the above experiment 100 times,

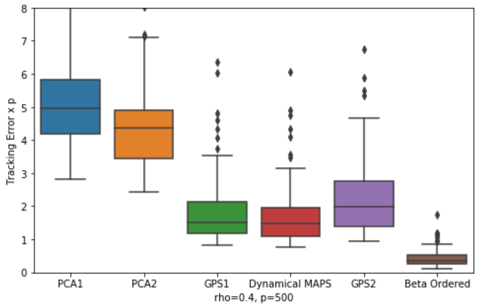
[6]The 11 sectors of the Global Industry Classification Standard are: Information Technology, Health Care, Financials, Consumer Discretionary, Communication Services, Industrials, Consumer Staples, Enery, Utilities, Real Estate, and Materials.



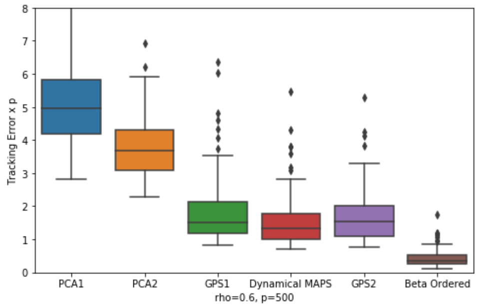
(a) $\rho = 0$



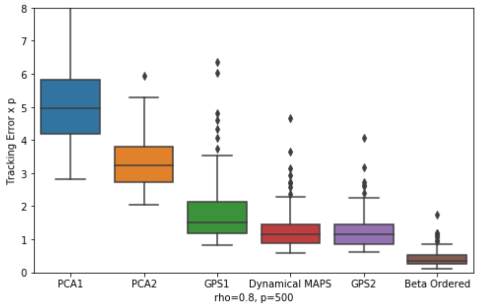
(b) $\rho = 0.2$



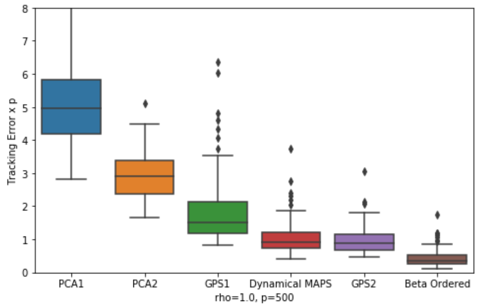
(c) $\rho = 0.4$



(d) $\rho = 0.6$



(e) $\rho = 0.8$



(f) $\rho = 1.0$

Figure 4: Tracking error results of simulation experiments for different estimators PCA1, PCA2, GPS1, Random Partition, Dynamical Maps, GPS2, and Beta Ordered. The pointwise correlation ρ is the correlation between betas in the two different time blocks. Figure 4b is the same as Figure 3.

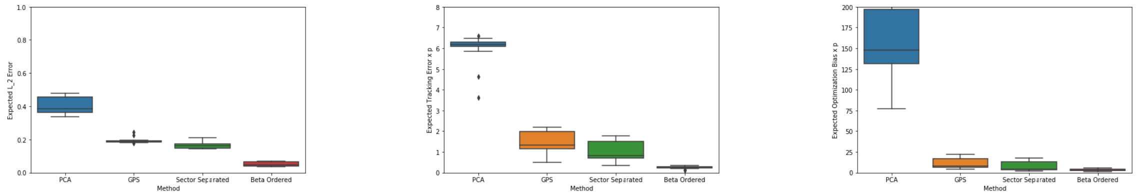
and take the average of the errors as a monte carlo estimate of the expectations

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

once for each t. We then display box plots for the resulting distribution of 24 expected errors of each type, corresponding to the 24 historical betas.

Figure 5 shows a similar story for all three error measures. The GPS estimator significantly outperforms the PCA estimator, and the Beta Ordered estimator, which assumes the ability to rank order partition the betas, is significantly the best.

The result of more interest is that a sector partition approximates a beta ordered partition well enough to improve on the GPS estimate. This approach takes advantage of the fact that betas of stocks in a common sector tend on average 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.



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

Figure 5: 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.

4.2.2. Double Data Block. In order to test the dynamical MAPS estimator that is designed to take advantage of serial correlation in the betas, we will generate a test bed of double data blocks of simulated market observations using the 24 WRDS historical betas for the same time period as before.

For each $t = 1, 2, \ldots, 12$, we generate 12 simulated monthly market returns for β_t and β_{t+12} according to

$$R_i^t = \beta_t X_i^t + Z_i^t, \quad i = 1, 2, ..., 12 \tag{4.7}$$

$$R_i^{t+12} = \beta_{t+12} X_i^{t+12} + Z_i^{t+12}, \quad i = 1, 2, ..., 12 \tag{4.8}$$

were X and Z are generated independently as before.

This provides, for each t, two $p \times 12$ "single block" returns matrices R^t and R^{t+12} each covering 12 months, and a combined "double block" $p \times 2n$ returns matrix $R_t = [R^t R^{t+12}]$ containing 24 consecutive monthly returns of the p stocks.

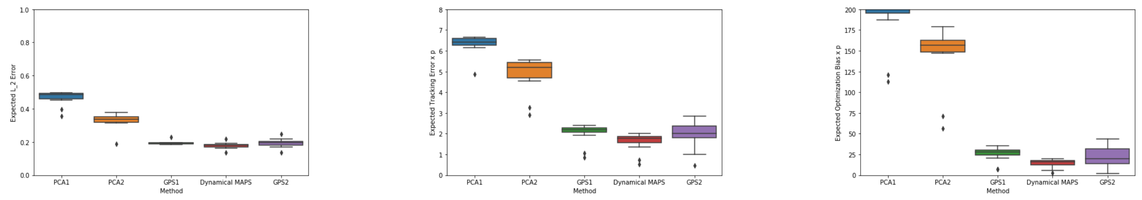
The estimation problem, given observation of the double block of data R_t, is to estimate the normalized beta vector $\beta^{t+12}/||\beta^{t+12}||$ corresponding to the most recent 12 months. This estimate then implies an estimated covariance matrix for that 12 month period according to equations (4.3) and (4.4), and allows us to measure the estimation error as before.

We compare the following estimators:

- h^s_{PCA}, the PCA estimator of R^{t+12}.
- h^d_{PCA}, the PCA estimator of the double block $R_t = [R^t R^{t+12}]$.
- $\hat{h}^s_q$ the GPS estimator of R^{t+12}.
- $\hat{h}^d_q$ the GPS estimator of $R_t = [R^t R^{t+12}]$.
- $\hat{h}_{L_D}$ the dynamical MAPS estimator defined on the double block $R_t = [R^t R^{t+12}]$ by Equation (2.16).

We will report our results using the same three error metrics as before: $\mathbb{E}[|| * - b||^2]$, $\mathbb{E}[p\mathcal{T}^2(*)]$, $\mathbb{E}[p\mathcal{E}^2_p(*)]$ for each of the five estimators. To obtain estimated expectations, we repeat the experiments 100 times and compute the average. The box plots summarize the distribution of the 12 overlapping double block expected errors.

The experiment shows that the Dynamical MAPS estimator outperforms the others, and illustrates the promise of Theorem 2.4, which is based on the hypothesis that betas exhibit some serial correlation. Another benefit of the Dynamical MAPS approach is to relieve the practitioner from the burden of choosing whether to use a GPS1 or GPS2 estimator when a double block of data is available.



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

Figure 6: Box plots for three kinds of expected error for (left to right) the PCA1, PCA2, GPS1, Dynamical MAPS, and GPS2 estimators, summarizing the distribution of 12 different expected errors for each estimator corresponding to 12 consecutive months of 2020. The experiment is conducted using 488 S&P 500 companies.

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, [8], 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

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

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 [6]. 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 [8].

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,

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

$$\limsup_{k\to\infty} (b, \underset{L_{p_k}}{\text{proj}}(h)) < (h,b)_\infty \tag{5.6}$$

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

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

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

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

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

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

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

$$\liminf_{k\to\infty} 1 - ||\underset{L_{p_k}}{\text{proj}}(h)||^2 > 1 - \limsup_{k\to\infty}||\underset{L_{p_k}}{\text{proj}}(h)||^2 > 0 \tag{5.9}$$

Using (5.8) and (5.9) we get,

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

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

Step 2: We have the following initial observation,

$$1 \geq ||\underset{<h,L_{p_k}>}{\text{proj}}(b)|| \geq ||\underset{<h>}{\text{proj}}(b)|| = (h,b) \tag{5.11}$$

and using that we get

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

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

$$\underset{<h,L_{p_k}>}{\text{proj}}(b) = \underset{<h-\underset{L_{p_k}}{\text{proj}}(h),L_{p_k}>}{\text{proj}}(b)$$

$$= \underset{L_{p_k}}{\text{proj}}(b) + \left(\frac{h-\underset{L_{p_k}}{\text{proj}}(h)}{||h-\underset{L_{p_k}}{\text{proj}}(h)||}, b\right)\frac{h-\underset{L_{p_k}}{\text{proj}}(h)}{||h-\underset{L_{p_k}}{\text{proj}}(h)||}$$

$$(5.12) \qquad = \underset{L_{p_k}}{\text{proj}}(b) + \gamma_{p_k}(h - \underset{L_{p_k}}{\text{proj}}(h))$$

$$(5.13) \qquad = \gamma_{p_k}(h + \frac{1}{\gamma_{p_k}}\underset{L_{p_k}}{\text{proj}}(b) - \underset{L_{p_k}}{\text{proj}}(h)).$$

We also have,

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

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,

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

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

$$|(\frac{1}{\tau_{p_k}} + 1)(h, b) - \frac{1}{\gamma_{p_k}}| = \left|\frac{(h, b)(1 - ||\underset{L_{p_k}}{\text{proj}}(h)||^2}{\psi_{p_k}^2 - ||\underset{L_{p_k}}{\text{proj}}(h)||^2} - \frac{1 - ||\underset{L_{p_k}}{\text{proj}}(h)||^2}{(h, b) - (\underset{L_{p_k}}{\text{proj}}(h), b)}\right|$$

$$(5.16) \qquad = |1 - ||\underset{L_{p_k}}{\text{proj}}(h)||^2|\left|\frac{(h, b)}{\psi_{p_k}^2 - ||\underset{L_{p_k}}{\text{proj}}(h)||^2} - \frac{1}{(h, b) - (\underset{L_{p_k}}{\text{proj}}(h), b)}\right|$$

Using parts (1) and (2) of Proposition 5.2 and the fact that $\psi_{p_k}^2$ converges to $(h, b)_\infty^2$ 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,

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

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,

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

$$+ \left|\left|\frac{\underset{L_{s_t}}{\text{proj}}(b)}{||\underset{L_{s_t}}{\text{proj}}(b)||} - \frac{\underset{L_{s_t}}{\text{proj}}(h)}{||\underset{L_{s_t}}{\text{proj}}(h)||}\right|\right|$$

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 Theorem 2.2(a) is an immediate application of Theorem 5.1.

Proof of the Theorem 2.2(a):. 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

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

■

To prove the remainder of Theorem 2.2 we need the following intermediate result concerning Haar random subspaces, proved in [8].

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 Haar random subspace of $\mathbb{R}^p$ that is independent of z_p. Assume the sequence $\{\dim \mathcal{H}_p\}$ is square root dominated.*

Then

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

Proof of the Theorem 2.2(b,c):. 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

$$
\begin{aligned}
&\lim_{p\to\infty} ||\underset{\mathcal{H}}{\mathrm{proj}}(b) + \underset{h-\underset{\mathcal{H}}{\mathrm{proj}}(h)}{\mathrm{proj}}(b) - \underset{<h>}{\mathrm{proj}}(b)||_2 \\
&\leq ||\underset{\mathcal{H}}{\mathrm{proj}}(b)||_2 + ||\underset{h-\underset{\mathcal{H}}{\mathrm{proj}}(h)}{\mathrm{proj}}(b) - \underset{<h>}{\mathrm{proj}}(b)||_2
\end{aligned}
\tag{5.20}
$$

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 $\underset{\mathcal{H}}{\mathrm{proj}}(h)$ converges to the origin in the l_2 norm. Hence we have $h - \underset{\mathcal{H}}{\mathrm{proj}}(h)$ 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,

$$
\begin{aligned}
&||\underset{\mathcal{H}}{\mathrm{proj}}(b) + \underset{<h-\underset{\mathcal{H}}{\mathrm{proj}}(h),q-\underset{\mathcal{H}}{\mathrm{proj}}(q)>}{\mathrm{proj}}(b) - \underset{<h,q>}{\mathrm{proj}}(b)|| \\
&\leq ||\underset{\mathcal{H}}{\mathrm{proj}}(b)|| + ||\underset{<h-\underset{\mathcal{H}}{\mathrm{proj}}(h),q-\underset{\mathcal{H}}{\mathrm{proj}}(q)>}{\mathrm{proj}}(b) - \underset{<h,q>}{\mathrm{proj}}(b)||
\end{aligned}
\tag{5.21}
$$

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 $\underset{\mathcal{H}}{\mathrm{proj}}(q)$ converges to the origin in the l_2 norm. Hence the basis elements of $< h - \underset{\mathcal{H}}{\mathrm{proj}}(h), q - \underset{\mathcal{H}}{\mathrm{proj}}(q) >$ 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,*

$$
\lim_{p\to\infty} ||\underset{L}{\mathrm{proj}}(b)|| = 1 \tag{5.22}
$$

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

$$
\begin{aligned}
1-||\mathrm{proj}_L(b)||^2 &= 1-\lim_{p\to\infty}\sum_{i=1}^{k_p}(b,v_i)^2\\
&=\sum_{i=1}^{p} b_i^2-\lim_{p\to\infty}\sum_{i=1}^{k_p}(b,v_i)^2\\
&=\lim_{p\to\infty}\frac{1}{||\beta||^2}\sum_{i=1}^{k_p}\Big(\sum_{j\in I_i}\beta_j^2-\frac{1}{|I_i|}\big(\sum_{j\in I_i}\beta_i\big)^2\Big)\\
&=\lim_{p\to\infty}\frac{1}{||\beta||^2}\sum_{i=1}^{k_p}\Big(\sum_{j\in I_i}\big(\beta_j-\frac{1}{|I_i|}\big(\sum_{j\in I_i}\beta_i\big)\big)^2 \qquad (5.23)
\end{aligned}
$$

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

$$
\begin{aligned}
\lim_{p\to\infty}\frac{1}{||\beta||^2}\sum_{i=1}^{k_p}\Big(\sum_{j\in I_i}\big(\beta_j-\frac{1}{|I_i|}\big(\sum_{j\in I_i}\beta_i\big)\big)^2 &\le \lim_{p\to\infty}\frac{1}{||\beta||^2}\sum_{i=1}^{k_p}|I_i|(a_i-c_i)^2\\
&\le \lim_{p\to\infty}\frac{\frac{Mp}{k_p}}{||\beta||^2}\sum_{i=1}^{k_p}(a_i-c_i)^2 \qquad (5.25)\\
&= \lim_{p\to\infty}\frac{M}{\frac{||\beta||^2}{p}}\frac{1}{k_p}(a_1-c_{k_p})^2 \qquad (5.26)
\end{aligned}
$$

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

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

and note that application of Lemma 5.4 shows that $||\mathrm{proj}_L(b)||$ converges to 1 as p tends to ∞. ■

5.4. Proof of Theorem 2.4(a). The proof of Theorem 2.4 requires the following proposition, from which part (a) of the theorem easily follows. The proof of the proposition, along with the more difficult proof of the the strict inequality of 2.4 (b), appear in [8].

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 Theorem 2.4(a). 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). ■

REFERENCES

[1] P. J. Bickel and E. Levina, *Covariance regularization by thresholding*, The Annals of Statistics, (2008), pp. 2577–2604.
[2] N. E. El Karoui, *Spectrum estimation for large dimensional covariance matrices using random matrix theory*, The Annals of Statistics, (2008), pp. 2757–2790.

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

[4] 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.

[5] 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.

[6] L. Goldberg, A. Papanicolaou, and A. Shkolnik, *The dispersion bias.* CDAR working paper https://cdar.berkeley.edu/publications/dispersion-bias, 2021.

[7] H. Gurdogan, *Eigenvector Shrinkage for Estimating Covariance Matrices*, PhD thesis, Florida State University, 2021.

[8] H. Gurdogan and A. Kercheval, *Multi anchor point shrinkage for the sample covariance matrix (extended version)*, arXiv 2109.00148, (2021).

[9] 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.

[10] T. L. Lai and H. Xing, *Statistical models and methods for financial markets*, Springer, 2008.

[11] 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.

[12] O. Ledoit and M. Wolf, *Honey, I shrunk the sample covariance matrix*, The Journal of Portfolio Management, 30 (2004), pp. 110–119.

[13] 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.

[14] H. Markowitz, *Portfolio selection*, The Journal of Finance, 7 (1952), pp. 77–91.

[15] B. Rosenberg, *Extra-market components of covariance in security returns*, Journal of Financial and Quantitative Analysis, 9 (1974), pp. 263–274.

[16] S. A. Ross, *The arbitrage theory of capital asset pricing*, Journal of economic theory, 13 (1976), pp. 341–360.

[17] 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.