Universal Portfolio Shrinkage

March 19, 2024

Abstract

We introduce a novel shrinkage methodology for building optimal portfolios in environments of *high complexity*, where the number of assets is comparable to or larger than the number of observations. Our universal portfolio shrinkage approximator (UPSA) is given in closed form, is easy to implement, and improves upon existing shrinkage methods. It exhibits an explicit two-fund separation, complementing the Markowitz portfolio with an optimal *complexity correction.* UPSA does not annihilate the low-variance principal components (PCs) of returns; instead, it optimally *reweights* them and produces a stochastic discount factor that substantially improves on its feasible PC-sparse counterparts.

1 Introduction

Efficient portfolios that optimally balance risk and return play a key role in asset pricing. However, in practically relevant scenarios involving thousands of stocks and hundreds of factors, classical estimators of the (Markowitz, 1952) portfolio are severely contaminated by noise. Despite their stellar in-sample performance, they typically fail out-of-sample and are often dominated by naively diversified portfolios (DeMiguel et al., 2009). The huge wedge between their in-sample (IS) and out-of-sample (OOS) performance is driven by *estimation complexity*. Since the number of parameters entering the portfolio construction typically

exceeds the number of observations, the Law of Large Numbers breaks down (Didisheim et al., 2023).[1]

An established way of reducing the wedge between IS and OOS performance is to optimize the bias-variance tradeoff through shrinkage. However, existing shrinkage methodologies either excessively constrain the admissible forms of shrinkage or target restrictive statistical objectives, such as the estimation error of the covariance matrix. Instead, optimal portfolio shrinkage should be built to target what investors care about: The OOS performance of the Stochastic Discount Factor (SDF). Our Universal Portfolio Shrinkage Approximator (UPSA) is precisely designed to tackle these issues. It is tractable, closed-form, easy to implement, and universal because it encompasses very general forms of shrinkage and easily adapts to the specifics of a particular economic objective.

To understand the nature of optimal shrinkage estimators developed in our paper, we start by noting that the Markowitz portfolio always admits an intuitive decomposition as a portfolio of principal component (PC) returns. Here, each individual portfolio weight is given by each PC's estimated risk-return tradeoff, i.e., the ratio of the PC's average return and sample variance. Inspired by the Arbitrage Pricing Theory (APT) of (Ross, 1976), many papers postulate that only top principal components of asset returns enter the SDF.[2] Intuitively, if risk premia are compensations for systematic risk, only risk factors that explain a large fraction of cross-sectional variation in returns should command non-negligible risk premia. Hence, low-variance PCs should have sufficiently small risk-premia to be safely ignored for the purpose of SDF construction.

As we argue in this paper, the above intuition breaks down when one has to estimate these principal components. Indeed, in realistic situations where the number N of assets is large, estimated low-variance PCs are strongly corrupted by noise. This leads to two conceptually distinct effects. First, even when some "true", unobservable low-variance PCs may offer a very good investment opportunity with a highly attractive risk-return tradeoff, statistical limits to arbitrage (Da et al., 2022) and limits to learning (Didisheim et al., 2023)

[1]For example, when the number of IS periods is smaller than the number of assets, the IS Sharpe ratio of the Markowitz Portfolio is not even finite.

[2]Following (Chamberlain and Rothschild, 1982), this assumption can be rationalized formally whenever the maximum Sharpe ratio portfolio investing in the low-variance PCs has a vanishing variance.

make it impossible to precisely isolate these opportunities out-of-sample. Second, incorrectly estimated in-sample low-variance PCs may also have significant exposure to the "true", unobservable high-variance PCs. In both cases, estimated low-variance PCs might offer important diversification opportunities for generating out-of-sample portfolio performance. Therefore, they should not be neglected in high-complexity environments where the number of assets is comparable to (or even larger than) the number of observations.

Granted that low-variance PCs may offer important diversification opportunities, a natural question is *how should estimated PCs be optimally weighted into a portfolio that delivers the highest out-of-sample economic value*? To tackle this question with a good degree of generality, we start from a broad family of *spectral shrinkage* estimators. These functions transform estimated PC sample variances with some potentially non-linear function f. The first key question we answer in this paper is *how to find the optimal, non-parametric shrinkage function f that maximizes the out-of-sample portfolio performance.* The second related important question we answer is about the *shape of the optimal shrinkage and whether it gives rise to SDFs with nontrivial exposure to low-variance PCs.*

Finding the optimal shrinkage function f without imposing overly restrictive assumptions on the shapes of admissible shrinkages or on the covariance matrix of returns may be challenging. Strikingly, we show that a large class of relevant portfolio shrinkage functions can be efficiently spanned using a tractable Universal Portfolio Shrinkage Approximator (UPSA), which is built from a basis of Ridge-penalized portfolios depending on a set of Ridge penalties.[3] The tractability of UPSA comes from the fact that its shrinkage function is given in closed form and that its computation only depends on the eigenvalues and eigenvectors of the sample covariance matrix of returns. Therefore, it is computationally scalable to even very large datasets. In addition, UPSA can be naturally modified to incorporate further desirable shrinkage features, such as strict positivity and monotonicity. Such a constrained version of UPSA (CUPSA) is built simply by forcing positivity of all weights of the Ridge-penalized portfolios, forming the basis for UPSA.

The monotonicity of CUPSA's shrinkage implies SDFs where the order of the variance

[3]The class of portfolio shrinkage functions universally approximated by UPSA is the class of continuous functions vanishing at infinity.

of each PC is the same as the one before shrinking. It also means that sample covariance matrices associated with a larger estimated risk will imply a larger risk after shrinking. Strict positivity further ensures that no PC gets eliminated after shrinking, i.e., CUPSA produces as Ridge a soft shrinkage thresholding and SDFs that are not sparse. These properties are essential for the ability of CUPSA to leverage information in low-variance PCs efficiently. In contrast, hard shrinkage thresholding violates both monotonicity and strict positivity, giving rise to sparse SDFs that implicitly assign an infinite risk to some low-variance PCs.

By construction, the CUPSA portfolio is equivalent to an optimal allocation, which distributes wealth across a family of Ridge-shrunk portfolio funds and is subject to a short-selling constraint on each fund. Therefore, CUPSA also has the interpretation of a Bayesian optimal portfolio, which extends the standard interpretation of Ridge-shrunk portfolios in the literature. While Ridge-shrunk portfolios are the optimal Bayesian portfolios of an investor with a Gaussian prior with scalar covariance matrix on expected returns, the CUPSA portfolio is the optimal Bayesian portfolio of an investor with a mixture of Gaussian prior on expected returns. In this interpretation, the CUPSA weights in each Ridge-shrunk portfolio are the corresponding probabilities in the Gaussian mixture prior, thus incorporating richer forms of prior uncertainty about expected returns. From this point of view, the superior performance of CUPSA portfolios relative to standard Ridge-shrunk portfolios may also be attributed to their ability to incorporate prior uncertainty about expected returns.

To emphasize the link between CUPSA and SDF estimation in our empirical analysis, we investigate the performance of CUPSA on a large set of managed portfolios from (Jensen et al., 2023), which are commonly thought to span a significant part of the risks in the SDF. We build several natural benchmarks for the CUPSA portfolio. The first is a simple Ridge-shrunk portfolio, in which the Ridge penalty is optimally selected through cross-validation. The second one is a Markowitz portfolio with a covariance matrix shrunk following the classic covariance spectral shrinkage approach in (Ledoit and Wolf, 2017). The third is a portfolio that incorporates both PC sparsity and Ridge shrinkage following (Kozak et al., 2020). In all our experiments, we find that CUPSA achieves a higher OOS Sharpe ratio than these three alternative methods. The associated SDF leads to lower out-of-sample pricing errors (Didisheim et al., 2023). Furthermore, when stratifying the anomalies in (Jensen et al.,

2023) into 13 themes, the CUPSA-SDF achieves significantly lower pricing errors across all of them.

To understand the origins of this out-performance, we use random matrix theory and derive an explicit connection between UPSA and the standard estimator of the efficient portfolio. The difference between the two is determined by the model complexity $c = N/T$ (number of assets/number of observations), as in (Kelly et al., 2022; Didisheim et al., 2023) and the covariance structure of our asset universe. High complexity $c > 0$ is responsible for the breakdown of the law of large numbers, leading to a divergence between the in-sample and out-of-sample moments. We find that CUPSA dominates alternative shrinkage estimators by diversifying and smoothing out the choice of the optimal shrinkage parameter across the multiple Ridge-shrunk portfolios from the UPSA basis.

To account for a potential non-stationarity, we train our shrinkage estimators using a rolling window; as a result, the selected shrinkage parameters fluctuate over time. Interestingly, we find that the time series average of CUPSA weights across Ridge penalties closely emulates the unconditional distribution of the time-varying optimal Ridge penalty across time. The single-Ridge-penalty shrinkage naively attempts to adjust to non-stationarity, with an optimal penalty oscillating substantially throughout the sample. In contrast, CUPSA shrinkage better adjusts to non-stationarity by diversifying and smoothing across multiple Ridge penalties to construct the optimal portfolio shrinkage.

To understand the role of low-variance PCs for portfolio performance, we investigate the behavior of all shrinkage methodologies by gradually increasing the number of estimated PCs used for portfolio construction. We find that the out-of-sample outperformance of the CUPSA SDF[4] monotonically increases with the number of PCs we allow it to incorporate. This monotonic pattern persists even after including the lowest-variance PCs and highlights a novel form of *shrinkage-based virtue of complexity*, consistent with earlier findings in (Kelly et al., 2022; Didisheim et al., 2023).

This pattern of outperformance is corroborated by a hump-shaped dependence of the portfolio Sharpe ratios on the number of PCs that we observe for all shrinkage methods.

[4]Outperformance is measured as the t-statistic of alpha (equivalently, the information ratio) from a multivariate regression of the CUPSA portfolio returns on a battery of alternative shrinkage estimators as well as standard factors.

That is, there exists an *ex-post* optimal number of PCs (henceforth, the *saturation point*) maximizing the out-of-sample portfolio performance. Contrary to previous empirical findings (see, e.g., (Kozak et al., 2020)), this ex-post "degree of PC-non-sparsity" is large (around 30 PCs). The power of the CUPSA methodology becomes apparent from the behavior of Sharpe ratios past the saturation point. We find that the performance of alternative shrinkage methodologies deteriorates quickly once we include lower-variance PCs because they cannot weigh these PCs efficiently. By contrast, UPSA performance stays essentially flat, and including lower-variance PCs does not hurt it. These findings naturally lead to the question: If sparsity is optimal ex-post, can one design an algorithm to find it ex-ante? Our extensive analysis suggests that empirically finding the "right" degree of sparsity that works well out-of-sample is challenging. These results indicate that the conventional APT wisdom advocating SDF–sparsity in the PC space might face significant difficulties associated with estimating the true factor structure of returns in the presence of complexity; as a result, a rational investor might be better off using all PCs combined optimally using smart, non-linear shrinkage.

2 Literature Review

Our work belongs to several large strands of literature that we outline below.

Covariance estimation: By its nature, our UPSA approach builds on a universal approximation of a large class of nonlinear shrinkage procedures for the covariance matrix of asset returns, which is directly applicable for optimizing a portfolio formation criterion. As a by-product of our methodology, we obtain an optimal nonlinear spectral shrinkage estimator of this covariance matrix, in which a corresponding shrinkage function optimizes the given portfolio objective.

Spectral shrinkage estimators of covariance matrices have a long tradition in statistics and have been studied by several authors in a number of important contributions. The main idea of spectral covariance matrix shrinkage was introduced in the influential lecture notes by (Stein, 1986) and is based on estimators that only transform the eigenvalues, but not the eigenvectors, of the sample covariance matrix. A pioneering contribution in this area is the

linear shrinkage estimator of (Ledoit and Wolf, 2004b), with its applications in finance to the estimation of, e.g., minimum variance (Ledoit and Wolf, 2003) and tracking (Ledoit and Wolf, 2004a) portfolios. In a series of further path-breaking papers, (Ledoit and Wolf, 2012), (Ledoit and Wolf, 2015), and (Ledoit and Wolf, 2020) proposed various optimal non-linear spectral shrinkage estimators exploiting the random matrix theory techniques introduced in (Ledoit and Péché, 2011). These estimators are designed to be asymptotically optimal (as $N, T \to \infty,\ N/T \to c$) for minimizing the Frobenius distance from the unknown covariance matrix or, equivalently, for minimizing the portfolio return variance (Ledoit and Wolf, 2017).[5] They are derived from the optimal shrinkage implied by random matrix theory, which is given by a function of the limit distribution of sample eigenvalues that are estimated consistently using nonparametric methods.

We crucially depart from the above approaches by introducing a universal approximator for a broad class of shrinkage functions that are analytically tractable and easily adaptable to optimize a quadratic portfolio objective, such as a (Markowitz, 1952)-type criterion[6]. Instead of trying to recover the (potentially complex) form of the asymptotically optimal shrinkage and then trying to estimate it consistently, we directly optimize the out-of-sample portfolio utility using our universal approximation[7]. This makes our approach explicitly designed for portfolio optimization and asset pricing. Importantly, while our optimal portfolio estimator depends on a corresponding spectral shrinkage estimator of the returns' covariance matrix, the associated optimal shrinkage also incorporates information from average returns whenever the optimized out-of-sample portfolio utility does. In this way, our approach optimally adjusts to effectively shrink in-sample estimated asset Sharpe ratios to maximize out-of-sample portfolio performance. To our knowledge, this portfolio shrinkage approach is novel and has never been studied in the literature before.

SDF estimation with PCs: Motivated by the emergence of the factor zoo (see (Cochrane, 2011), and (Harvey et al., 2016)), many papers attempted to find a characteristics-sparse

[5]The Frobenius distance measures the Euclidean distance between two matrices A, B: $\|A - B\|_{Fr}^2 = \sum_{i,j}(A_{ij} - B_{ij})^2$.

[6]In the context of linear shrinkage for portfolio optimization, (Pedersen et al., 2020) emphasize the significance of selecting the correct mean-variance objective.

[7]As (Didisheim et al., 2023) show, when properly defined, minimizing the (Hansen and Richard, 1987) OOS distance is equivalent to maximizing the Sharpe ratio of the SDF.

representation of the SDF.[8] Recent research, based on the ideas of APT, proposed instead to look for a PC-sparse representation of the SDF constructed from a few (typically, less than six) principal components of factors. See, for example, (Kozak et al., 2018), (Kozak et al., 2020), (Lettau and Pelger, 2020), (Kelly et al., 2020), (Gu et al., 2021), (Bryzgalova et al., 2023b), (Giglio and Xiu, 2021).

Of particular relevance to us is the paper (Kozak et al., 2020), which argues that a good SDF approximation can be constructed by selecting the top few PCs of factors and applying simple Ridge shrinkage to their covariance matrix. Thus, the SDF is *sparse in the space of PCs*. In this paper, using a different dataset (we use 153 factors from (Jensen et al., 2023)), we find that low-variance estimated PCs are important contributors to SDF performance. The out-of-sample out-performance increases monotonically in the number of PCs. Furthermore, our non-linear shrinkage methodology dominates the simple Ridge. While the latter corresponds to a prior with a fixed degree of uncertainty, our optimal shrinkage may be interpreted as capturing heterogeneous beliefs of investors with varying degrees of prior uncertainty.

Several recent papers argue that the emergence of the factor zoo is associated with the existence of the so-called weak factors, whose risk premia are too small to be efficiently identifiable [9]. To address the weak factor problem, (Lettau and Pelger, 2020) develop a novel covariance shrinkage methodology called Risk Premium PCA (RP-PCA). This methodology still advocates a PC-sparse SDF but with PCs computed for the shrunk covariance matrix. This shrinkage introduces an important bias in the PCs, tilting them towards the vector of their sample means. In particular, their estimator does not belong to the spectral family. The RP-PCA aims to correct the bias (induced by complexity $c = N/T > 0$) in estimating PCs. (Lettau and Pelger, 2020) prove that this bias correction is indeed efficient for high-variance PCs but cannot be used to fix low-variance PCs because they are severely contaminated by noise; hence, they build their SDF from a few bias-corrected top PCs. By contrast, our approach keeps all of the original PCs (including the low-variance ones) and, instead, re-weights them optimally through eigenvalue shrinkage. These low-variance PCs matter

[8] See, e.g., (Fama and French, 1993), (Hou et al., 2015), (Fama and French, 2015), and (Barillas and Shanken, 2018).

[9] See, e.g., (Bryzgalova et al., 2023a), (Preite et al., 2022)

precisely because they capture the exposures to the *weak factors*. Shrinkage corrects some of the bias in estimating these weak factor risk premia.

Statistical Pricing Frictions and Complexity: Our paper also addresses statistical limits to efficient estimation in Finance. It is particularly suited for dealing with situations where the number of model parameters and training samples are of the same order of magnitude. A sequence of recent papers shows how classical statistical theory needs to be adjusted when dealing with such situations of estimation complexity. See, (Martin and Nagel, 2021), (Kelly et al., 2022), (Da et al., 2022), and (Didisheim et al., 2023). In particular, (Martin and Nagel, 2021) emphasizes the importance of employing both shrinkage techniques and out-of-sample (OOS) testing in Bayesian high-dimensional models. Similarly, (Kelly et al., 2022) and (Didisheim et al., 2023) highlight theoretically and empirically the advantages of complex models in asset pricing for achieving superior out-of-sample performance. This holds both for forecasting asset returns (Kelly et al., 2022) and constructing SDFs (Didisheim et al., 2023). Our paper contributes to the literature on complexity by introducing a robust shrinkage methodology to mitigate high-dimensional noise. Furthermore, we offer an unbiased estimator for OOS performance, aiding in model selection and helping to bridge the complexity wedge (Didisheim et al., 2023) between in-sample (IS) and OOS performance.

3 Optimal Portfolio Shrinkage

We consider a set N of assets (factors) whose excess returns follow a stochastic process $F_t \in \mathbb{R}^N,\ t \geq 0$. In a perfect information environment, an economic agent maximizing quadratic utility[10]

$$U(R_t^\pi) \ = \ R_t^\pi \ - \ \frac{1}{2}(R_t^\pi)^2 \tag{1}$$

of portfolio returns

$$R_t^\pi \ = \ \pi' F_t \tag{2}$$

[10]Our analysis is readily applicable to non-quadratic utilities. However, in these cases, the starting point for shrinkage deviates from the traditional Markowitz solution. Instead, the function f should be applied to the IS solution derived from the non-quadratic utility.

would select the efficient portfolio

$$\pi_* \;=\; E[FF']^{-1}\,E[F]\,, \tag{3}$$

achieving the expected utility

$$E[U(R_t^{\pi})] \;=\; \frac{1}{2}E[F]'E[FF']^{-1}E[F]\,. \tag{4}$$

A real-world economic agent with access to T in-sample observations of F_t can instead compute finite-sample moments

$$\begin{aligned}\bar{E}[FF'] \;&=\; \frac{1}{T}\sum_{t=1}^{T} F_t F_t' \\ \bar{E}[F] \;&=\; \frac{1}{T}\sum_{t=1}^{T} F_t,\end{aligned} \tag{5}$$

and construct a simple, empirical counterpart of (3), given by

$$\bar{\pi} \;=\; \bar{E}[FF']^{-1}\bar{E}[F]\,. \tag{6}$$

The corresponding in-sample (IS) utility is given by

$$\bar{u} \;=\; \frac{1}{T}\sum_{t=1}^{T} U(R_t^{\bar{\pi}}) \;=\; \frac{1}{2}\bar{E}[F]'\bar{E}[FF']^{-1}\bar{E}[F]\,, \tag{7}$$

while the out-of-sample (OOS) expected utility is given by

$$u^{OOS} \;=\; E[U(R_t^{\bar{\pi}})],\; t > T. \tag{8}$$

When $N/T \neq 0$, *complexity* leads to a breakdown of the law of large numbers, and empirical and theoretical moments diverge,

$$\bar{E}[F] \;\not\to\; E[F],\; \bar{E}[FF'] \;\not\to E[FF']\,. \tag{9}$$

The exact out-of-sample behavior of (6) depends on subtle properties of the stochastic process F_t. When F_t are independent and identically distributed over time, random matrix theory methods can be used to characterize these quantities in the limit when $N, T \to \infty,\ N/T \to c$. See, (Didisheim et al., 2023). The key insight from these theoretical results is that, for $c > 0$, there is a potentially large *complexity wedge*

$$Wedge \ = \ \bar{u} \ - \ u^{OOS} \ > \ 0. \tag{10}$$

(Didisheim et al., 2023) refer to this wedge as *limits to learning* and show how this wedge originates in the misestimation of factor moments (9). The common approach in the literature for dealing with this misestimation is the shrinkage of the covariance matrix.

Let $\bar{E}[FF'] = U\,\mathrm{diag}(\lambda)U'$ be the eigenvalue decomposition of the empirical covariance matrix, and $R^{PC}_{i,t} = U_i' F_t$ be the returns of these principal components. We also use $\bar{R}^{PC}_i = \bar{E}[R^{PC}_{i,t}]$ to denote the in-sample mean returns of these PCs. In this case, we can rewrite portfolio returns as

$$R^{\bar{\pi}}_t \ = \ \sum_{i=1}^{N} \frac{\bar{R}^{PC}_i}{\lambda_i} R^{PC}_{i,t}\,. \tag{11}$$

In other words, the estimated efficient portfolio return is the sum of PC returns, with each PC weighted by its estimated risk-return tradeoff. Empirically, when N is large enough, we often observe that these tradeoffs, $\frac{\bar{R}^{PC}_i}{\lambda_i}$, are very large for small λ_i. As a result, the estimated efficient portfolio severely overweights low-variance PCs, leading to poor OOS performance. A common approach for dealing with instabilities induced by small eigenvalues is to use the Ridge-penalized covariance matrix (see, e.g., (Kozak et al., 2020), (Didisheim et al., 2023)),

$$\bar{\pi}(z) \ = \ (zI + \bar{E}[FF'])^{-1}\bar{E}[F], \tag{12}$$

leading to the following decomposition of portfolio returns:

$$R^{\bar{\pi}(z)}_t \ = \ \sum_{i=1}^{N} \frac{\bar{R}^{PC}_i}{\lambda_i + z} R^{PC}_{i,t} \ = \ \sum_{i=1}^{N} \frac{\bar{R}^{PC}_i}{\lambda_i} \frac{1}{1 + z/\lambda_i} R^{PC}_{i,t}\,. \tag{13}$$

The formula (13) shows how a Ridge penalty acts as a "soft" thresholding of the eigenvalues, with the shrinkage factor $\frac{1}{1+z/\lambda_i}$ effectively annihilating the contributions of low-variance PCs. Two limiting cases of Ridge shrinkage are $\lim z \to 0$ (the so-called Ridgeless case)

$$\lim_{z\to 0} \bar{\pi}(z) \;=\; (\bar{E}[FF'])^{+}\bar{E}[F] \tag{14}$$

where $(\bar{E}[FF'])^{+}$ is the pseudo-inverse of the covariance matrix. When $N < T$, the Ridgeless portfolio is equal to the Markowitz portfolio (11). The infinite Ridge limit $z \to \infty$, converges to the simple "momentum" portfolio (for example, see (Arnott et al., 2023) and (Gupta and Kelly, 2019)) that completely ignores the covariance matrix and invests proportionally to in-sample mean returns:

$$\lim_{z\to\infty} z\bar{\pi}(z) \;=\; \lim_{z\to\infty} (I + z^{-1}\bar{E}[FF'])^{-1}\bar{E}[F] \;=\; \bar{E}[F]\,. \tag{15}$$

By varying z between 0 and ∞, we span a spectrum of potential SDFs. Initially, with no shrinkage, the covariance matrix and all its associated noise have a direct impact on the optimal portfolio. As we progress towards full shrinkage, this influence wanes. Ultimately, in the full shrinkage regime, the covariance matrix is completely disregarded, with only the means playing a significant role.

Motivated by the Arbitrage Pricing Theory (APT) of (Ross, 1976), some papers (see, e.g., (Kozak et al., 2020)) use hard thresholding of eigenvalues, only retaining top principal components in (13) (see (Chamberlain and Rothschild, 1982) for the underlying theory). This "one-size-fits-all" shrinkage approach is potentially highly inefficient. Ideally, we would like to have an estimator that optimally defines the contribution of each PC based on its estimated OOS risk-return tradeoff. The goal of this paper is to develop such an algorithm.

Formally, an estimator that only shrinks the weights of all PCs in (13) without modifying the PCs themselves is commonly referred to as a spectral shrinkage estimator. This class of estimators was introduced in the influential paper by (Stein, 1986). A generic spectral shrinkage estimator is defined by a function f applied to the eigenvalues of the sample

covariance matrix, whereby $\bar{E}[FF']$ is replaced with

$$f(\bar{E}[FF']) \;=\; U\,\mathrm{diag}(f(\lambda))U'\,. \tag{16}$$

Let

$$\bar{\pi}(f) \;=\; \underbrace{f(\bar{E}[FF'])}_{shrunk\ \ inverse\ \ covariance\ \ matrix}\bar{E}[F] \tag{17}$$

be the *f-spectral shrinkage estimator of the (feasible) efficient portfolio* $E[FF']^{-1}E[F]$. Then, we can rewrite portfolio returns $R_f(f) \;=\; R_t^{\bar{\pi}(f)}$ as

$$R_t(f) \;=\; \bar{\pi}(f)'F_t \;=\; \sum_{i=1}^{N} \underbrace{f(\lambda_i)\,\bar{R}_i^{PC}}_{shrunk\ \ PC\ \ weights} R_{i,t}^{PC}\,. \tag{18}$$

We can now formally define the optimal spectral shrinkage estimator.

Definition 1 (Optimal Non-linear Shrinkage) *Let* $F_{IS} = \{F_1, \cdots, F_T\}$ *be the in-sample factor returns. The optimal spectral shrinkage estimator is a function* $f:\ \mathbb{R}\times\mathbb{R}^{N\times T}\to\mathbb{R}$, *such that* $f(\lambda; F_{IS})$ *solves the* ***OOS*** *utility maximization problem*

$$\max_f\ E\left[U\left(R_t(f)\right)\right]\ ,\ t>T\,. \tag{19}$$

The key aspect of the optimal spectral shrinkage is the dependency of the non-linear function f on the in-sample observations F_{IS}. Thus, while the actual function $f(\cdot; F_{IS})$ applied to the eigenvalues of $\bar{E}[FF']$ has only one argument $\lambda\in\mathbb{R}$, the shrinkage operator, f, has, in fact, $NT+1$ arguments.[11] With stationary data and low complexity, when $N/T\to 0$, Lemma 7 implies that shrinkage is sub-optimal and, hence, the optimal $f(\lambda; F_{IS}) = \lambda$ is independent of the in-sample data. However, some shrinkage is always optimal[12] when $c = N/T > 0$.

[11] (Stein, 1986) introduced the infeasible optimal spectral shrinkage for the Frobenius norm objective. This infeasible estimator depends on both the in-sample data and the true, unobservable covariance matrix of F. In the online Appendix, we derive an analog of Stein's infeasible estimator for the portfolio optimization problem (19).

[12] See Corollary 4.

To approach the problem of finding the optimal shrinkage, we first need to compute the expected OOS utility in (19), which seems impossible because neither $E[F]$ nor $E[FF']$ are observable. To overcome this issue, we follow an indirect approach and compute an approximation based on a classical technique known as Leave-One-Out (LOO). LOO is based on a simple observation that when F_t are independent and identically distributed, one can compute an unbiased estimate of the OOS performance of a portfolio by dropping any observation t and then evaluating OOS performance on that removed observation. For any t, define the LOO estimators of the empirical moments as follows:

$$\begin{aligned} \bar{E}_{T,t}[FF'] &= \frac{1}{T} \sum_{\tau \neq t, 1 \leq \tau \leq T} F_\tau F_\tau' \\ \bar{E}_{T,t}[F] &= \frac{1}{T} \sum_{\tau \neq t, 1 \leq \tau \leq T} F_\tau . \end{aligned} \tag{20}$$

Given these LOO estimators of the empirical moments, we can define the analog of the spectral shrinkage estimator (17):

$$\bar{\pi}_{T,t}(f) \;=\; f(\bar{E}_{T,t}[FF'])\bar{E}_{T,t}[F]\,. \tag{21}$$

The dropping of the observation F_t allows us to evaluate the OOS performance, $\bar{\pi}_{T,t}(f)'F_t$, while staying within the in-sample data $F_\tau, \tau \in [1,T]$. Our ultimate objective is to measure the expected OOS performance of $\bar{\pi}(f)$. To achieve this goal, we can build unbiased estimators of OOS moments of $R_t(f) \;=\; \bar{\pi}(f)'F_t$ by averaging the realized performance of $\bar{\pi}_{T,t}(f)'F_t$ across t, as is shown in the following lemma.

Lemma 1 *Suppose that F_t are interchangeable/exchangeable sequence. Let*

$$R_{T,t}(f) \;=\; \bar{\pi}_{T,t}(f)'F_t\,,\;\; t = 1, \cdots, T\,. \tag{22}$$

Then,

$$U^{OOS}_{LOO}(f) \;=\; \frac{1}{T}\sum_{\tau=1}^{T} U(R_{T,\tau}(f)) \tag{23}$$

is an unbiased estimator of OOS expected utility:

$$E[U(R_t(f))] \;=\; E\left[U_{LOO}^{OOS}(f)\right]\,, \tag{24}$$

where, $t > T$.

The formula (24) motivates the following feasible version of the infeasible problem (19).[13]

Definition 2 (Optimal Non-linear Feasible Shrinkage) *The optimal* **feasible** *spectral shrinkage estimator is a function* f *solving the utility maximization problem*

$$\max_f U_{LOO}^{OOS}(f)\,. \tag{25}$$

The formula (25) defines a feasible, directly observable objective for optimal shrinkage. However, at first sight, the maximization problem (25) still looks complex. Indeed, even computing the objective requires evaluating the function f on the eigenvalues of T different matrices $\bar{E}_{T,t}[FF']$, $t = 1, \cdots, T$. The main theoretical result of our paper is an explicit, tractable, analytical solution to (25) that we derive in the next section.

4 Universal Portfolio Shrinkage Approximator

We now briefly switch our focus to the Ridge shrunk efficient portfolio. Our first key insight is that one can rewrite the feasible OOS expected utility estimator (24) for $f_z(\lambda) = \frac{1}{z+\lambda}$ in terms of the inverse of just one matrix, $\bar{E}[FF']$. Let

$$\bar{\pi}_{T,\tau}(f_z) \;=\; (\bar{E}_{T,\tau}[FF'] + zI)^{-1}\bar{E}_{T,\tau}[F] \tag{26}$$

[13] An important theoretical question is whether the estimator $U_{LOO}^{OOS}(f)$ is consistent: Is it true that, as $T \to \infty$, $U_{LOO}^{OOS}(f) \to E[U(R_T(f))]$ in probability. Such a result would imply that maximizing $U_{LOO}^{OOS}(f)$ directly is equivalent to maximizing the true out-of-sample expected performance. The literature has established such a result for LOO estimators for the linear regression problem. See, e.g., (Hastie et al., 2019) and (Patil et al., 2021). Establishing it for the utility maximization problem in our setting is technically more involved but can be achieved using the results from (Didisheim et al., 2023). We leave this important question for future research.

be the (21) estimator for $f_z(\lambda) = \frac{1}{z+\lambda}$ and $R_{T,\tau}(f_z)$ the corresponding portfolio return (22). The following lemma provides an explicit link between LOO and simple in-sample returns.

Lemma 2 (LOO Ridge Performance) *We have*

$$R_{T,\tau}(f_z) = \underbrace{\frac{1}{1-\psi_\tau(z)}}_{complexity\ \ multiplier} \Big(R_\tau(f_z) \ - \ \underbrace{\psi_\tau(z)}_{overfit}\Big), \tag{27}$$

where, $R_\tau(f_z)$ *is the in-sample return at time* τ*,*

$$R_\tau(f_z) \ = \ \bar{\pi}(f_z)' F_\tau, \, \tau \leq T, \tag{28}$$

and

$$\psi_\tau(z) \ = \ \frac{1}{T} F_\tau' (zI + \bar{E}[FF'])^{-1} F_\tau \, . \tag{29}$$

The quantity ψ_τ plays a key role in our analysis. It is responsible for *complexity corrections,* originating in the high dimensionality of F_t, when number of assets is comparable to number of observations. Complexity corrections manifest themselves through the two terms in (27). The *overfit* term accounts for the fact that the in-sample mean of the efficient portfolio return overestimates the true mean. The *complexity multiplier* accounts for the fact that the in-sample covariance matrix underestimates the true amount of risk in the portfolio. By a direct calculation based on the Sherman-Morrison formula, we have that $\psi_\tau(z) \in (0,1)$ and, hence, the multiplier $\frac{1}{1-\psi_\tau(z)}$ is always above one, showing precisely by how much true out-of-sample variance is higher than the in-sample variance.[14] One can derive the following bound for ψ_τ.

Lemma 3 *Let* $c = N/T$ *be the model complexity. Assuming all factor returns,* F_t*, are*

[14] We have

$$\psi_\tau \ = \ \frac{T^{-1} F_\tau' (zI + \bar{E}_{T,\tau}[FF'])^{-1} F_\tau}{1 + T^{-1} F_\tau' (zI + \bar{E}_{T,\tau}[FF'])^{-1} F_\tau} \tag{30}$$

bounded by a constant K in absolute value. Then,

$$\psi_\tau(z) \;<\; \min\{1, z^{-1}\, c\, K^2\}\,. \tag{31}$$

In particular, ψ_τ vanishes when c is small.

The key insight from Lemmas 2 and 3 is that, when complexity is large, estimation errors accumulate across N factors, leading to a breakdown of the law of large numbers: Even when T is large, errors may stay significant, proportional to $c = N/T$. The difference between the fully in-sample return $R_\tau(f_z)$ and the out-of-sample return $R_{T,\tau}(f_z)$ comes from two effects. First, $R_\tau(f_z)$ has a higher mean in-sample return than the OOS $R_{T,\tau}(f_z)$ because of the overfit term in (27). Second, it has a lower in-sample volatility due to the complexity multiplier in (27). Both effects imply that, without accounting for complexity corrections, the in-sample-based estimates might give a biased, overly optimistic view of the performance of efficient portfolios and their risk-return tradeoffs. The bias and the underestimation of risk can be severe when the complexity $c = N/T$ is large.

Our next key observation in this paper is that the simple algebraic structure of (27) allows us to compute all expressions analytically, only involving the full sample covariance matrix $\bar{E}[FF']$. It then seems natural to extend our analysis from the single Ridge function $f_z(\lambda)$ to functions representable as linear combinations of the simple Ridge. One important issue with combining Ridge portfolios with different values of z is that their scales (and, hence, volatilities) differ drastically because $f_z(\lambda) \sim 1/z$. In order to deal with this scale heterogeneity, we introduce normalization constants

$$c_z = \frac{\frac{1}{N}\operatorname{tr}(\bar{E}[FF']) + z}{\frac{1}{N}\operatorname{tr}(\bar{E}[FF'])}. \tag{32}$$

The constant $\frac{1}{N}\operatorname{tr}(\bar{E}[FF']) = \frac{1}{N}\sum_{i=1}^N \|F_i\|^2$ is the average in-sample second moment of factors. Using the crude approximation $\bar{E}[FF'] \approx \frac{1}{N}\operatorname{tr}(\bar{E}[FF'])I$, we get

$$c_z f_z(\bar{E}[FF']) \;=\; c_z(zI+\bar{E}[FF'])^{-1} \;\approx\; \frac{\frac{1}{N}\operatorname{tr}(\bar{E}[FF']) + z_i}{\frac{1}{N}\operatorname{tr}(\bar{E}[FF'])}(zI+\frac{1}{N}\operatorname{tr}(\bar{E}[FF'])I)^{-1} \;=\; \frac{1}{\frac{1}{N}\operatorname{tr}(\bar{E}[FF'])}\,, \tag{33}$$

implying that, under this approximation, the scale of $c_z f_z(\bar{E}[FF'])$ is independent of z[15]. We can now introduce the relevant ridge ensembles.

Definition 3 *Let $Z = (z_i)_{i=1}^L$ be a grid of Ridge penalties, and $W = (w_i)_{i=1}^L$ a collection of weights.*

$$f_{Z,W}(\lambda) \;=\; \sum_{i=1}^{L} \underbrace{(z_i + \lambda)^{-1}}_{Ridge}\; \underbrace{w_i}_{weight}\;. \tag{34}$$

We refer to $\mathcal{F}(Z) = \{f_{Z,W}(\lambda) : W \in \mathbb{R}^L\}$ as the Ridge ensemble, and to $\mathcal{F}_C(Z) = \{f_{Z,W}(\lambda) : W \in \mathcal{S}_+^L\}$ as the constrained Ridge ensemble, where $\mathcal{S}_+^L = \{W \in \mathbb{R}_+^L, \sum_{i=1}^N w_i/c_{z_i} = 1\}$.

The Ridge ensemble is a rich, parametric family of functions. Since the functions f from this ensemble are linear in W, the OOS utility estimator (25) is quadratic in these weights, as is shown by the following result.

Lemma 4 *Let*

$$\begin{aligned} \bar{\mu}(Z) \;&=\; \left(\frac{1}{T}\sum_{t=1}^{T} R_{T,t}(f_{z_i})\right)_{i=1}^{L} \in \mathbb{R}^L \\ \bar{\Sigma}(Z) \;&=\; \left(\frac{1}{T}\sum_{t=1}^{T} R_{T,t}(f_{z_i}) R_{T,t}(f_{z_j})\right)_{i,j=1}^{L} \in \mathbb{R}^{L\times L} \end{aligned} \tag{35}$$

to be the LOO-based estimators of the OOS means and covariances of the Ridge components of the Ridge ensemble. Then, we have

$$R_{T,\tau}(f_{Z,W}) \;=\; \sum_{i=1}^{L} w_i R_{T,\tau}(f_z). \tag{36}$$

Therefore, the feasible estimator (25) of the OOS utility is given by

$$U_{LOO}^{OOS}(f_{Z,W}) \;=\; W'\bar{\mu}(Z) \;-\; 0.5 W'\,\bar{\Sigma}(Z)\,W\,. \tag{37}$$

[15]The desired effect could be achieved with any scaling factor c_z that is proportional with z. Specifically, choosing $c_z = z$ aligns precisely with the actions of a Bayesian agent. For further details, see Lemma 6.

Lemma 4 shows how the OOS utility can be computed explicitly in terms of the estimated OOS moments (35). As a result, Lemma 4 implies that finding the optimal spectral shrinkage inside the Ridge ensemble amounts to solving the OOS (based on LOO) Markowitz problem, with the original N-dimensional vector of asset returns F_t replaced with the L-dimensional vector of shrunk LOO returns, $(R_{T,t}(f_{z_i}))_{i=1}^{L},\ i = 1, \cdots, L$. This simple asset space transformation implies that the optimization problem (25) admits an explicit, interpretable, closed-form solution *if we restrict the class of functions f in (25) to $\mathcal{F}(Z)$*. We refer to this solution as (Constrained) Universal Portfolio Shrinkage Approximator, (C)UPSA. Formally, we define the UPSA and CUPSA estimators as solutions to the following (constrained) versions of (25):

$$\begin{aligned} f_{UPSA} &= \arg\max_{f\in\mathcal{F}(Z)} U_{LOO}^{OOS}(f) \\ f_{CUPSA} &= \arg\max_{f\in\mathcal{F}_C(Z)} U_{LOO}^{OOS}(f) \end{aligned} \tag{38}$$

The key implication of the above discussion is that the non-parametric *optimization over functions* in (38) is equivalent to a closed form, explicit *optimization over weight vectors W*.

Theorem 1 (UPSA and CUPSA) *We have*

$$\begin{aligned} f_{UPSA}(\lambda) &= f_{Z,W_{UPSA}}(\lambda), \\ f_{CUPSA}(\lambda) &= f_{Z,W_{CUPSA}}(\lambda) \end{aligned} \tag{39}$$

with

$$\begin{aligned} W_{UPSA} &= \bar{\Sigma}(Z)^{-1}\bar{\mu}(Z) \\ W_{CUPSA} &= \arg\max_{W\in\mathcal{S}_+^L}(W'\bar{\mu}(Z) - 0.5W'\bar{\Sigma}(Z)W). \end{aligned} \tag{40}$$

The name "Universal Approximation" naturally leads us to the question: How rich is the Ridge ensemble? What kind of non-linear functions can be approximated with functions from $\mathcal{F}(Z)$ and $\mathcal{F}_C(Z)$? It turns out that these ensembles have a *universal approximation property*, as is shown by the following lemma.

Lemma 5 *Any continuous function $f(x)$ on a compact interval can be uniformly approximated by a function $f \in \mathcal{F}(Z)$ if the grid Z is sufficiently large and dense.*

Furthermore, for any matrix monotone-decreasing function[16] *with $\lambda f(\lambda)$ being uniformly bounded for $\lambda > 0$, there exists a unique constant, $\kappa_*(f) > 0$, such that $\kappa_*(f)f(\lambda)$ can be uniformly approximated on compact subsets of $(0, +\infty)$ by a function $\tilde{f} \in \mathcal{F}_C(Z)$ if the grid Z is sufficiently large and dense.*

Lemma 5 justifies the term "Universal Approximation." Since any non-linear shrinkage f can be approximated by a Ridge ensemble, the economic agent maximizing any utility function can achieve approximately optimal performance by using a combination of shrinkages from the Ridge ensemble.

Corollary 2 (The Universal Approximation Property) *Let*

$$\begin{aligned} f^*(\lambda) &= \arg\max\{U_{LOO}^{OOS}(f) : \ f \text{ is continuous}\} \\ f_C^*(\lambda) &= \arg\max\{U_{LOO}^{OOS}(f) : \ f \text{ is matrix monotone}, \sup_{\lambda>0} |\lambda f(\lambda)| < \infty, \text{and } \kappa_*(f) = 1\} \end{aligned} \tag{41}$$

Then, for any $\varepsilon > 0$, we can make the grid Z sufficiently large and dense, so that

$$\begin{aligned} U_{LOO}^{OOS}(f_{UPSA}) &\geq U_{LOO}^{OOS}(f^*) - \varepsilon \\ U_{LOO}^{OOS}(f_{CUPSA}) &\geq U_{LOO}^{OOS}(f_C^*) - \varepsilon \end{aligned} \tag{42}$$

Simply put, the Universal Approximation Property implies that the simple Ridge shrinkage functions can serve as a basis for approximating arbitrary, non-linear shrinkage estimators. The result for CUPSA is particularly important. Indeed, requiring that the shrinkage function belong to the class of matrix monotone functions imposes natural, economic risk-taking constraints on the shrinkage estimator. Effectively, it requires that any increase in the realized risk $\bar{E}[FF']$ should be associated with lower risk-taking, represented by $f(\bar{E}[FF'])$. The normalization $\sum_i w_i/c_{z_i} = 1$ (which is equivalent to $\kappa_*(f) = 1$) ensures that CUPSA

[16] A function f is called matrix monotone decreasing if $f(A) - f(B)$ is positive semi-definite whenever $B - A$ is positive semi-definite. It is known that any function $f \in \mathcal{F}_C(Z)$ is matrix monotone decreasing. The fact that the converse is true is highly non-trivial and follows from the celebrated (Löwner, 1934) theorem.

shrinks eigenvalues by building a convex combination of the rescaled simple Ridge shrinkages $c_{z_i} f_{z_i}$. Since f_{CUPSA} is always positive and monotone increasing in λ, it preserves positivity and the order of the empirical eigenvalues. Economically, this means that estimated PCs with high in-sample risk are assigned a higher, positive denominator in their "shrunk" risk-return tradeoffs in the decomposition (18).

The closed-form solution of Theorem 1 provides a tractable characterization of the solution in terms of the *variability in performance among the individual components of the Ridge ensemble.* Namely, non-linear shrinkage is only optimal when the Ridge portfolio returns $R_t(f_{z_i}),\ i = 1, \cdots, L$ exhibit a sufficient amount of variability in risk-return tradeoffs across z_i. It is this variability that produced potential diversification gains, implying that combining multiple Ridge penalties is beneficial. In other words, provided that the momentum portfolio (15) and the Markowitz portfolio (14) are sufficiently different (diversification benefits across assets) then there should also be a benefit to Ridge shrinkage diversification and NLS.

4.1 Implications of Complexity

The gain from using the Ridge ensemble is determined by the diversification benefits from using the optimal Ridge weights in (40). The formula (27) implies that to understand these benefits, we need to study the behavior of $\psi_t(z)$. Although $\psi_t(z)$ typically varies with t, it is possible to apply asymptotic principles from Random Matrix Theory to eliminate this time dependence. We will need the following result from (Didisheim et al., 2023) to achieve this.

Proposition 3 ((Didisheim et al., 2023)) *Suppose that $F_t = \lambda + \Psi^{1/2} X_t$, where X_t are i.i.d. mean zero, unit-variance variables with uniformly bounded forth moments, and the eigenvalue distribution of $\Psi \in \mathbb{R}^{N \times N}$ converges as $N \to \infty$. Then, the limits*

$$\begin{aligned} m(-z; c) &= \lim_{N,T\to\infty,\ N/T\to c} N^{-1} \operatorname{tr}((\bar{E}[FF'] + zI)^{-1}) \\ \psi(z; c) &= \lim_{N,T\to\infty,\ N/T\to c} \psi_t(z) \end{aligned} \tag{43}$$

exist in probability and are independent of t and of the expected risk premia vector λ.

Furthermore, the asymptotic overfit is given by

$$\psi(z;c) \;=\; c(1 - zm(-z;c))\,. \tag{44}$$

4.1.1 Optimality of Non-Linear Shrinkage

Proposition 3 allows us to drastically simplify the calculations of (35) and highlight explicitly how complexity impacts the optimal non-linear shrinkage. The first consequence is that we can now show the following Corollary:

Corollary 4 (Non-Zero Shrinkage is Always Optimal) *Under the hypothesis of Proposition 3,*

$$\sup_{z>0} U^{OOS}_{LOO}(f_z) \;>\; U^{OOS}_{LOO}(f_0)\,, \tag{45}$$

and the supremum is always achieved for some $z_* > 0$. The gains from ridge shrinkage are bounded from above by $0.5\,\frac{(-2\psi(0;c)+\psi(0,c)^2)}{(1-\psi(0,c))^2}$.

Corollary 4 establishes optimality of Ridge shrinkage, contingent upon a non-zero overfit $\psi(z;c)$. By Lemma 3, the magnitude of the overfit is controlled by the complexity $c = \frac{N}{T} > 0$. In scenarios where $\psi(z;c)$ is close to zero, the benefits of shrinkage evaporate.

4.1.2 Two-Fund Separation

Corollary 4 shows that some form of shrinkage is always optimal. This raises the question: Do we really need the whole Ridge ensemble to construct UPSA, or is a single, optimally chosen z_* sufficient? The following Theorem provides an answer to this question, deriving the optimal UPSA weights.

Theorem 5 (Two-Fund Separation) *Under the hypothesis of Proposition 3, suppose that $z_0 = 0+$ (so that Ridgeless, (14), is the first element of the Ridge ensemble). Let $\psi(Z) = (\psi(z_i;c))_{i=1}^L$ be the vector of overfits for different values of z, and $\Sigma_{IS}(Z) = (\bar{E}[R(f_{z_i})R(f_{z_j})])_{i,j=1}^L$ the in-sample Ridge covariance matrix. Let also $\delta_{z_0} \;=\; (1, 0, \cdots, 0) \in \mathbb{R}^L$, and let $D(Z) \;=\; \text{diag}(\frac{1}{1-\psi(Z)}) \in \mathbb{R}^{L\times L}$, be the complexity multiplier. Then, for some explicit constants $\alpha,\ \beta >$*

0 :

$$W_{UPSA} \;=\; \alpha\,\delta_{z_0} \;+\; \beta\,D(Z)^{-1}\Sigma_{IS}(Z)^{-1}\psi(Z), \tag{46}$$

so that the UPSA efficient portfolio return is given by

$$\bar{\pi}(f_{UPSA}) \;=\; \alpha \underbrace{\bar{\pi}(f_0)}_{Markowitz} + \beta \underbrace{\bar{\pi}^{\psi}}_{complexity\ \ correction}, \tag{47}$$

where

$$\bar{\pi}^{\psi} \;=\; \sum_{z\in Z} \underbrace{\bar{\pi}(f_z)}_{Ridge\ \ portfolio\ \ (12)} (D(Z)^{-1}\Sigma_{IS}(Z)^{-1}\psi(Z))(z)\,. \tag{48}$$

Theorem 5 implies a surprising result: Even in a fully stationary, i.i.d. environment with constant risk premia, complexity leads to a systematic deviation from the conventional efficient portfolio theory, with a closed-form correction defined by the vector of overfits, $\psi(Z)$. By Proposition 3, this vector of overfits depends exclusively on the eigenvalue distribution of the true (unobservable) asset covariance matrix, $E[FF']$. The extent of this adjustment is critically linked to the magnitude of the overfit and the corresponding complexity corrections (see Lemma 2). Greater complexity necessitates a more substantial adjustment to the in-sample Markowitz portfolio. In the high complexity regime when $c = N/T$ is large, $\bar{\pi}^{\psi}$ in (47) dominates, tilting the optimal portfolio further away from the naive, in-sample estimator.

4.2 Economic Interpretations of CUPSA

CUPSA, the constrained version of UPSA, imposes discipline on the individual Ridge weights, minimizing instabilities due to potential degeneracies in the Ridge covariance matrix $\bar{\Sigma}(Z)$ in (35). As we now explain, the constraint of nonnegative weights summing up to one (see Definition 3) implies an important interpretation of CUPSA as a form of Bayesian posterior, aggregating a dispersed prior.

We follow (Kozak et al., 2020) and note that the Ridge-penalized optimal portfolio is, in fact, optimal for an economic agent who (irrationally) believes that the estimated covariance

matrix $\bar{E}[FF'] - \bar{E}[F]\bar{E}[F]'$ is correct (that is, the agent believes that $\bar{E}[FF'] - \bar{E}[F]\bar{E}[F]' = E[FF'] - E[F]E[F]'$), but is uncertain about the mean vector $E[F]$.[17] Here, we extend this observation to a special case of the $\mathcal{F}_C(Z)$ ensemble.

Lemma 6 *Consider an economic agent who (irrationally) believes that* $\Sigma = \bar{E}[FF'] - \bar{E}[F]\bar{E}[F]' = E[FF'] - E[F]E[F]'$. *She only cares about the mean, building a portfolio proportional to the posterior mean estimate,* $\pi^{mean} = E[F_{T+1}|F_{IS}]$. *The agent believes that* $F_t = \mu + \varepsilon_t$ *where* $\varepsilon_t \sim N(0, \Sigma)$ *is i.i.d., and the prior on* μ *is a Gaussian mixture:* μ *is sampled from* $N(0, z_i I)$ *distribution with probability* $(w_i/z_i)/\bar{w}$, *where* $\bar{w} = \sum_j (w_j/z_j)$. *Then,*

$$E[\mu|F_{IS}] \;=\; \frac{1}{\bar{w}} \sum_{i=1}^{L} w_i (z_i I + \Sigma)^{-1} \bar{E}[F]\,. \tag{49}$$

Note that the behavior of the Bayesian agent in the above lemma is in line with CUPSA (Theorem 1) when $c_{z_i} = z_i$. This is because

$$E[\mu|F_{IS}] \;=\; \sum_{i=1}^{L} \frac{\tilde{w}_i}{c_{z_i}} c_{z_i} (z_i I + \Sigma)^{-1} \bar{E}[F]\,, \tag{50}$$

where $\tilde{w}_i = w_i/\bar{w}$. By construction $\sum_{i=1}^{L} \frac{\tilde{w}_i}{c_{z_i}} = 1$, hence the constraint for the CUPSA ensemble, $\mathcal{F}_C(Z)$, is satisfied. In economic terms, the rational Bayesian agent's posterior should not scale with uncertainty (z_i). If uncertainty about the prior is large she should use the sample mean as the posterior estimate. This rationale guides our decision to design CUPSA as scale-independent.

The Gaussian mixture prior from Lemma 6 can be viewed as an extension of the simpler, single$-z$ prior in (Kozak et al., 2020). It is intuitive to expect that a typical market participant does not have a strong view of the exact degree of uncertainty about the mean vector μ, justifying the mixture prior. Alternatively, CUPSA can also be interpreted as an aggregation of beliefs of market participants with diverse degrees of prior uncertainty. One could imagine that an over-confident hedge fund manager who believes in outperforming the market would use a small z reflecting a tight prior, while a risk-averse retail investor

[17] In Appendix C, we offer an alternative, richer setting for bayesian updating.

might have a more dispersed prior. In equilibrium, (49) might represent the "true" expected returns aggregating these diverse priors and reflecting the strong heterogeneity of market participants.

5 Empirics

5.1 Data

We utilize the monthly frequency characteristic-managed portfolios (factors) from (Jensen et al., 2023)[18]. This comprehensive dataset contains monthly returns for $N = 153$ factors $F_t \in \mathbb{R}^{153}$, which are constructed from publicly traded stocks in the United States, covering the period from 1971-11-30 to 2022-12-31[19]. Each factor represents a capped value weighted long-short portfolio that is based on a distinctive characteristic[20], such as momentum, value, or reversal.

5.2 Methodology

We estimate portfolio weights using a rolling window of $T = 120$ months (10 years)[21] and rebalance every month. We then use the constructed portfolios and SDFs to price monthly factor returns. We fix the grid of Ridge penalties,[22] $z \in [10^i : i \in \{-10, -9, \cdots, -1\}]$ and construct Ridge-shrunk Markowitz portfolio weights with $Z = (z_i)_{i=1}^L$:

$$\begin{aligned} \bar{\pi}_t(f_Z) &= (\bar{\pi}_t(z_i))_{i=1}^L, \\ \bar{\pi}_t(f_{z_i}) &= c_{z_i}(t-T+1,t)(\bar{E}[FF'](t-T+1,t) + z_i I)^{-1}\bar{E}[F](t-T+1,t), \end{aligned} \tag{51}$$

[18] The data is accessible online at jkpfactors.

[19] Our findings are robust across different datasets, size stratifications, and construction methodologies. Many of our checks are available in the appendix.

[20] The exhaustive list of these 153 characteristics is detailed online in jkpfactors.

[21] Results are robust to the choice of a rolling window. Many robustness checks are available in the appendix.

[22] Our results are robust to the choice of the grid Z. Results for alternative grid choices are available upon request. However, optimal outcomes are achieved when the Z grid spans from the smallest to the largest eigenvalues of the covariance matrix

where $\bar{E}[F](t-T+1,t), \bar{E}[FF'](t-T+1,t)$ are sample means (5) estimated with the rolling window $[t-T+1,t]$. The Ridge scaling coefficients are computed by

$$c_{z_i} = \frac{\frac{1}{N}\operatorname{tr}(\bar{E}[FF'](t-T+1,t)) + z_i}{\frac{1}{N}\operatorname{tr}(\bar{E}[FF'](t-T+1,t))}, \tag{52}$$

where tr denotes the sum of the diagonal elements of a matrix, and N is the number of factors we use in the portfolio. Results are robust to Ridge scaling. Setting $c_{z_t} = 1$ does not alter results by much. However, we do this to make the shrinkage function **scale-invariant** and comparable to (Ledoit and Wolf, 2020)[23]. Given a vector of weights $W = (w_i)_{i=1}^L$, we construct the Ridge shrinkage approximator from Lemma 4:

$$\bar{\pi}_t(f_{Z,W}) = \sum_{i=1}^{L} w_i \bar{\pi}_t(f_{z_i}). \tag{53}$$

With the estimates $\bar{E}[F](t-T+1,t), \bar{E}[FF'](t-T+1,t)$, we compute leave-one-out (LOO) returns using formula (27). We use these LOO returns to compute $\bar{\mu}(Z)(t-T+1,t)$, $\bar{\Sigma}(Z)(t-T+1,t)$ from Lemma 4. Finally, we recover the optimal weight vectors $W_{\text{CUPSA}}(t-T+1,t)$ and $W_{\text{UPSA}}(t-T+1,t)$ using Theorem 1[24]. Everywhere in the sequel, we focus exclusively on CUPSA. While UPSA also achieves performance superior to that of classical shrinkage methodologies (such as Ridge), we find that CUPSA strictly dominates UPSA in almost every experiment we run.[25] The superiority of CUPSA is consistent with the intuitive economic interpretation of the positivity and normalization constraints that CUPSA imposes on the weights W. Indeed, as we explain above, these constraints are equivalent to imposing monotonicity in risk (more risk = larger estimated covariance matrix) and a normalization constraint. See the discussion after Corollary 2.

Given our estimated weight vector $W_{CUPSA}(t-T+1,t)$ for the window $[t-T+1,t]$, we calculate the out-of-sample return as

$$R_{t+1}(f_{Z,W_{CUPSA}(t-T+1,t)}) \;=\; \bar{\pi}_t(f_{Z,W_{CUPSA}(t-T+1,t)})' F_{t+1}. \tag{54}$$

[23] This is explained in greater detail in (33).
[24] The code for LOO CUPSA is available here.
[25] Results for UPSA are available upon request.

By construction, $\bar{\pi}_t(f_{Z,W_{CUPSA}(t-T+1,t)})$ only depends on factor returns during the $[t-T+1,t]$ time interval and, hence, portfolio returns (54) are indeed OOS. We compare the performance of (54) with that of three main spectral shrinkage benchmarks

- **LW** (Ledoit-Wolf): $R_{t+1}(f_{LW}(t-T+1,t))$, where $f_{LW}(t-T+1,t)$ is the optimal non-parametric non-linear shrinkage[26] from (Ledoit and Wolf, 2020), computed using the data in the $[t-T+1,t]$ time interval.[27]

- **Best z**: $R_{t+1}(f_{z_*(t-T+1,t)})$, the best LOO-based Ridge shrinkage utilizing the optimal penalty

$$z_*(t-T+1,t) = \arg\max_{z\in Z} U_{LOO}^{OOS}(f_z)\,. \tag{55}$$

- **KNS**: $R_{t+1}(f_{KNS}(t-T+1,t))$, is the return of the portfolio that applies Ridge shrinkage (55) to the **top 5 PCs**[28] of the covariance matrix. This is similar to (Kozak et al., 2020) who advocate for the use of a small number of PCs in the SDF.

$$R_{t+1}(f_{KNS}(t-T+1,t)) \;=\; R_{t+1}^{PC}(5, f_{z_*(t-T+1,t)}) \tag{56}$$

  where, R_{t+1}^{PC} is defined in (60).

To test the statistical significance of CUPSA out-performance, we run the following

[26] This algorithm minimizes the distance between the true and empirical covariance matrix. Distance is defined by using the Frobenius norm.

[27] We use the analytical formula of (Ledoit and Wolf, 2020) to do non-linear shrinkage conditional on Frobenius norm.

[28] An alternative could be using least absolute shrinkage and selection operator (LASSO; (Tibshirani, 1996)). These results are available on demand. Furthermore, (Quaini and Trojani, 2022) demonstrate that, under specific technical conditions, sequential shrinkage—initially employing Ridge followed by Lasso—is equivalent to the application of Elastic Net shrinkage. This development paves the way for the integration of Lasso with UPSA. We consider this extension as a topic for future research.

regression

$$
\begin{aligned}
&R_{t+1}(f_{Z,W_{CUPSA}(t-T+1,t)}) \\
&= \alpha + \beta_{LW} R_{t+1}(f_{LW}(t-T+1,t)) + \beta_{z_*} R_{t+1}(f_{z_*(t-T+1,t)}) + \beta_{KNS} R_{t+1}(f_{KNS}(t-T+1,t)) \\
&+ \beta_{MKT} MKT_{t+1} + \beta_{SMB} SMB_{t+1} + \beta_{HML} HML_{t+1} \\
&+ \beta_{CMA} CMA_{t+1} + \beta_{RMA} RMA_{t+1} + \beta_{MOM} MOM_{t+1} + \varepsilon_{t+1},
\end{aligned}
\tag{57}
$$

where $R_{t+1}(f_{z_*})$ is the "Best z" portfolio return; $R_{t+1}(f_{LW}(t-T+1,t))$ is the LW (Ledoit and Wolf, 2020) shrinkage portfolio return, and $R_{t+1}(f_{KNS})$ is the portfolio based on the KNS shrinkage methodology described above, (56). In addition, we use the returns of the five Fama-French factors, (Fama and French, 2015), and momentum, (Jegadeesh and Titman, 1993), as controls.[29] Their returns are denoted by $MKT_{t+1}, SMB_{t+1}, HML_{t+1}, CMA_{t+1}, RMA_{t+1}$, and MOM_{t+1}, respectively.

5.3 Non-Linearly Shrinking The Cross Section

The emergence of the factor zoo (Cochrane, 2011; Harvey et al., 2016), and the failure to find a characteristic-sparse representation of the SDF (Bryzgalova et al., 2023a) has led many researchers to look for other forms of sparsity. Based on the ideas of APT, several papers proposed to look for a PC-sparse representation of the SDF constructed from a few (typically, less than six) principal components of factors. In particular, (Kozak et al., 2020) argue that both PC-sparsity (annihilation of low-variance PCs) and shrinkage of the estimated eigenvalues of the remaining PCs are necessary to construct efficient SDFs. In this section, we provide evidence of a significant PC-based *virtue of complexity* of the CUPSA-SDF compared to alternative benchmarks: The OOS out-performance (measured as the t-statistic of α in the regression (57)) of the CUPSA-SDF is monotone increasing in the number of PCs and keeps increasing even when we add very low-variance PCs.

We compute PCs by decomposing factor returns covariance matrix $\bar{E}[FF'](t-T+1,t)$:[30]

[29] The data is from the website of Kenneth French.

[30] It is crucial that we perform the eigenvalue decomposition in-sample and use them to construct the SDF OOS: Using the infeasible OOS PCs would drastically boost performance due to the look-ahead bias. The

$$\bar{E}[FF'](t-T+1,t) = \bar{U}(t-T+1,t)\,\mathrm{diag}(\bar{\lambda}(t-T+1,t))\bar{U}(t-T+1,t)'\,, \tag{58}$$

where the eigenvalues $\bar{\lambda}$ are ordered to be decreasing: $\bar{\lambda}_1 \geq \cdots \geq \bar{\lambda}_N$. Denoting by $\bar{U}_i(t-T+1,t)$ the i-th column of $\bar{U}(t-T+1,t)$, we define the OOS returns on the $i-th$ IS principle component as

$$R^{PC}_{i,\tau}(t-T+1,t) \;=\; \bar{U}_i(t-T+1,t)'F_\tau\,. \tag{59}$$

Subsequently, we apply the CUPSA and optimal Ridge shrinkage methods to an incrementally expanding subset of PCs. Namely, for each $I = 1,\cdots,N$, we define $R^{PC}_\tau(I) = (R^{PC}_{i,\tau}(t-T+1,t))^I_{i=1}$, and then compute the in-sample Ridge portfolios based on the in-sample covariance matrix $R^{PC}_\tau(I)$ (computed using $\tau \in [t-T+1,t]$).[31] We then apply all our shrinkage methodologies (CUPSA, LW, KNS, and Best z) to these returns and study their out-of-sample behavior, defined as

$$R^{PC}_{t+1}(I, f(t-T+1,t)) \;=\; \pi^{PC}(I)(f)'R^{PC}_{t+1}(I) \tag{60}$$

where $f \in \{CUPSA,\ LW,\ KNS,\ Best\ z\}$. Importantly, all these portfolio returns are computed purely out-of-sample.

Figure 1 reports the resulting OOS Sharpe ratios. Our first observation is that all shrinkage methods saturate at around 30 PCs. This number is much higher than that for the SDF constructed in (Kozak et al., 2020), who argue that 5 PCs are sufficient to span the SDF.[32] Furthermore, the Sharpe ratios for all shrinkage methodologies deteriorate past the saturation point. By contrast, CUPSA stands out as the unique shrinkage estimator that is the least sensitive to low-variance PCs: Its performance is close to flat past the saturation point.

reason is that in a high-complexity regime, in-sample PCs are severely corrupted by noise. See, e.g., (Lettau and Pelger, 2020).

[31] Note that this matrix can also be computed directly through $\bar{U}(t-T+1,t)$ and $\mathrm{diag}(\bar{\lambda}(t-T+1,t))$.

[32] (Kozak et al., 2020) use a different (much smaller) set of factors; using a larger set of factors automatically leads to a higher complexity.

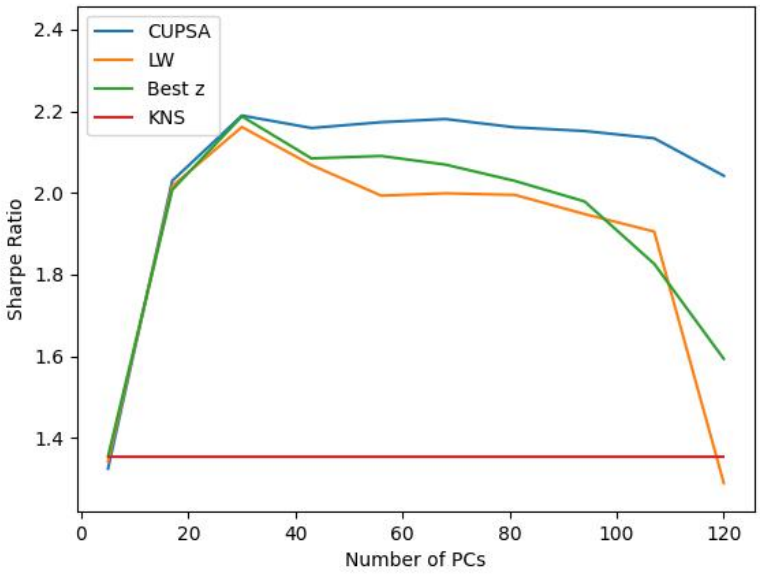

Figure 1: The figure shows the Sharpe Ratio of PCs for CUPSA (Theorem 1), LW (Ledoit and Wolf, 2020), KNS (56), and "Best z" (z_* of (55)), as the number of PCs grows. This is done by using (59). Results are from the period 1981-12-31 to 2022-12-31. Portfolios are estimated with a rolling window of $T = 120$ months and re-balanced monthly.

To gain a deeper insight into the nature of the non-linear shrinkage, we run the regression, (57), for the PC-based CUPSA portfolio returns, gradually increasing the number of PCs used, as set out in Equation (59). The results of this regression are reported in Figure 2. The t-statistics of alpha reveal a clear *virtue of complexity*.[33] The t-statistic of alpha keeps increasing even as we add the lowest-variance PCs.

Some of the observed outperformance in Figure 2 is driven by the divergence between the performance of alternative shrinkage estimators happening after the inclusion of the low-variance PCs. This divergence illustrates that a key power of CUPSA lies in its ability to efficiently weigh low-variance PCs based on their estimated OOS risk-return tradeoff. By contrast, other shrinkage estimators suffer (sometimes severely) from the associated estimation noise. These findings have important implications for our general understanding of factor structure and the search for PC-sparse SDFs motivated by the APT of (Ross, 1976): Even if some form of PC-sparsity is indeed optimal *ex-post*, finding the optimal degree of sparsity that works out-of-sample is difficult. Our extensive experiments suggest that the in-sample-optimal number of PCs varies over time, is unstable, and leads to inferior out-of-sample performance.

In Appendix B, we report the results of multiple robustness checks for various alternative factor datasets and rebalancing frequencies. In all these experiments, CUPSA significantly dominates alternative shrinkage methodologies.

5.4 Asset Pricing Implications: The CUPSA-SDF

Classic asset pricing theory (see, e.g., (Hansen and Jagannathan, 1991)) establishes an important connection between efficient portfolios and the tradable stochastic discount factor. By direct calculation, the infeasible portfolio $\pi_* = E[FF']^{-1}E[F]$ (see (3)) can be used to define the unique tradable SDF

$$M^*_{t+1} \;=\; 1 \;-\; \pi'_* F_{t+1} \tag{61}$$

[33] The virtue of complexity (Kelly et al., 2022; Didisheim et al., 2023) is the fact that the cost of statistical estimation error for complex models is lower than the gains from their better expressive ability. Formally, it states that *more complex models work better OOS*.

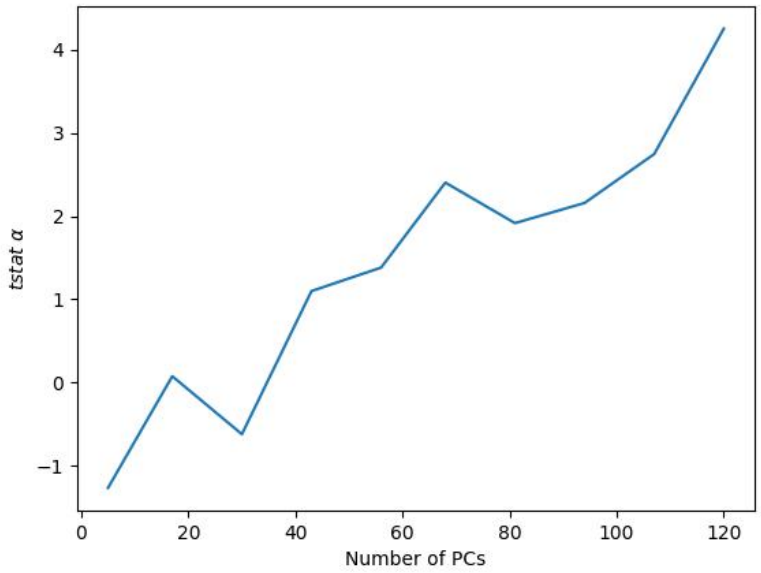

Figure 2: We plot the Heteroskedasticity-adjusted (with five lags) t-statistics of α of PC portfolios as the number of PCs grows. The regression is done using (57) with the corresponding PC portfolios. PC portfolio returns are derived from (59). Results are from the period 1981-12-31 to 2022-12-31. Portfolios are estimated to have a rolling window of $T = 120$ months and are re-balanced monthly.

satisfying the *zero pricing errors condition*

$$E[F_{i,t+1}\, M_{t+1}] \;=\; 0,\; i=1,\cdots,N\,. \tag{62}$$

Mechanically, the same calculation implies that the naive IS Markowitz portfolio $\bar{\pi}(f_0) = \bar{E}[FF']^{-1}\bar{E}[F]$ gives zero IS pricing errors: With $M_{t+1}(f_0) = 1 - \bar{\pi}(f_0)'F_{t+1}$, we have

$$\bar{E}[F_{i,t+1}\, M_{t+1}(f_0)] \;=\; 0,\; i=1,\cdots,N\,. \tag{63}$$

However, complexity implies that the OOS pricing errors are non-zero when $c > 0$. *To minimize OOS pricing errors, we need to build portfolios $\bar{\pi}$ that work OOS.* Given that CUPSA is our best feasible counter-part for the efficient portfolio *that is specifically trained to optimize OOS performance*, we can build the corresponding SDF:

$$M_{t+1}(CUPSA) \;=\; 1 \;-\; \bar{\pi}(f_{CUPSA})'F_{t+1}\,. \tag{64}$$

Intuitively, we expect CUPSA to produce small OOS pricing errors (while at the same time severely failing the IS pricing equation (63)). The goal of this section is to test this prediction.

Following (Hansen and Jagannathan, 1991), we use the Hansen-Jagannathan distance as a test statistic for measuring the OOS performance of SDFs. This distance is computed as follows. Given the out-of-sample period[34] of size T_{OOS}, we define

$$\bar{E}_{OOS}[X] \;=\; \frac{1}{T_{OOS}}\sum_{t=T+1}^{T_{OOS}} X_t\,, \tag{65}$$

and then the OOS pricing errors are defined as

$$PE_{OOS}(i) \;=\; \bar{E}_{OOS}[F_{i,t+1}M_{t+1}]\;,\; PE_{OOS} \;=\; (PE_{OOS}(i))_{i=1}^{N}\,. \tag{66}$$

The Hansen-Jagannathan distance is then defined using a weight matrix A (judiciously

[34] In the complex regime where $c = N/T > 0$, it is crucial to work only with OOS quantities.

chosen by the researcher) as

$$D^{HJ}_{OOS}(A) \;=\; (PE_{OOS})'A(PE_{OOS})\,. \tag{67}$$

If our goal is to price all asset returns F_{t+1} jointly, (Hansen and Jagannathan, 1991) advocate the use of of the weight matrix $A = E[FF']^{-1}$. However, since the latter is not observable, the computation of a correct HJ distance depends in a very subtle fashion on the choice of the matrix A.

As (Didisheim et al., 2023) argue, with non-zero complexity $c = N/T$, the most intuitive choice of A is the *OOS error matrix* $A = \bar{E}_{OOS}[FF']^{-1}$. Indeed, given a candidate estimator π_t of the infeasible portfolio π_*, we can define the estimated SDF, $M_{t+1} = 1 - \pi_t' F_{t+1}$, and evaluate its performance by computing D^{HJ}_{OOS}. In this case, as (Didisheim et al., 2023) show, the distance D^{HJ}_{OOS} with $A = \bar{E}_{OOS}[FF']^{-1}$ coincides with a constant minus the squared Sharpe ratio of the $\pi_t' F_{t+1}$ portfolio. Thus, the best-performing portfolio OOS also automatically achieves the lowest OOS pricing errors.

This result of (Didisheim et al., 2023) implies that the large gains in the Sharpe ratio produced by CUPSA (see Figure 1) should translate directly into significantly lower pricing errors. In other words, CUPSA-SDF should be better able to price the cross-section of factor returns. We now take a deeper look into the precise nature of these pricing error reductions.

Our goal is to understand how CUPSA achieves it and where the improvements are most noticeable. To do this, we use the (Jensen et al., 2023) approach and aggregate the 153 factors into 13 intuitive themes: Skewness, Profitability, Low Risk, Value, Investment, Seasonality, Debt Issuance, Size, Accruals, Low Leverage, Profit Growth, Momentum, and Quality. We follow this approach and compute theme-specific pricing errors for the CUPSA-SDF and its competitors, defined as

$$M_{t+1}(f(t-T+1,t)) = 1 - \alpha_f R_{t+1}(f(t-T+1,t)), \tag{68}$$

where $f \in \{CUPSA,\ LW,\ KNS,\ Best\ z\}$ and the optimal scaling[35] α_f is

$$\alpha_f \ = \ \frac{\bar{E}_{OOS}[R_{t+1}(f(t-T+1,t))]}{\bar{E}_{OOS}[R_{t+1}(f(t-T+1,t))^2]}. \tag{69}$$

For factors $i \in theme_j$, we define the OOS pricing error vector for $theme_j$ as

$$PE_j(f) \ = \ (\bar{E}_{OOS}[F_{i,t+1}M_{t+1}(f(t-T+1,t))])_{i \in theme_j}, \tag{70}$$

where $\bar{E}_{OOS}$ is the expectation over the full OOS sample period. Next, we aggregate pricing errors using the OOS factor covariance matrix of $theme_j$ as weights

$$D^{HJ}_{theme_j}(f) \ = \ PE_j(f)' \bar{E}_{OOS}[F_{theme_j} F'_{theme_j}]^{-1} PE_j(f), \tag{71}$$

where,

$$F_{theme_j} \ = \ (F_i)_{i \in theme_j}. \tag{72}$$

Figure 3 reports these errors for all themes. We make several observations. First, the pricing error reductions achieved by shrinkage are very large. CUPSA dominates all alternatives for every single theme. We note that the low-risk and momentum factors are commonly viewed as difficult to price because they are "anomalies" and do not reflect compensation for risk. Figure 3 suggests that these results might be driven by inefficient shrinkage.

5.5 Understanding CUPSA

How does CUPSA achieve its performance? How does it select the optimal weights W_{CUPSA}? To answer these questions, we report in Figures 4 and 5 the dynamics of the weight vector, $W_{CUPSA}(t-T+1,t)$ and the optimal Ridge penalty $z_*(t-T+1,t)$ over time.

Both z_* and the shrinkage values $z \in Z$ with non-zero W_{CUPSA} weights seem to be similar. Both are relatively stable over time around some Ridge shrinkage value. However,

[35] The optimal calling is derived in (Didisheim et al., 2023)

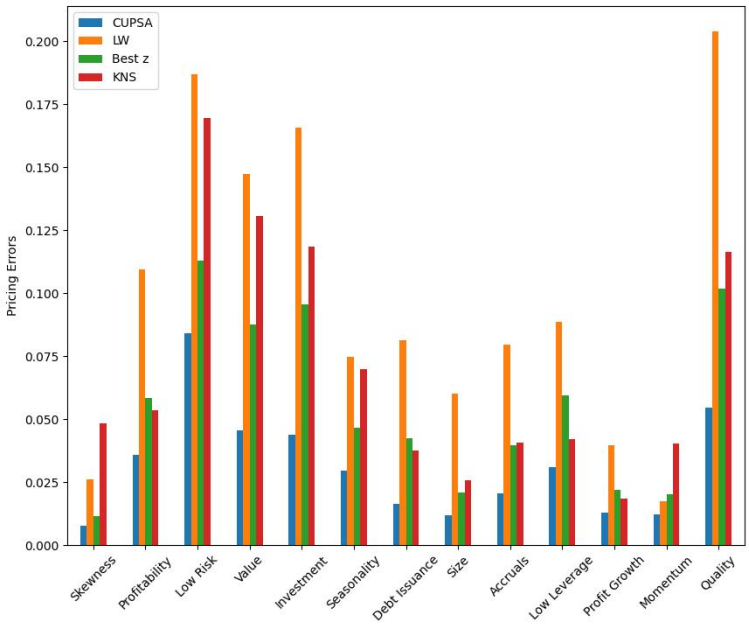

Figure 3: This figure shows the OOS HJ distance, (71), using SDFs from (68), for CUPSA (Theorem 1), LW (Ledoit and Wolf, 2020), KNS (56), and "Best z" (z_* of (55)), over the period 1981-12-31 to 2022-12-31,. Pricing errors are aggregated over themes as in (Jensen et al., 2023). Portfolios are estimated with a rolling window of $T = 120$ months and re-balanced monthly.

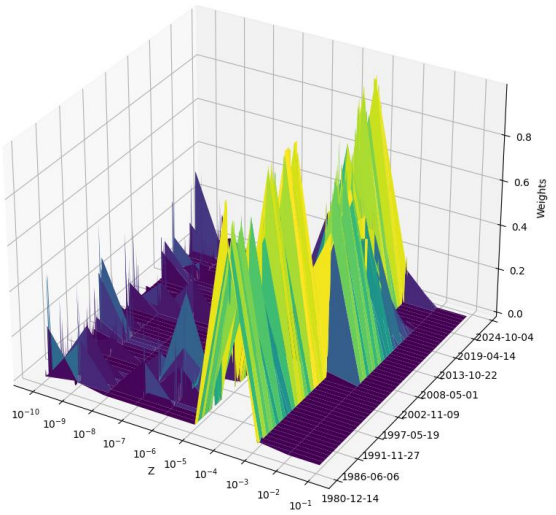

Figure 4: Weights for each Ridge portfolio associated with the CUPSA strategy, $W_{CUPSA}(t-T+1,t)$. The weights are determined using Theorem 1, computed over the period 1981-12-31 to 2022-12-31. Portfolios are estimated with a rolling window of $T = 120$ months and re-balanced monthly.

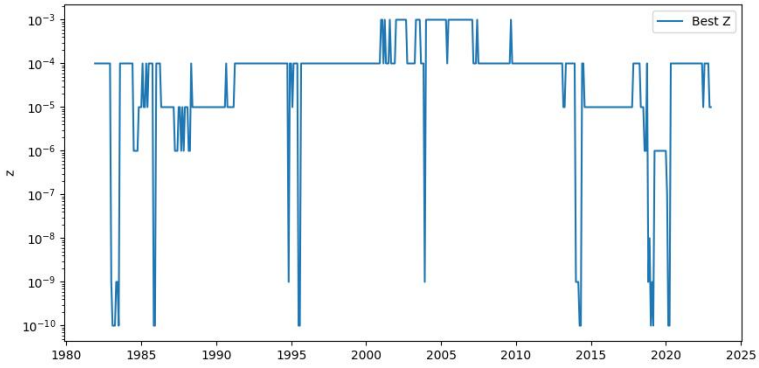

Figure 5: The figure shows the time series of optimal Ridge shrinkage, $z_*(t-T+1,t)$. Optimal Ridge shrinkage is chosen using Corollary 4 and computed over the period 1981-12-31 to 2022-12-31. Portfolios are estimated with a rolling window of $T = 120$ months and re-balanced monthly.

while z_* for "Best z" oscillates between different values, CUPSA selects an optimal convex combination of these same Ridge portfolios. As we argue in Section 4.2, this behavior is consistent with a time-varying degree of uncertainty about factor risk premia that CUPSA is able to capture by efficiently blending the different shrinkages together. These effects become particularly apparent in Figure 6 showing the histogram of the number of times a given Ridge penalty has been chosen over the whole period, overlapped with the average Ridge weights for the CUPSA strategy, $W_{CUPSA}(t-T+1,t)$. As one can see, the two distributions are almost identical, suggesting that, on average, non-linear and linear shrinkage exhibit very similar behavior. The critical distinction, however, lies in the fact that non-linear shrinkage can strategically navigate through different levels of z to secure diversification benefits. In contrast, the optimal z_* is confined to a single shrinkage level, resulting in going back and forth between different shrinkages in the vain hope of finding the ideal one.

In order to gain a deeper understanding of the optimal shrinkage function, we report the historical average of $1/f_{CUPSA}(\lambda)$ and $1/f_{z_*}(\lambda)$ in Figure 7. For large eigenvalues, $1/f(\lambda)$ is approximately linear: Both CUPSA and Best z find it optimal to keep the large eigenvalues as they are: $1/f(\lambda) \approx \lambda$ when λ is large. By contrast, the two shrinkage functions differ drastically in the way they treat low eigenvalues. While Best z simply shrinks them to a constant (completely annihilating information contained in the dispersion of these eigenvalues), CUPSA implements an optimal, convex shrinkage function that only flattens for very low eigenvalues. This ability to efficiently and non-parametrically transform lower eigenvalues is key to the performance of CUPSA.

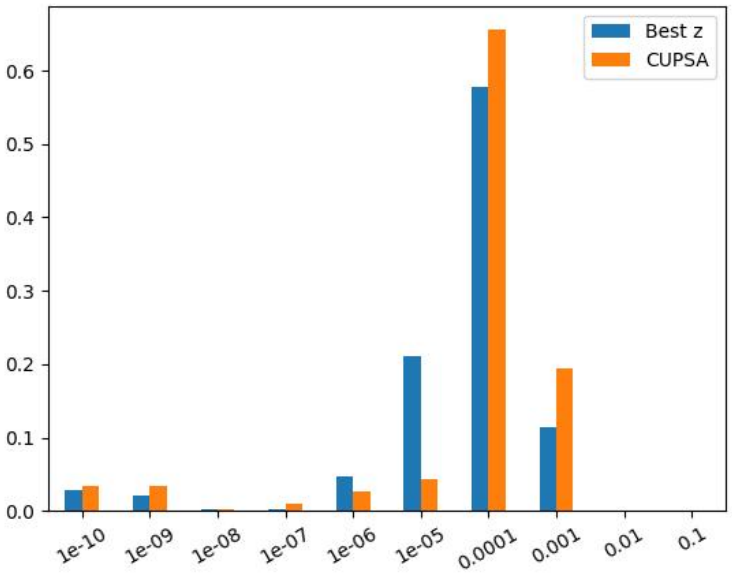

Figure 6: Histogram of the number of times a given Ridge penalty, $z_*(t-T+1,t)$, has been chosen, overlapped with the average Ridge weights, $W_{CUPSA}(t-T+1,t)$, for the CUPSA strategy. Optimal Ridge is chosen using Corollary 4, and CUPSA weights are determined using Theorem 1. Both are computed over the period 1981-12-31 to 2022-12-31. Portfolios are estimated with a rolling window of $T = 120$ months and re-balanced monthly.

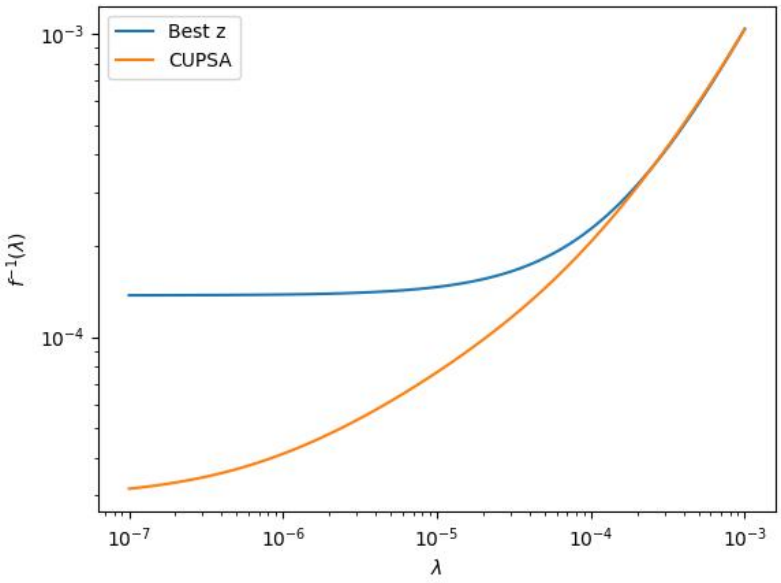

Figure 7: Historical averages of $1/f_{CUPSA}(\lambda)$ and $1/f_{z_*}(\lambda)$. Optimal Ridge is chosen using Corollary 4, and CUPSA weights are determined using Theorem 1. Both are computed over the period 1981-12-31 to 2022-12-31.

6 Conclusions

Both the problem of finding an efficient portfolio and the problem of finding a stochastic discount factor (SDF) that correctly prices all securities face the same, purely statistical, hurdle: Complexity. Whether we deal with thousands of single stocks in an unconditional setting or with hundreds of factors to construct conditional SDF, we need to estimate the number of parameters (the vector of means and the covariance) that drastically exceed the number of observations. Conventional ways of dealing with this statistical complexity involve imposing a form of sparsity on the data-generating process, reducing the dimensionality of the problem. While the characteristics-based sparsity has largely failed in capturing the complex predictive relationships in economic and financial variables (Giannone et al., 2021; Jensen et al., 2023; Kelly et al., 2022), several papers (see, e.g., (Kozak et al., 2018, 2020)) argue that the cross-section of asset returns can be characterized using an SDF that is sparse in the space of Principal Components: A PC-sparse SDF obtained through an extreme form of shrinkage, annihilating all but a few top PCs of the hundreds of factors discovered in

the asset pricing literature. In this paper, we introduce a novel, non-linear, constrained universal portfolio shrinkage approximator (CUPSA) that, instead of completely removing low-variance PCs, optimally weights them, considering their estimated out-of-sample risk-return tradeoffs. We empirically evaluate CUPSA by using it to construct the conditional SDF from a large set of factors (characteristics-based portfolios from (Jensen et al., 2023)). We find that (1) CUPSA significantly outperforms other portfolio shrinkage methodologies; (2) exhibits a *virtue of complexity*, with its out-performance (relative to standard shrinkage estimators) monotonically increasing in the number of PCs used for the SDF construction. The ability of CUPSA to exploit low-variance PCs depends on its capacity to weight these PCs efficiently, adjusting for their risk-return tradeoffs. While standard shrinkage estimators (e.g., Ridge) suggest that the optimal SDF should be PC-sparse when selected ex-post, our results imply that finding the right degree of sparsity ex-ante might be infeasible due to estimation complexity. Instead, one should keep the low-variance PCs but shrink them efficiently.

References

Arnott, Robert D, Vitali Kalesnik, and Juhani T Linnainmaa, "Factor momentum," *The Review of Financial Studies*, 2023, *36* (8), 3034–3070.

Barillas, Francisco and Jay Shanken, "Comparing asset pricing models," *The Journal of Finance*, 2018, *73* (2), 715–754.

Bryzgalova, Svetlana, Jiantao Huang, and Christian Julliard, "Bayesian solutions for the factor zoo: We just ran two quadrillion models," *The Journal of Finance*, 2023, *78* (1), 487–557.

___, Victor DeMiguel, Sicong Li, and Markus Pelger, "Asset-Pricing Factors with Economic Targets," *Available at SSRN 4344837*, 2023.

Chamberlain, Gary and Michael Rothschild, "Arbitrage, factor structure, and mean-variance analysis on large asset markets," 1982.

Cochrane, John H, "Presidential address: Discount rates," *The Journal of finance*, 2011, *66* (4), 1047–1108.

Da, Rui, Stefan Nagel, and Dacheng Xiu, "The Statistical Limit of Arbitrage," Technical Report, Chicago Booth 2022.

DeMiguel, Victor, Lorenzo Garlappi, Francisco J Nogales, and Raman Uppal, "A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms," *Management Science*, 2009, *55*, 798–812.

Didisheim, Antoine, Shikun Barry Ke, Bryan T Kelly, and Semyon Malamud, "Complexity in factor pricing models," Technical Report, National Bureau of Economic Research 2023.

Fama, Eugene F and Kenneth R French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 1993, *33*, 3–56.

___ and ___, "A five-factor asset pricing model," *Journal of financial economics*, 2015, *116* (1), 1–22.

Giannone, Domenico, Michele Lenza, and Giorgio E Primiceri, "Economic predictions with big data: The illusion of sparsity," *Econometrica*, 2021, *89* (5), 2409–2437.

Giglio, Stefano and Dacheng Xiu, "Asset pricing with omitted factors," *Journal of Political Economy*, 2021, *129* (7), 1947–1990.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, "Autoencoder asset pricing models," *Journal of Econometrics*, 2021, *222* (1), 429–450.

Gupta, Tarun and Bryan Kelly, "Factor momentum everywhere," *The Journal of Portfolio Management*, 2019, *45* (3), 13–36.

Hansen, Lars Peter and Ravi Jagannathan, "Implications of security market data for models of dynamic economies," *Journal of political economy*, 1991, *99* (2), 225–262.

___ **and Scott F Richard**, "The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models," *Econometrica: Journal of the Econometric Society*, 1987, pp. 587–613.

Harvey, Campbell R, Yan Liu, and Hao Zhu, "... and the cross-section of expected returns," *Review of Financial Studies*, 2016, *29*, 5–68.

Hastie, Trevor, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," *arXiv preprint arXiv:1903.08560*, 2019.

Hou, Kewei, Chen Xue, and Lu Zhang, "Digesting anomalies: An investment approach," *Review of Financial Studies*, 2015, *28*, 650–705.

Jegadeesh, Narasimhan and Sheridan Titman, "Returns to buying winners and selling losers: Implications for stock market efficiency," *Journal of Finance*, 1993, *48*, 65–91.

Jensen, Theis Ingerslev, Bryan Kelly, and Lasse Heje Pedersen, "Is there a replication crisis in finance?," *The Journal of Finance*, 2023, *78* (5), 2465–2518.

Kelly, Bryan, Seth Pruitt, and Yinan Su, "Instrumented Principal Component Analysis," *Working paper*, 2020.

Kelly, Bryan T, Semyon Malamud, and Kangying Zhou, "The virtue of complexity in return prediction," Technical Report, National Bureau of Economic Research 2022.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, "Interpreting factor models," *The Journal of Finance*, 2018, *73* (3), 1183–1223.

___ **,** ___ **, and** ___ , "Shrinking the cross-section," *Journal of Financial Economics*, 2020, *135* (2), 271–292.

Ledoit, Olivier and Michael Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *Journal of Empirical Finance*, 2003, *10*, 603–621.

___ **and** ___ , "Honey, I shrunk the sample covariance matrix," *Journal of Portfolio Management*, 2004, *30*, 110–119.

___ **and** ___ , "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of multivariate analysis*, 2004, *88* (2), 365–411.

___ **and** ___ , "Nonlinear shrinkage estimation of large-dimensional covariance matrices," *The Annals of Statistics*, 2012, *40* (2), 1024–1060.

___ **and** ___ , "Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions," *Journal of Multivariate Analysis*, 2015, *139*, 360–384.

___ **and** ___ , "Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks," *The Review of Financial Studies*, 2017, *30* (12), 4349–4388.

___ **and** ___ , "Analytical nonlinear shrinkage of large-dimensional covariance matrices," *The Annals of Statistics*, 2020, *48* (5), 3043–3065.

___ **and Sandrine Péché**, "Eigenvectors of some large sample covariance matrix ensembles," *Probability Theory and Related Fields*, 2011, *150*, 233–264.

Lettau, Martin and Markus Pelger, "Factors that fit the time series and cross-section of stock returns," *The Review of Financial Studies*, 2020, *33* (5), 2274–2325.

Löwner, Karl, "Über monotone matrixfunktionen," *Mathematische Zeitschrift*, 1934, *38* (1), 177–216.

Markowitz, Harry, "Portfolio Selection," *The Journal of Finance*, 1952, *7* (1), 77–91.

Martin, Ian WR and Stefan Nagel, "Market efficiency in the age of big data," *Journal of Financial Economics*, 2021.

Murphy, Kevin P, "Conjugate Bayesian analysis of the Gaussian distribution," *def*, 2007, *1* ($2\sigma 2$), 16.

Patil, Pratik, Yuting Wei, Alessandro Rinaldo, and Ryan Tibshirani, "Uniform consistency of cross-validation estimators for high-dimensional ridge regression," in "International Conference on Artificial Intelligence and Statistics" PMLR 2021, pp. 3178–3186.

Pedersen, Lasse Heje, Abhilash Babu, and Ari Levine, "Enhanced Portfolio Optimization," *Working paper, AQR and Copenhagen Business School*, 2020.

Preite, Massimo Dello, Raman Uppal, Paolo Zaffaroni, and Irina Zviadadze, "What is Missing in Asset-Pricing Factor Models?," 2022.

Quaini, Alberto and Fabio Trojani, "Proximal Estimation and Inference," *arXiv preprint arXiv:2205.13469*, 2022.

Ross, Stephen A., "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, 1976, *13*, 341–360.

Rudin, Walter, *Principles of Mathematical Analysis*, 3 ed., New York: McGraw-Hill, 1976. See Chapter 7 for the Stone-Weierstrass Theorem.

Stein, Charles, "Lectures on the theory of estimation of many parameters," *Journal of Soviet Mathematics*, 1986, *34*, 1373–1403.

Tibshirani, Robert, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1996, *58* (1), 267–288.

A Proofs

Proof of Lemma 1. Interchangeability implies that the joint distributions $((F_i)_{i\neq\tau,1\leq i\leq T}, F_\tau)$ and $((F_i)_{i,1\leq i\leq T}, (F_t)_{t>T}$ are the same. Hence, the joint distributions of $(\bar{\pi}_{T,\tau}(f), F_\tau)$ and $(\bar{\pi}(f), (F_t)_{t>T})$ are also the same. Therefore,

$$\begin{aligned} E[U(R_{T,\tau}(f))] &= E[U(\bar{\pi}_{T,\tau}(f)'F_\tau)] \\ &= E[U(\bar{\pi}(f)'F_t)] \\ &= E[U(R_t(f)] \end{aligned} \tag{73}$$

and,

$$\begin{aligned} E\left[U^{OOS}_{LOO}(f)\right] &= E[\frac{1}{T}\sum_{\tau=1}^{T} U(R_{T,\tau}(f))] \\ &= \frac{1}{T}\sum_{\tau=1}^{T} E[U(R_T(f)] \\ &= E[U(R_t(f)]. \end{aligned} \tag{74}$$

This concludes the proof of Lemma 1. □

Proof of Lemma 2.

$$R_{T,\tau}(f_z) = F'_\tau\ \bar{\pi}_{T,t}(f_z) = F'_\tau(zI + \bar{E}_{T,\tau}[FF'])^{-1}\bar{E}_{T,\tau}[F_t]. \tag{75}$$

Therefore, it suffices to calculate

$$F'_\tau(zI + \bar{E}_{T,\tau}[FF'])^{-1}\bar{E}_{T,\tau}[F_t] = \frac{1}{T}\sum_{t\neq\tau}^{T} F'_\tau(zI + \bar{E}_{T,\tau}[FF'])^{-1}F_t. \tag{76}$$

Placing $A = \bar{E}[FF'] + zI$, $u = F_\tau$, and $v = -\frac{1}{T}F_\tau$ in Lemma 8 gives us:

$$(zI + \bar{E}_{T,\tau}[FF'])^{-1} = (zI + \bar{E}[FF'])^{-1} + \frac{1}{T}\frac{(zI + \bar{E}[FF'])^{-1}F_\tau F'_\tau(zI + \bar{E}[FF'])^{-1}}{1 - \frac{1}{T}F'_\tau(zI + \bar{E}[FF'])^{-1}F_\tau} \tag{77}$$

Multiplying F_τ to both sides gives

$$\begin{aligned} F_\tau'(zI+\bar{E}_{T,\tau}[FF'])^{-1} &= F_\tau'(zI+\bar{E}[FF'])^{-1}+\frac{1}{T}\frac{F_\tau'(zI+\bar{E}[FF'])^{-1}F_\tau F_\tau'(zI+\bar{E}[FF'])^{-1}}{1-\frac{1}{T}F_\tau'(zI+\bar{E}[FF'])^{-1}F_\tau} \\ &= \frac{F_\tau'(zI+\bar{E}[FF'])^{-1}}{1-\frac{1}{T}F_\tau'(zI+\bar{E}[FF'])^{-1}F_\tau}\,. \end{aligned} \tag{78}$$

Combining (76) and (77), we get

$$\frac{1}{T}\sum_{t\neq\tau}^{T}F_\tau'(zI+\bar{E}_{T,\tau}[FF'])^{-1}F_t = \frac{1}{T}\sum_{t\neq\tau}^{T}\frac{F_\tau'(zI+\bar{E}[FF'])^{-1}F_t}{1-\frac{1}{T}F_\tau'(zI+\bar{E}[FF'])^{-1}F_\tau} \tag{79}$$

If we define $\psi_\tau(z)=\frac{1}{T}F_\tau'(zI+\bar{E}[FF'])^{-1}F_\tau$, this concludes the proof of Lemma 2. □

Proof of Lemma 3.

$$\begin{aligned} \psi_\tau(z) &= \frac{1}{T}F_\tau'(zI+\bar{E}[FF'])^{-1}F_\tau \\ &\leq \frac{1}{T}\|F_\tau'\|^2\|(zI+\bar{E}[FF'])^{-1}\| \\ &\leq \frac{1}{T}F_\tau'F_\tau z^{-1} \\ &\leq \frac{1}{T}K^2Pz^{-1} \\ &\leq cK^2z^{-1} \end{aligned} \tag{80}$$

and

$$\begin{aligned} \psi_\tau(z) &= \frac{1}{T}F_\tau'(zI+\bar{E}[FF'])^{-1}F_\tau \\ &\underbrace{=}_{Lemma\ 8}\frac{T^{-1}F_\tau'(zI+\bar{E}_{T,\tau}[FF'])^{-1}F_\tau}{1+T^{-1}F_\tau'(zI+\bar{E}_{T,\tau}[FF'])^{-1}F_\tau} \\ &\leq 1 \end{aligned} \tag{81}$$

The proof of Lemma 3 is complete. □

Proof of Lemma 4. We optimize directly on $U_{LOO}^{OOS}(f_{Z,W})$ (Lemma 1), hence

$$\begin{aligned} U_{LOO}^{OOS}(f_{Z,W}) &= \frac{1}{T}\sum_{\tau=1}^{T} U(R_{T,\tau}(f_{Z,W})) \\ &= \frac{1}{T}\sum_{\tau=1}^{T}(R_{T,\tau}(f_{Z,W}) - \frac{1}{2}R_{T,\tau}(f_{Z,W})^2) \\ &= \frac{1}{T}\sum_{\tau=1}^{T}(W'R_{T,\tau}(f_Z) - \frac{1}{2}W'R_{T,\tau}(f_Z)R_{T,\tau}(f_Z)'W) \\ &= W'\big(\frac{1}{T}\sum_{\tau=1}^{T} R_{T,\tau}(f_Z)\big) - \frac{1}{2}W'\big(\frac{1}{T}\sum_{\tau=1}^{T} R_{T,\tau}(f_Z)R_{T,\tau}(f_Z)'\big)W \\ &= W'\bar{\mu}(Z) - \frac{1}{2}W'\bar{\Sigma}(Z)W \end{aligned} \tag{82}$$

This completes the proof of Lemma 4. □

Proof of Lemma 5. The proof of the first statement of the lemma relies on an application of the Stone-Weierstrass Theorem; see, e.g., Rudin (1976). Consider the algebra of functions generated by the Ridge ensemble $\{\Theta_z : z > 0\}$, where $\Theta_z(x) := (z+x)^{-1}$ for any $x \geq 0$. These functions are bounded, strictly monotonically decreasing, and continuous on $\mathbb{R}_+$. Using the identity

$$\Theta_{z_1}(x) - \Theta_{z_2}(x) = (z_2 - z_1)\Theta_{z_1}(x)\Theta_{z_2}(x), \tag{83}$$

it follows that on any compact interval $[a, b]$ the linear span of the Ridge ensemble is dense in the algebra generated by the Ridge ensemble. Moreover, it is easy to see that the Ridge ensemble separates points on any compact interval $[a, b]$, and it vanishes nowhere. As a consequence, for any compact interval $[a, b]$ the algebra generated by the Ridge ensemble is dense in $C(a, b)$ – by the Stone-Weierstrass Theorem – and the claim follows. To prove the second statement of the lemma, let f be a real-valued, non-negative, matrix monotone decreasing function on $\mathbb{R}_+$ so that $g := -f$ is negative and matrix monotone increasing. Using (Löwner, 1934) Theorem, it follows that there exist constants $a \in \mathbb{R}$, $b > 0$ and a

positive finite measure μ on $\mathbb{R}_+$ such that, for $\lambda \geq 0$:

$$g(\lambda) = a + \lambda b + \int_0^\infty \frac{\lambda}{z+\lambda} d\mu(z) = a + \lambda(b + \mu(\mathbb{R}_+)) - \int_0^\infty \frac{z}{z+\lambda} d\mu(z)\,. \tag{84}$$

By monotone convergence, $0 = \lim_{\lambda\to\infty} \int_0^\infty \frac{z}{z+\lambda} d\mu(z)$. Therefore,

$$0 = \lim_{\lambda\to\infty} g(\lambda) = \lim_{\lambda\to\infty} \left(a + \lambda \left(b + \mu(\mathbb{R}_+)\right)\right)\ , \tag{85}$$

which implies $a = 0$ and $b + \mu(\mathbb{R}_+) = 0$. This gives the representation:

$$f(\lambda) = \int_0^\infty \frac{z}{z+\lambda} d\mu(z) =: \int_0^\infty \frac{1}{z+\lambda} d\nu(z)\ , \tag{86}$$

for a positive measure ν on $\mathbb{R}_+$ having Radon-Nykodim derivative $\frac{d\nu}{d\mu}(z) = z$ with respect to μ. Hence, function f is the Stieltjes transform of positive measure ν. By assumption, $\lambda f(\lambda)$ is bounded, and hence there exists a constant $K > 0$ such that:

$$K \geq \sup_{\lambda\geq 0} \lambda f(\lambda) \geq \lim_{\lambda\to\infty} \int_0^\infty \frac{\lambda}{z+\lambda} d\nu(z) = \int_0^\infty d\nu(z)\ , \tag{87}$$

using in the last identity the monotone convergence theorem. Hence, f is the Stielties transform of positive finite measure ν. Equivalently, we can also write:

$$f(\lambda) = \nu(\mathbb{R}_+) \int_0^\infty \frac{1}{z+\lambda} dP(z)\ , \tag{88}$$

using probability measure $P := \nu/\nu(\mathbb{R}_+)$ on $\mathbb{R}_+$. Next, let c_z be a strictly positive scaling function such that $\int_0^\infty c_z^{-1} d\nu(z) < \infty$, which is a condition obviously satisfied by any scaling function such that $c_\cdot^{-1}$ is bounded. It then follows:

$$f(\lambda) = \int_0^\infty \frac{c_z}{z+\lambda} c_z^{-1} d\nu(z) =: \int_0^\infty \frac{c_z}{z+\lambda} d\tilde{\nu}(z)\ , \tag{89}$$

for a positive finite measure $\tilde{\nu}$ on $\mathbb{R}_+$ having Radon-Nykodim derivative $\frac{d\tilde{\nu}}{d\nu}(z) = c_z^{-1}$ with respect to ν. Equivalently, we can also write:

$$f(\lambda) = \tilde{\nu}(\mathbb{R}_+) \int_0^\infty \frac{c_z}{z+\lambda} d\tilde{P}(z)\ , \tag{90}$$

using probability measure $\tilde{P} := \tilde{\nu}/\tilde{\nu}(\mathbb{R}_+)$ on $\mathbb{R}_+$. Pick $\lambda_0 > 0$. By assumption, $c(z) = (z+a)/a$ for some $a > 0$ and, hence $c(z)$ is continuous, and $\psi(z) = c_z/(z+\lambda)$ is bounded, $\psi(z) < A_0(\lambda)$, and has a bounded derivative for $\lambda > \lambda_0$:

$$|\psi'(z)| \;=\; a^{-1}\frac{|a-\lambda|}{(z+\lambda)^2} \;\leq\; a^{-1}|a-\lambda|\lambda^{-2} \;\leq\; a^{-1}\max(1,|a/\lambda_0-1|)\lambda_0^{-1} \;=\; A_1(\lambda_0) \quad (91)$$

Let also $A(\lambda_0) = \max(A_0(\lambda_0), A_1(\lambda_0))$. Then, the integral can be approximated by Rieman sums. Pick an $\varepsilon > 0$. Pick a large constant $Z > 0$, such that $\tilde{P}([Z,+\infty) < \varepsilon$. Let also $0 = z_1 < \cdots < z_n = Z,\ z_{n+1} = +\infty$ be a partition of $[0,Z]$ with $|z_{i+1} - z_i| < \varepsilon$ for all i. Then,

$$|\frac{c_z}{z+\lambda} - \frac{c_{z_i}}{z_i+\lambda}| \;\leq\; \varepsilon A(\lambda_0)\ ,\ z \in [z_i, z_{i+1}]\,. \quad (92)$$

Let $p_i = \tilde{P}([z_i, z_{i+1}))$. Then,

$$\begin{aligned}&\left|\frac{1}{\tilde{\nu}(\mathbb{R}_+)}f(\lambda) - \sum_{i=1}^{n}\frac{c_{z_i}}{z_i+\lambda}p_i\right| \\ &= \;|\sum_i \int_{z_i}^{z_{i+1}}(\frac{c_z}{z+\lambda} - \frac{c_{z_i}}{z_i+\lambda})d\tilde{P}(z)| \\ &\leq\; \varepsilon A(\lambda_0)\,.\end{aligned} \quad (93)$$

□

Remark 6 (Remarks on scaling) From the above proof, real-valued, non-negative, monotonically decreasing functions on $\mathbb{R}_+$ that vanish at $+\infty$ have the (pointwise) representation:

$$f(\lambda) = \int_0^\infty \frac{z}{z+\lambda}d\mu(z) = \mu(\mathbb{R}_+)\int_0^\infty \frac{z}{z+\lambda}dP(z)\ , \quad (94)$$

for a corresponding finite positive measure μ. If furthermore, $f(\lambda) = O(1/\lambda)$ as $\lambda \to \infty$ then:

$$f(\lambda) = \int_0^\infty \frac{1}{z+\lambda}d\tilde{\mu}(z) = \tilde{\mu}(\mathbb{R}_+)\int_0^\infty \frac{z}{z+\lambda}d\tilde{P}(z)\ , \quad (95)$$

for a finite positive measure $\tilde{\mu}$ on $\mathbb{R}_+$ having Radon-Nykodim derivative $\frac{d\tilde{\mu}}{d\mu}(z) = z$ with

respect to μ. In particular, this shows that function f has various equivalent representations using different pairs of kernels and finite measures on $\mathbb{R}_+$. Furthermore, note that since in our setting f is the argument of a maximum Sharpe ratio optimization problem, scaling f by a positive constant does not affect the optimal portfolio solution, i.e., a normalization is necessary to identify the optimally shrunk portfolio across real-valued, non-negative, monotonically decreasing shrinkage functions. Since $z/(\lambda + z) \in [0, 1)$, we adopt the normalization $\mu(\mathbb{R}_+) = 1$ and write:

$$f(\lambda) = \int_0^\infty \frac{z}{z+\lambda} dP(z) \ , \tag{96}$$

for some associated probability measure P on $\mathbb{R}_+$. The natural discrete approximation to integral (97) using a discrete grid $z_1, \ldots, z_n$ with probability weights $w_1, \ldots, w_n$ is:

$$f(\lambda) \approx \sum_{i=1}^n \frac{z_i}{z_i+\lambda} w_i \ . \tag{97}$$

Proof of Theorem 5. Note that $\bar{E}[R(f_Z)] = \bar{E}[F]'(ZI + \bar{E}[FF'])^{-1}\bar{E}[F]$ while

$$\bar{E}[R(f_Z)R(f_Z)'] = \bar{E}[F]'(Z_1 I + \bar{E}[FF'])^{-1}\bar{E}[FF'](Z_2 I + \bar{E}[FF'])^{-1}\bar{E}[F] \tag{98}$$

Thus, if $Z_0 = 0$, we have

$$\begin{aligned}(\bar{E}[R(f_Z)R(f_Z)']e_0)_i &= \bar{E}[F]'(z_i I + \bar{E}[FF'])^{-1}\bar{E}[FF'](0I + \bar{E}[FF'])^{-1}\bar{E}[F] \\ &= \bar{E}[F]'(z_i I + \bar{E}[FF'])^{-1}\bar{E}[F] = \bar{\mu}_{IS}(z_i),\end{aligned} \tag{99}$$

implying that

$$\bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(Z) = e_0, \tag{100}$$

so that no shrinkage is optimal.

$$\begin{aligned}
\bar{\Sigma}(Z)^{-1}\bar{\mu}(Z) &= D(Z)^{-1}\big(\bar{\Sigma}_{IS}(Z) + \psi(Z)\psi(Z)' - (\bar{\mu}_{IS}(z)\psi(Z)' + \psi(Z)\bar{\mu}_{IS}(z'))\big)^{-1}(\bar{\mu}_{IS}(z) - \psi(Z)) \\
&= D(Z)^{-1}\big(\bar{\Sigma}_{IS}(Z) + \psi(Z)\psi(Z)' - (\bar{\mu}_{IS}(z)\psi(Z)' + \psi(Z)\bar{\mu}_{IS}(z'))\big)^{-1}\bar{\mu}_{IS}(z) \\
&- D(Z)^{-1}\big(\bar{\Sigma}_{IS}(Z) + \psi(Z)\psi(Z)' - (\bar{\mu}_{IS}(z)\psi(Z)' + \psi(Z)\bar{\mu}_{IS}(z'))\big)^{-1}\psi(Z) \\
&= D(Z)^{-1}term1 - D(Z)^{-1}term2
\end{aligned} \tag{101}$$

For $term2$ we have

$$\begin{aligned}
term2 &= \big(\bar{\Sigma}_{IS}(Z) + \psi(Z)\psi(Z)' - (\bar{\mu}_{IS}(z)\psi(Z)' + \psi(Z)\bar{\mu}_{IS}(z'))\big)^{-1}\psi(Z) \\
&\underbrace{=}_{Lemma\ 8} \frac{\big(\bar{\Sigma}_{IS}(Z) - (\bar{\mu}_{IS}(z)\psi(Z)' + \psi(Z)\bar{\mu}_{IS}(z'))\big)^{-1}\psi(Z)}{1 + \psi(Z)'\big(\bar{\Sigma}_{IS}(Z) - (\bar{\mu}_{IS}(z)\psi(Z)' + \psi(Z)\bar{\mu}_{IS}(z'))\big)^{-1}\psi(Z)} \\
&= \frac{1}{d_{2,1}}\big(\bar{\Sigma}_{IS}(Z) - (\bar{\mu}_{IS}(z)\psi(Z)' + \psi(Z)\bar{\mu}_{IS}(z'))\big)^{-1}\psi(Z).
\end{aligned} \tag{102}$$

Then

$$\begin{aligned}
&\big(\bar{\Sigma}_{IS}(Z) - (\bar{\mu}_{IS}(z)\psi(Z)' + \psi(Z)\bar{\mu}_{IS}(z'))\big)^{-1}\psi(Z) \\
&\underbrace{=}_{Lemma\ 8} \big(\bar{\Sigma}_{IS}(Z) - \bar{\mu}_{IS}(z)\psi(Z)'\big)^{-1}\psi(Z) \\
&+ \frac{\big(\bar{\Sigma}_{IS}(Z) - \bar{\mu}_{IS}(z)\psi(Z)'\big)^{-1}\psi(Z)\bar{\mu}_{IS}(z)'\big(\bar{\Sigma}_{IS}(Z) - \bar{\mu}_{IS}(z)\psi(Z)'\big)^{-1}\psi(Z)}{1 - \mu_{IS}(z)'\big(\bar{\Sigma}_{IS}(Z) - \bar{\mu}_{IS}(z)\psi(Z)'\big)^{-1}\psi(Z)} \\
&= \frac{\big(\bar{\Sigma}_{IS}(Z) - \bar{\mu}_{IS}(z)\psi(Z)'\big)^{-1}\psi(Z)}{1 - \mu_{IS}(z)'\big(\bar{\Sigma}_{IS}(Z) - \bar{\mu}_{IS}(z)\psi(Z)'\big)^{-1}\psi(Z)} \\
&= \frac{1}{d_{2,2}}\big(\bar{\Sigma}_{IS}(Z) - \bar{\mu}_{IS}(z)\psi(Z)'\big)^{-1}\psi(Z).
\end{aligned} \tag{103}$$

Finally,

$$
\begin{aligned}
&\big(\bar{\Sigma}_{IS}(Z) \;-\; \bar{\mu}_{IS}(z)\psi(Z)'\big)^{-1}\psi(Z) \\
&\underbrace{=}_{Lemma\ 8} \;\; \bar{\Sigma}_{IS}(Z)^{-1}\psi(Z) \\
&+\; \frac{\bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(z)\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(z)} \\
&=\; \bar{\Sigma}_{IS}(Z)^{-1}\psi(Z) + \; \frac{e_0\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(Z)'e_0} \\
&=\; \bar{\Sigma}_{IS}(Z)^{-1}\psi(Z) + \; e_0\frac{\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(0)}.
\end{aligned}
\tag{104}
$$

Therefore,

$$
\begin{aligned}
term2 \;&=\; \frac{1}{d_{2,1}d_{2,2}}\left(\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z) + \; e_0\frac{\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(0)}\right) \\
&=\; \alpha_2 e_0 + \beta_2\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)
\end{aligned}
\tag{105}
$$

For term1 the calculations will be slightly more involved:

$$
\begin{aligned}
term1 \;&=\; \big(\bar{\Sigma}_{IS}(Z) \;+\; \psi(Z)\psi(Z)' \;-\; (\bar{\mu}_{IS}(z)\psi(Z)' \;+\; \psi(Z)\bar{\mu}_{IS}(z)')\big)^{-1}\bar{\mu}_{IS}(z) \\
&\underbrace{=}_{Lemma\ 8} \;\; \big(\bar{\Sigma}_{IS}(Z) \;+\; \psi(Z)\psi(Z)' \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\bar{\mu}_{IS}(z) \\
&+\; \frac{\big(\bar{\Sigma}_{IS}(Z)+\psi(Z)\psi(Z)'-\psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\bar{\mu}_{IS}(z)\psi(Z)'\big(\bar{\Sigma}_{IS}(Z)+\psi(Z)\psi(Z)'-\psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\bar{\mu}_{IS}(z)}{1-\psi(Z)'\big(\bar{\Sigma}_{IS}(Z) \;+\; \psi(Z)\psi(Z)' \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\bar{\mu}_{IS}(z)} \\
&=\; \frac{\big(\bar{\Sigma}_{IS}(Z) \;+\; \psi(Z)\psi(Z)' \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\bar{\mu}_{IS}(z)}{1-\psi(Z)'\big(\bar{\Sigma}_{IS}(Z) \;+\; \psi(Z)\psi(Z)' \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\bar{\mu}_{IS}(z)} \\
&=\; \frac{1}{d_1}\big(\bar{\Sigma}_{IS}(Z) \;+\; \psi(Z)\psi(Z)' \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\bar{\mu}_{IS}(z).
\end{aligned}
\tag{106}
$$

Next

$$\begin{aligned}
&\left(\bar{\Sigma}_{IS}(Z) \;+\; \psi(Z)\psi(Z)' \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\right)^{-1}\bar{\mu}_{IS}(z) \\
&\underbrace{=}_{Lemma\ 8} \left(\bar{\Sigma}_{IS}(Z) \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\right)^{-1}\bar{\mu}_{IS}(z) \\
&- \frac{\left(\bar{\Sigma}_{IS}(Z) \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\right)^{-1}\psi(Z)\psi(Z)'\left(\bar{\Sigma}_{IS}(Z) \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\right)^{-1}\bar{\mu}_{IS}(z)}{1+\psi(Z)'\left(\bar{\Sigma}_{IS}(Z) \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\right)^{-1}\psi(Z)} \\
&= term11 \;-\; term12
\end{aligned} \tag{107}$$

Note that

$$\begin{aligned}
\left(\bar{\Sigma}_{IS}(Z) \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\right)^{-1}\psi(Z) \;&\underbrace{=}_{Lemma\ 8}\; \frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(z)} \\
&= \frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(Z)'e_0} \\
&= \frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(0)}.
\end{aligned} \tag{108}$$

Plugging back

$$\begin{aligned}
term12 \;&=\; \frac{\frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(z)}{(1-\psi(0))^2}}{1+\frac{\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(0)}} \\
&=\; \frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(z)}{(1-\psi(0))^2+(1-\psi(0))\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)} \\
&=\; \frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)\psi(Z)'e_0}{(1-\psi(0))^2+(1-\psi(0))\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)} \\
&=\; \frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)\psi(0)}{(1-\psi(0))^2+(1-\psi(0))\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}.
\end{aligned} \tag{109}$$

Furthermore

$$\begin{aligned}
term11 \;&=\; \left(\bar{\Sigma}_{IS}(Z) \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\right)^{-1}\bar{\mu}_{IS}(z) \\
&\underbrace{=}_{Lemma\ 8}\; \bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(z) + \frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)\bar{\mu}_{IS}(z)'\bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(z)}{1-\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(z)} \\
&=\; e_0 + \frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(0)}.
\end{aligned} \tag{110}$$

Note that d_1 is just $1 - \psi(Z)'(term11 + term12)$. This means

$$\begin{aligned} d_1 &= 1 - \psi(Z)'e_0 + \frac{\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(0)} - \frac{\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)\psi(0)}{(1-\psi(0))^2 + (1-\psi(0))\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)} \\ &= 1 - \psi(0) + \frac{\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(0)} - \frac{\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)\psi(0)}{(1-\psi(0))^2 + (1-\psi(0))\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)} \\ &= 1 - \psi(0) + \frac{qf}{1-\psi(0)} - \frac{qf\psi(0)}{(1-\psi(0))^2 + (1-\psi(0))qf} \\ &= 1 - \psi(0) + \frac{qf}{1-\psi(0)}(1 - \frac{\psi(0)}{1-\psi(0)+qf}) \\ &= 1 - \psi(0) + \frac{qf}{1-\psi(0)}(\frac{1-2\psi(0)+qf}{1-\psi(0)+qf}) \end{aligned} \tag{111}$$

Hence

$$\begin{aligned} term1 &= \frac{1}{d_1}\left(e_0 + \frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(0)} - \frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)\psi(0)}{(1-\psi(0))^2 + (1-\psi(0))\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}\right) \\ &= \alpha_1 e_0 + \beta_1 \bar{\Sigma}_{IS}(Z)^{-1}\psi(Z) \end{aligned} \tag{112}$$

$$\begin{aligned} \bar{\Sigma}(Z)^{-1}\bar{\mu}(Z) &= D(Z)^{-1}(term1 \ - \ term2) \\ &= D(Z)^{-1}(\alpha e_0 + \beta\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)) \end{aligned} \tag{113}$$

This concludes the proof of Theorem 5 □

Proof of Lemma 6. First note that

$$E[\mu|F_{IS}] = \sum_{i=1}^{L}(w_i/z_i)/\bar{w}E[\mu_i|F_{IS}] \tag{114}$$

where μ_i is sampled from $N(0, z_i I)$. We know from Murphy (2007), that under a Gaussian Mixture prior

$$E[\mu_i|F_{IS}] = (\frac{1}{z_i}I + \Sigma^{-1})^{-1}\Sigma^{-1}\bar{E}[F] = z_i(z_i I + \Sigma)^{-1}\bar{E}[F]. \tag{115}$$

Plugging back into (114) we get

$$E[\mu|F_{IS}] \;=\; \frac{1}{\bar{w}}\sum_{i=1}^{L}(w_i)(z_iI+\Sigma)^{-1}\bar{E}[F] \tag{116}$$

□

Lemma 7 *Let* $D(Z) = \text{diag}(\frac{1}{1-\psi(Z)})$, *be the complexity multiplier. Under the hypothesis of Proposition 3, we have*

$$\begin{aligned}
\bar{\mu}(Z) \;&=\; D(Z)(\bar{\mu}_{IS}(z) - \underbrace{\psi(Z)}_{overfit}) \\
\bar{\Sigma}(Z) \;&=\; D(Z)\big(\bar{\Sigma}_{IS}(Z) \;+\; \underbrace{\psi(Z)\psi(Z)'}_{variance\ \ overfit} \;-\; \underbrace{(\bar{\mu}_{IS}(z)\psi(Z)' \;+\; \psi(Z)\bar{\mu}_{IS}(z)')}_{mean\ \ overfit}\big)D(Z)
\end{aligned} \tag{117}$$

where,

$$\begin{aligned}
\bar{\mu}_{IS}(z) \;&=\; \bar{E}[R(f_Z)], \\
\bar{\Sigma}_{IS}(Z) \;&=\; \bar{E}[R(f_Z)R(f_Z)']
\end{aligned} \tag{118}$$

are the in-sample mean and second moment of the Ridge portfolios, respectively.

Proof of Lemma 7. For the mean, using formula (27) and Proposition 3 to replace ψ_t with its limit that is independent of t, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}R_{T,t}(f_z) \;&=\; \frac{1}{T}\sum_{t=1}^{T}\;\frac{1}{1-\psi_t(z)}\big(R_t(f_z) \;-\; \psi_t(z)\big) \\
&=\; \frac{1}{1-\psi(z;c)}\big(\frac{1}{T}\sum_{t=1}^{T}R_t(f_z) \;-\; \psi(z;c)\big) \\
&=\; \frac{1}{1-\psi(z;c)}\big(\bar{\mu}_{IS}(z) \;-\; \psi(z;c)\big).
\end{aligned} \tag{119}$$

In the matrix form, we can rewrite this as

$$\bar{\mu}(Z) \;=\; \text{diag}(\frac{1}{1-\psi(Z)})\big(\bar{\mu}_{IS}(z) \;-\; \psi(Z)\big)\,. \tag{120}$$

Similarly, for the second moment matrix

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} R_{T,t}(f_{z_i})R_{T,t}(f_{z_j}) &= \frac{1}{T}\sum_{t=1}^{T} \frac{1}{1-\psi_t(z_i)}\big(R_t(f_{z_i}) - \psi_t(z_i)\big)\frac{1}{1-\psi_t(z_j)}\big(R_t(f_{z_j}) - \psi_t(z_j)\big) \\
&= \frac{1}{T}\sum_{t=1}^{T} \frac{1}{1-\psi(z_i;c)}\big(R_t(f_{z_i}) - \psi(z_i;c)\big)\frac{1}{1-\psi(z_j)}\big(R_t(f_{z_j}) - \psi(z_j)\big) \\
&= \frac{1}{(1-\psi(z_j))(1-\psi(z_i;c))}\frac{1}{T}\sum_{t=1}^{T}\big(R_t(f_{z_i})R_t(f_{z_j}) + \psi(z_i;c)\psi(z_j)\big) \\
&- \frac{1}{(1-\psi(z_j))(1-\psi(z_i;c))}\frac{1}{T}\sum_{t=1}^{T}\big(R_t(f_{z_j})\psi(z_i;c) + R_t(f_{z_i})\psi(z_j)\big) \\
&= \frac{1}{(1-\psi(z_j))(1-\psi(z_i;c))}\big(\bar{\Sigma}_{IS}(z_i,z_j) + \psi(z_i;c)\psi(z_j)\big) \\
&- \frac{1}{(1-\psi(z_j))(1-\psi(z_i;c))}\big(\bar{\mu}_{IS}(z_i)\psi(z_i;c) + \bar{\mu}_{IS}(z_i)\psi(z_j)\big).
\end{aligned} \tag{121}$$

In the matrix form, this gives

$$\bar{\Sigma}(Z) = \mathrm{diag}(\frac{1}{1-\psi(Z)})\big(\bar{\Sigma}_{IS}(Z) + \psi(Z)\psi(Z)' - \bar{\mu}_{IS}(z)\psi(Z)' - \psi(Z)\bar{\mu}_{IS}(z)'\big)\,\mathrm{diag}(\frac{1}{1-\psi(Z)}) \tag{122}$$

This concludes the proof of Lemma 7. □

Proposition 7 (In-Sample Shrinkage is Not Optimal) *Assume $R_\tau(f_z)$ is the in-sample return at time τ,*

$$R_\tau(f_z) = \bar{\pi}(f_z)'F_\tau,\ \tau \le T. \tag{123}$$

Suppose

$$\bar{\mu}_{IS}(Z) = (\bar{E}[R_\tau(f_{z_i})])_{i=1}^{L} \quad \bar{\Sigma}_{IS}(Z) = (\bar{E}[R_\tau(f_{z_i})R_\tau(f_{z_i})'])_{i=1}^{L} \tag{124}$$

are the IS mean and covariance of the Ridge-shrunk portfolios, respectively. Now define

$$a(z) \;=\; \bar{\mu}_{IS}(z) - 0.5\bar{\Sigma}_{IS}(z) - 0.5 \tag{125}$$

to be the in-sample quadratic utility. $a'(z)$ *is negative for* $z > 0$ *and* $a'(0) = a''(0) = 0$. *Hence,* $a(z)$ *obtains its maximum at* $z = 0$.

Proof of Proposition 7. Let

$$\begin{aligned} a(z) \;&=\; \bar{\mu}_{IS}(z) - 0.5\bar{\Sigma}_{IS}(z) - 0.5 \\ &=\; \bar{E}[F]'(zI + \bar{E}[FF'])^{-1}\bar{E}[F] - 0.5 \\ &-\; 0.5\bar{E}[F]'(zI + \bar{E}[FF'])^{-1}\bar{E}[FF'](zI + \bar{E}[FF'])^{-1}\bar{E}[F] - 0.5 \\ &=\; \bar{E}[F]'(zI + \bar{E}[FF'])^{-2}(zI + \bar{E}[FF'] - 0.5\bar{E}[FF'])\bar{E}[F] - 0.5 \\ &=\; \bar{E}[F]'(zI + \bar{E}[FF'])^{-2}(zI + 0.5\bar{E}[FF'])\bar{E}[F] - 0.5. \end{aligned} \tag{126}$$

We compute the derivative of $a(z)$

$$\begin{aligned} a'(z) \;&=\; \bar{\mu}'_{IS}(z) - 0.5\bar{\Sigma}'_{IS}(z) \\ &=\; -z\bar{E}[F]'(zI + \bar{E}[FF'])^{-2}\bar{E}[F] \\ &+\; z\bar{E}[F]'(zI + \bar{E}[FF'])^{-3}\bar{E}[FF']\bar{E}[F] \\ &=\; z\bar{E}[F]'(zI + \bar{E}[FF'])^{-3}(\bar{E}[FF'] - (zI + \bar{E}[FF']))\bar{E}[F] \\ &=\; -z^2\bar{E}[F]'(zI + \bar{E}[FF'])^{-3}\bar{E}[F]. \end{aligned} \tag{127}$$

Therefore, $a'(z)$ is negative for $z > 0$ and $a'(0) = a''(0) = 0$.

□

Proof of Corollary 4. We have

$$
\begin{aligned}
&\frac{d}{dz}U_{LOO}^{OOS}(f_z) \\
&= \frac{d}{dz}(\bar{\mu}(z) - 0.5\bar{\Sigma}(z)) \\
&= \frac{d}{dz}(\frac{\bar{\mu}_{IS}(z) - \psi(z;c)}{1-\psi(z;c)} - 0.5\frac{\bar{\Sigma}_{IS}(z) + \psi^2(z) - 2\bar{\mu}_{IS}(z)\psi(z;c)}{(1-\psi(z;c))^2}) \\
&= \frac{d}{dz}(\frac{\bar{\mu}_{IS}(z) - \psi(z;c) - \bar{\mu}_{IS}(z)\psi(z;c) + \psi(z;c)^2}{(1-\psi(z;c))^2} \\
&- 0.5\frac{\bar{\Sigma}_{IS}(z) + \psi^2(z) - 2\bar{\mu}_{IS}(z)\psi(z;c)}{(1-\psi(z;c))^2}) \\
&= \frac{d}{dz}(\frac{\bar{\mu}_{IS}(z) - \psi(z;c)}{(1-\psi(z;c))^2} - 0.5\frac{\bar{\Sigma}_{IS}(z) - \psi^2(z)}{(1-\psi(z;c))^2}) \\
&= \frac{d}{dz}(\frac{\bar{\mu}_{IS}(z) - 0.5\bar{\Sigma}_{IS}(z) - \psi(z;c) + 0.5\psi^2(z)}{(1-\psi(z;c))^2}) \\
&= \frac{d}{dz}(\frac{\bar{\mu}_{IS}(z) - 0.5\bar{\Sigma}_{IS}(z) + 0.5(1-\psi(z;c))^2 - 0.5}{(1-\psi(z;c))^2}) \\
&= \frac{d}{dz}(\frac{\bar{\mu}_{IS}(z) - 0.5\bar{\Sigma}_{IS}(z) - 0.5}{(1-\psi(z;c))^2}) \\
&= \frac{d}{dz}(\frac{a(z)}{b(z)}),
\end{aligned}
\tag{128}
$$

where we have defined $b(z) = (1-\psi(z;c))^2$. Note that from Lemma 7, $a(z)$ is decreasing in z and $a(z) < 0$. Therefore the OOS quadratic utility objective can be written as

$$
(a(z)/b(z))' = \frac{a'(z)b(z) - a(z)b(z)'}{b(z)^2} \tag{129}
$$

If we show that the derivative is positive in $z = 0$, then that means there is a value in shrinkage. It suffices to show $\frac{b'(0)}{b(0)} > \frac{a'(0)}{a(0)}$ and $b'(0) > 0$.

$$
b'(0) = -2\psi'(z)(1-\psi(z;c)) \tag{130}
$$

Lemma 8 ensures that $\psi(z;c) < 1$. On the other hand, from the definition of $\psi(z;c) = \frac{1}{T}F_\tau'(zI + \bar{E}[FF'])^{-1}F_\tau$, increasing z leads to a decrease $\psi(z;c)$. This concludes the Proof of

Corollary 4. Finally,

$$\begin{aligned} U_{LOO}^{OOS}(f_z) - U_{LOO}^{OOS}(f_0) &= \int_0^z \frac{d}{dx}(\frac{a(x)}{b(x)})dx \\ &= \int_0^z \frac{a'(x)b(x) - a(x)b(x)'}{b(x)^2}dx \\ &= \int_0^z (-a(x))\frac{b(x)' - b(x)(a'(x)/a(x))}{b(x)^2}dx \\ &\leq \int_0^z (-a(x))\frac{b(x)'}{b(x)^2}dx \\ &\leq \sup_x(-a(x))\left(b(0)^{-1} - b(z)^{-1}\right) \\ &\leq \sup_x(-a(x))\left(b(0)^{-1} - 1\right) \\ &= 0.5\,\frac{(-2\psi(0;c) + \psi(0,c)^2)}{(1-\psi(0,c))^2} \end{aligned} \tag{131}$$

because $\sup_x(-a(x)) = (-a(\infty)) = 0.5$. □

A.1 Auxiliary Lemmas

Lemma 8 (Sherman-Morrison Formula) *Suppose $A \in \mathbb{R}^{n\times n}$ is an invertible square matrix and $u,\ v \in \mathbb{R}^n$ are column vectors. Then $A + \tilde{u}v'$ is invertible if $1 + v'A^{-1}u \neq 0$. In this case,*

$$(A + uv')^{-1} = A^{-1} - \frac{A^{-1}uv'A^{-1}}{1 + v'A^{-1}u} \tag{132}$$

Lemma 9 (PCs and Shrinkage) *Assume $\bar{E}[FF'] = UDU'$ and*

$$\bar{\pi}(z) \ = \ (\bar{E}[FF'])^{-1}\bar{E}[F]. \tag{133}$$

be the Ridge-shrunk Markowitz portfolio. Define the PC portfolios, $F^{PC} = U'F$, then the

corresponding Ridge shrunk Markowitz portfolio is

$$\begin{aligned} \bar{\pi}^{PC}(z) &= (\bar{E}[F^{PC}F^{PC\prime}])^{-1}\bar{E}[F^{PC}] \\ &= (U'\bar{E}[FF']U)^{-1}U'\bar{E}[F] \\ &= (D)^{-1}U'\bar{E}[F] \end{aligned} \tag{134}$$

The OOS performances of these two portfolios are equal

$$\begin{aligned} F'_{OOS}\bar{\pi}(z) &= F^{PC\prime}_{OOS}\bar{\pi}^{PC}(z) \\ &= F'_{OOS}U(D)^{-1}U'\bar{E}[F] \\ &= F'_{OOS}(\bar{E}[FF'])^{-1}\bar{E}[F] \end{aligned} \tag{135}$$

Lemma 10 (Trace Bound for CUPSA) *The sum of shrunk eigenvalues derived from the CUPSA method (as defined in Equation 3) is constrained to be no greater than the total sum of the eigenvalues of the empirical covariance matrix.*

$$\sum_{i=1}^{N} f^{-1}_{Z,W}(\lambda_i) \leq \sum_{j=1}^{N} \lambda_j \tag{136}$$

Proof of Lemma. We know

$$f_{Z,W}(\lambda) = \sum_{i=1}^{L} w_i \frac{c_{z_i}}{z_i + \lambda}, \tag{137}$$

where

$$c_{z_i} = \frac{N z_i + \sum_{j=1}^{N} \lambda_j}{\sum_{j=1}^{N} \lambda_j}. \tag{138}$$

By the Cauchy-Schwartz inequality

$$(\sum_{i=1}^{L} w_i \frac{c_{z_i}}{z_i + \lambda_j})(\sum_{i=1}^{L} w_i \frac{z_i + \lambda_j}{c_{z_i}}) \geq (\sum_{i=1}^{L} w_i)^2 = 1 \tag{139}$$

Then

$$\begin{aligned}\sum_{i=1}^{N} f_{Z,W}^{-1}(\lambda_i) &= \sum_{j=1}^{N} \frac{1}{\sum_{i=1}^{L} w_i \frac{c_{z_i}}{z_i+\lambda_j}} \\ &\leq \sum_{j=1}^{N}\sum_{i=1}^{L} w_i \frac{z_i+\lambda_j}{c_{z_i}} \\ &= \sum_{i=1}^{L} w_i \frac{N z_i + \sum_{j=1}^{N}\lambda_j}{c_{z_i}} \\ &= \sum_{j=1}^{N} \lambda_j.\end{aligned} \tag{140}$$

□

B Additional Results

B.1 Other Rolling Windows

B.1.1 Monthly Returns Results

In this section, we show performance across other rolling windows T for the (Jensen et al., 2023) value-weighted capped monthly returns. Results are fairly similar across rolling windows: There is monotonicity in the Sharpe ratio up to a point and then extreme decline for all shrinkage methods except CUPSA.

T=60

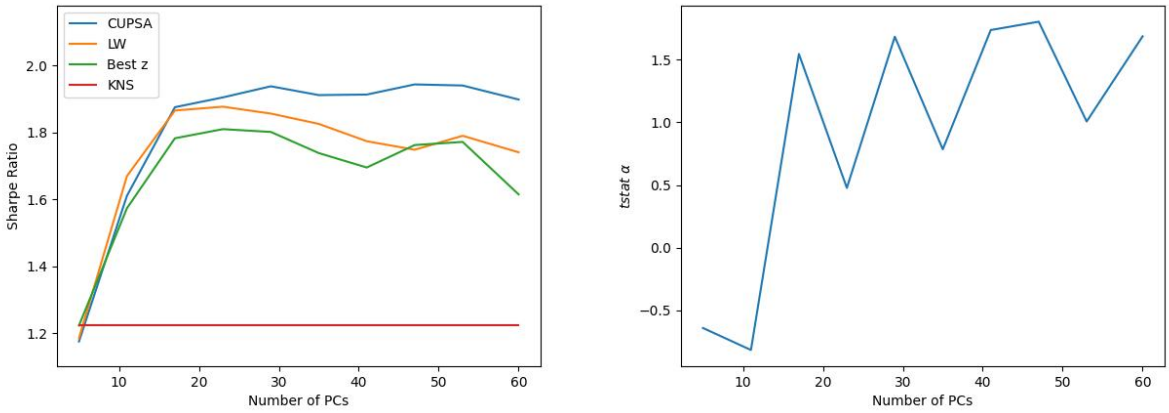

Figure 8: The left plot shows the annualized Sharpe ratio for CUPSA (Theorem 1), LW (Ledoit and Wolf, 2020), KNS (56), and "Best z" (z_* of (55)), as the number of PCs grows. This is done by using (59). The right plot is the Heteroskedasticity-adjusted (with five lags) t-statistics of α of PC portfolios as the number of PCs grows. The regression is done using (57) with the corresponding PC portfolios. Results are from the period 1976-12-31 to 2022-12-31. Portfolios are estimated with a rolling window of $T = 60$ months and re-balanced monthly.

T=360

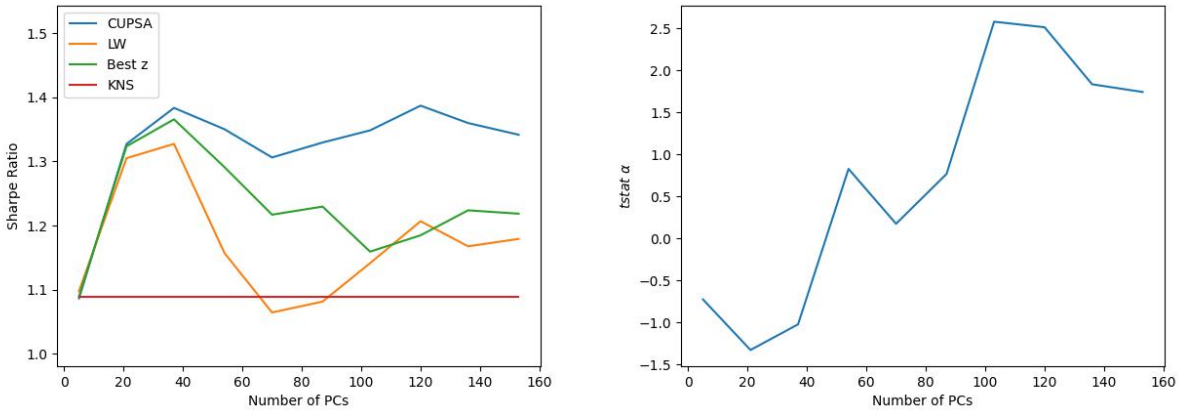

Figure 9: The left plot shows the annualized Sharpe ratio for CUPSA (Theorem 1), LW (Ledoit and Wolf, 2020), KNS (56), and "Best z" (z_* of (55)), as the number of PCs grows. This is done by using (59). The right plot is the Heteroskedasticity-adjusted (with five lags) t-statistics of α of PC portfolios as the number of PCs grows. The regression is done using (57) with the corresponding PC portfolios. Results are from the period 2001-12-31 to 2022-12-31. Portfolios are estimated with a rolling window of $T = 360$ months and re-balanced monthly.

B.1.2 Weekly Returns Results

We repeat the PC exercise for (Jensen et al., 2023) value-weighted capped weekly returns. Weekly returns are calculated by compounding the daily return series [36]. Weekly results[37] for CUPSA are quite different from monthly. The Sharpe ratio from CUPSA shrinkage keeps increasing even with the inclusion of small PCs. We believe this is evidence of CUPSA's ability to capture weak factors once they appear in the data. In other words, if the *virtue of complexity* (Didisheim et al., 2023) exists in the cross-section of returns, CUPSA is more capable of capturing it compared to other shrinkage methodologies.

T=50

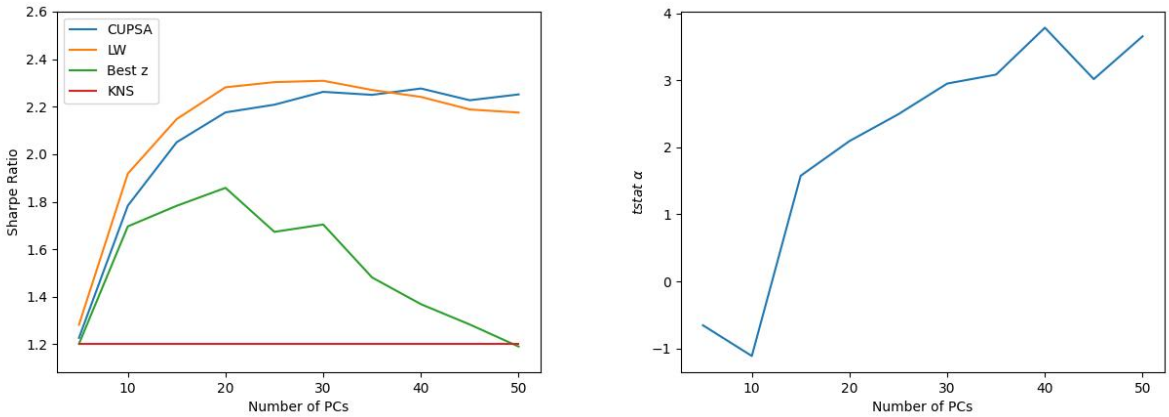

Figure 10: The left plot shows the annualized Sharpe ratio for CUPSA (Theorem 1), LW (Ledoit and Wolf, 2020), KNS (56), and "Best z" (z_* of (55)), as the number of PCs grows. This is done by using (59). The right plot is the Heteroskedasticity-adjusted (with five lags) t-statistics of α of PC portfolios as the number of PCs grows. The regression is done using (57) with the corresponding PC portfolios. Results are from the period 1972-10-31 to 2022-12-31. Portfolios are estimated with a rolling window of $T = 50$ weeks and re-balanced weekly.

[36] These factors are also available on the https://jkpfactors.com.

[37] For the calculation of t-statistics and Sharpe ratios, weekly returns are aggregated into monthly returns.

T=250

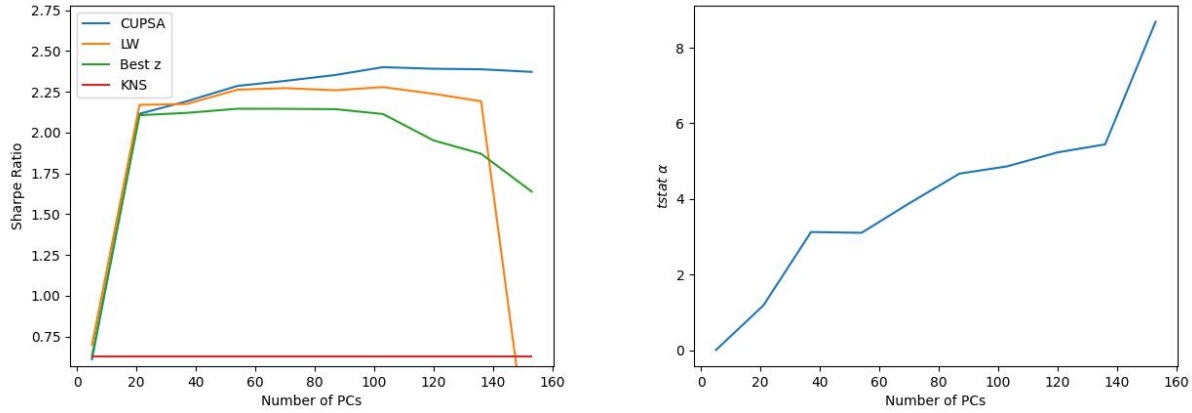

Figure 11: The left plot shows the annualized Sharpe ratio for CUPSA (Theorem 1), LW (Ledoit and Wolf, 2020), KNS (56), and "Best z" (z_* of (55)), as the number of PCs grows. This is done by using (59). The right plot is the Heteroskedasticity-adjusted (with five lags) t-statistics of α of PC portfolios as the number of PCs grows. The regression is done using (57) with the corresponding PC portfolios. Results are from the period 1976-08-31 to 2022-12-31. Portfolios are estimated with a rolling window of $T = 250$ weeks and re-balanced weekly.

T=500

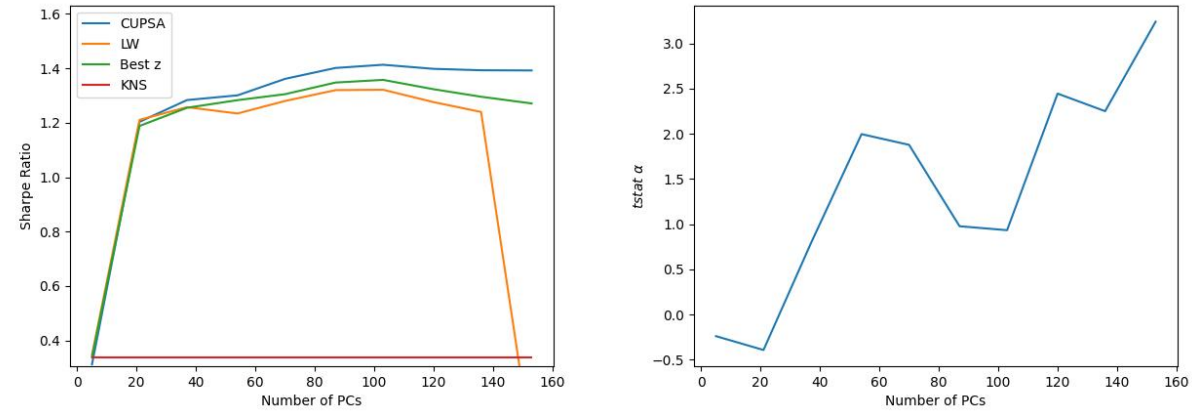

Figure 12: The left plot shows the annualized Sharpe ratio for CUPSA (Theorem 1), LW (Ledoit and Wolf, 2020), KNS (56), and "Best z" (z_* of (55)), as the number of PCs grows. This is done by using (59). The right plot is the Heteroskedasticity-adjusted (with five lags) t-statistics of α of PC portfolios as the number of PCs grows. The regression is done using (57) with the corresponding PC portfolios. Results are from the period 1981-06-30 to 2022-12-31. Portfolios are estimated with a rolling window of $T = 500$ weeks and re-balanced weekly.

B.2 Results for Rank-Managed Portfolios

B.2.1 The Data

We build monthly frequency factors from publicly traded stocks in the US in the style of (Kozak et al., 2020). We use the dataset from (Jensen et al., 2023), which contains a comprehensive set of stock-level predictors previously discussed in the Finance literature. The dataset is populated by 153 characteristics from each stock from 1963 to 2023. We use NYSE/AMEX/NASDAQ securities with a CRSP share code 10, 11, or 12. We exclude nano and micro stocks as classified by (Jensen et al., 2023) (stocks with market capitalization below the 20th percentile). We remove the characteristics with the lowest coverage (this is especially important for the early parts of the sample). We do this to have homogeneous portfolios over time and avoid losing many stock-month observations. We kept the 131 characteristics with the least number of missing observations. Furthermore, we drop stock-month observations where more than 30% is missing. We then follow the exact steps in Kozak et al. (2020) and build rank-managed factors. In this dataset, we see much more monotonicity in the PC space, especially for CUPSA.

B.2.2 Small+Large+Mega Results

Our results are even better with the Kozak et al. (2020) factor construction. The CUPSA Sharpe ratios are much more monotonic in the PC space, and the t-statistics of α are larger. In the case of the $T = 360$ months (30 years) rolling window, there is clear *virtue of complexity*: Even the smallest PCs are improving performance.

T=60

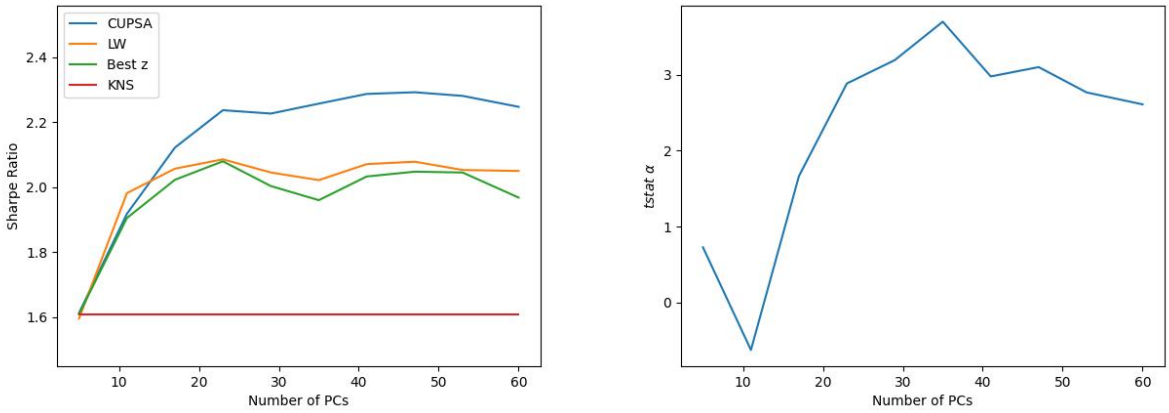

Figure 13: The left plot shows the annualized Sharpe ratio for CUPSA (Theorem 1), LW (Ledoit and Wolf, 2020), KNS (56), and "Best z" (z_* of (55)), as the number of PCs grows. This is done by using (59). The right plot is the Heteroskedasticity-adjusted (with five lags) t-statistics of α of PC portfolios as the number of PCs grows. The regression is done using (57) with the corresponding PC portfolios. Results are from the period 1968-03-31 to 2022-12-31. Portfolios are estimated to have a rolling window of $T = 60$ months and are re-balanced monthly.

T=120

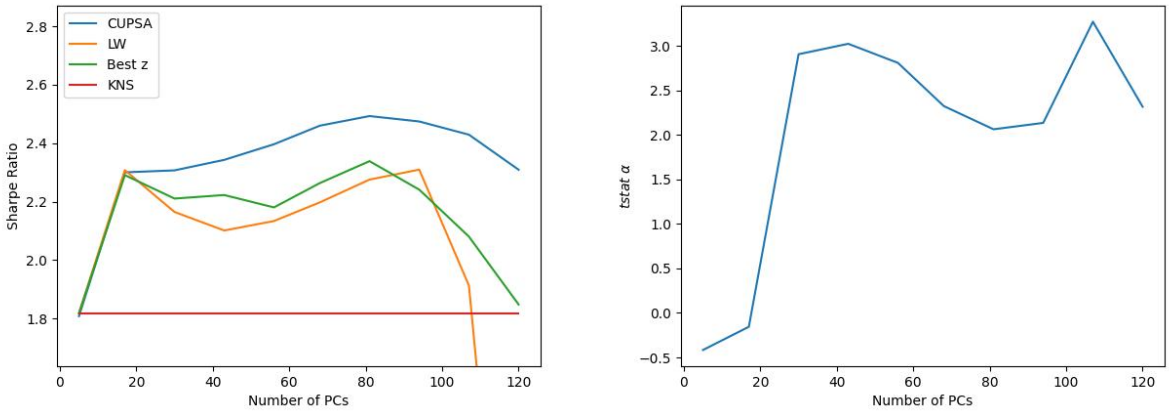

Figure 14: The left plot shows the annualized Sharpe ratio for CUPSA (Theorem 1), LW (Ledoit and Wolf, 2020), KNS (56), and "Best z" (z_* of (55)), as the number of PCs grows. This is done by using (59). The right plot is the Heteroskedasticity-adjusted (with five lags) t-statistics of α of PC portfolios as the number of PCs grows. The regression is done using (57) with the corresponding PC portfolios. Results are from the period 1973-03-31 to 2022-12-31. Portfolios are estimated to have a rolling window of $T = 120$ months and are re-balanced monthly.

T=360

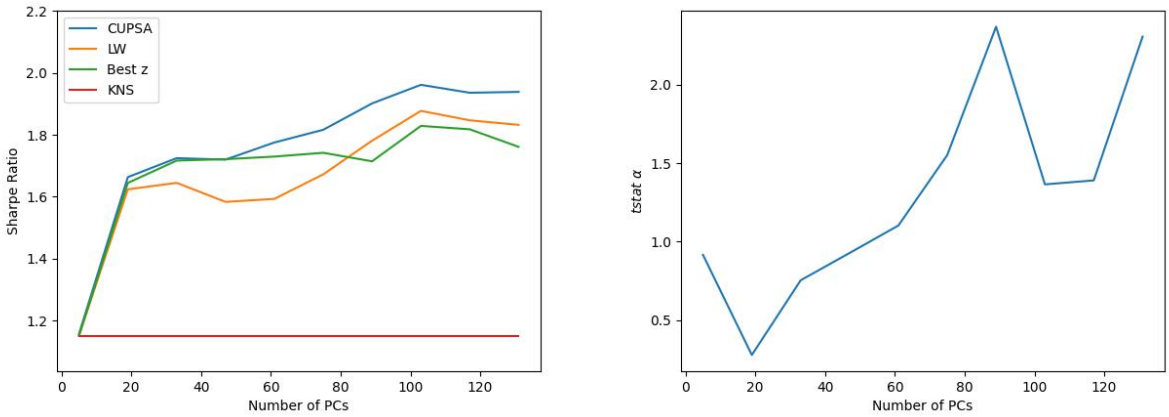

Figure 15: The left plot shows the annualized Sharpe ratio for CUPSA (Theorem 1), LW (Ledoit and Wolf, 2020), KNS (56), and "Best z" (z_* of (55)), as the number of PCs grows. This is done by using (59). The right plot is the Heteroskedasticity-adjusted (with five lags) t-statistics of α of PC portfolios as the number of PCs grows. The regression is done using (57) with the corresponding PC portfolios. Results are from the period 1993-03-31 to 2022-12-31. Portfolios are estimated to have a rolling window of $T = 360$ months and are re-balanced monthly.

B.2.3 Robustness Check: Mega + Large Results

In this section, We build rank-managed portfolios using the most liquid stocks. I first remove small stocks[38], then follow the same steps as Kozak et al. (2020) to build a new set of factors. CUPSA is the superior shrinkage methodology even when limiting portfolios to the most liquid stocks. *Virtue of complexity* is also clearly present for the $T = 360$ months rolling window.

T=60

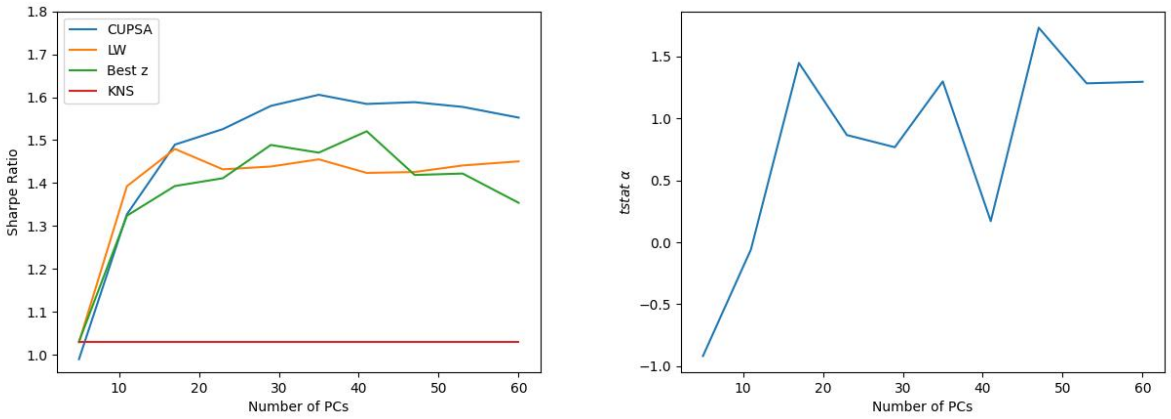

Figure 16: The left plot shows the annualized Sharpe ratio for CUPSA (Theorem 1), LW (Ledoit and Wolf, 2020), KNS (56), and "Best z" (z_* of (55)), as the number of PCs grows. This is done by using (59). The right plot is the Heteroskedasticity-adjusted (with five lags) t-statistics of α of PC portfolios as the number of PCs grows. The regression is done using (57) with the corresponding PC portfolios. Results are from the period 1968-03-31 to 2022-12-31. Portfolios are estimated to have a rolling window of $T = 60$ months and are re-balanced monthly.

[38]Stocks with market capitalization between the 50th and 20th percentiles, as defined in Jensen et al. (2023).

T=120

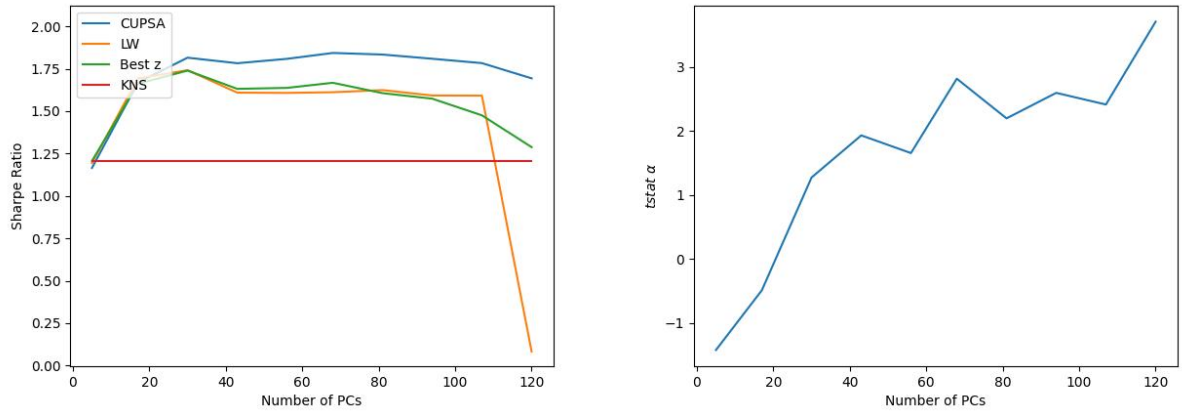

Figure 17: The left plot shows the annualized Sharpe ratio for CUPSA (Theorem 1), LW (Ledoit and Wolf, 2020), KNS (56), and "Best z" (z_* of (55)), as the number of PCs grows. This is done by using (59). The right plot is the Heteroskedasticity-adjusted (with five lags) t-statistics of α of PC portfolios as the number of PCs grows. The regression is done using (57) with the corresponding PC portfolios. Results are from the period 1973-03-31 to 2022-12-31. Portfolios are estimated to have a rolling window of $T = 120$ months and are re-balanced monthly.

T=360

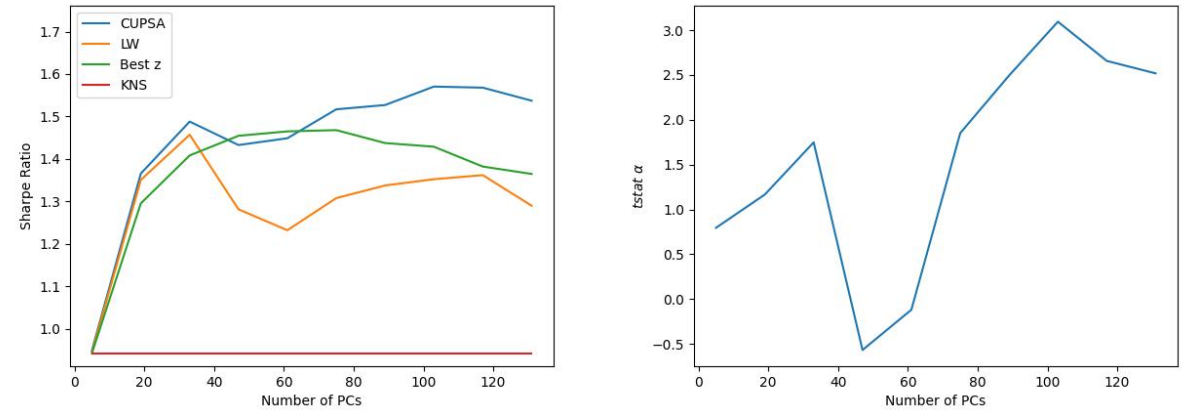

Figure 18: The left plot shows the annualized Sharpe ratio for CUPSA (Theorem 1), LW (Ledoit and Wolf, 2020), KNS (56), and "Best z" (z_* of (55)), as the number of PCs grows. This is done by using (59). The right plot is the Heteroskedasticity-adjusted (with five lags) t-statistics of α of PC portfolios as the number of PCs grows. The regression is done using (57) with the corresponding PC portfolios. Results are from the period 1993-03-31 to 2022-12-31. Portfolios are estimated with a rolling window of $T = 360$ months and re-balanced monthly.

B.3 Correlation Across Shrinkage Values

Drawing on the insights from Theorem 1, we know that $CUPSA$ generates performance benefits by efficiently combining various "simple Ridge" portfolios. Thus, the superior performance of CUPSA suggests significant diversification gains (i.e., low correlations) across Ridge portfolios. Figure 19 confirms this intuition: average correlations between low-z (i.e., Markowitz) and high-z (i.e., factor momentum (15)) are indeed low.

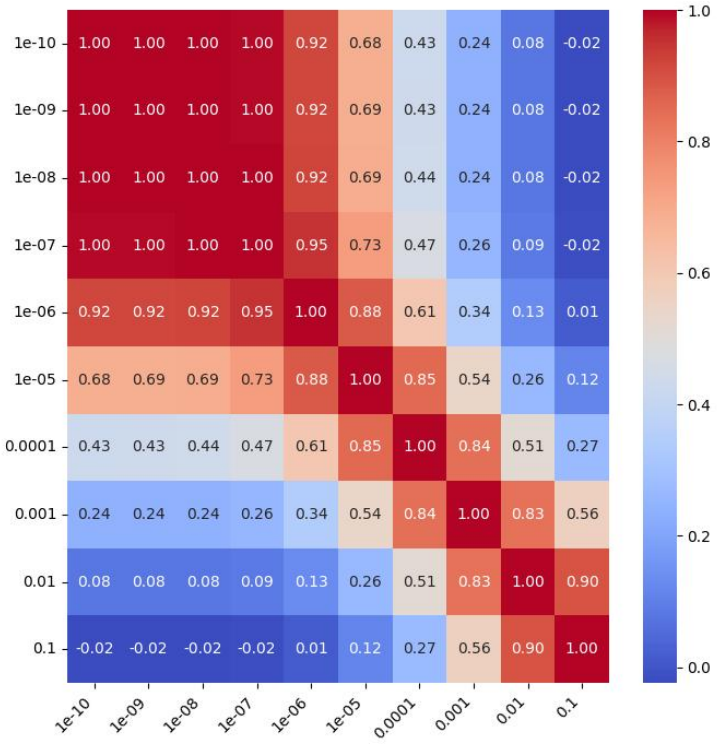

Figure 19: Correlation of OOS returns of efficient portfolios with different levels of Ridge shrinkage, $Corr(R_{t+1}(f_Z), R_{t+1}(f_Z))$. Correlations are computed over the period 1981-12-31 to 2022-12-31. Portfolios are estimated to have a rolling window of $T = 120$ months and are re-balanced monthly.

C Bayesian Interpretation

Let $F_1, \ldots, F_T$ be an IID Gaussian random sequence with unknown expectation μ and known positive definite covariance matrix Σ. The unknown Markowitz portfolio is given by:

$$b := \Sigma^{-1}\mu \tag{141}$$

Uncertainty about expected returns is described by a prior π such that:

$$\mu \sim_\pi N(0, g(\Sigma)) \tag{142}$$

for a strictly positive matrix function g. The resulting prior on the Markovitz portfolio is:

$$b \sim_\pi N(0, \Sigma^{-1} g(\Sigma) \Sigma^{-1}) \tag{143}$$

This specification also has a direct interpretation in terms of principal component factors. Indeed, given spectral decomposition $\Sigma = QDQ'$ and principal component returns $\{P_t := Q'F_t : t = 1, \ldots, T\}$, following prior on expected principal component returns holds:

$$\mu_P \sim_\pi N(0, g(D)) \tag{144}$$

The resulting prior on principal component Sharpe ratios are:

$$SR_P := D^{-1/2}\mu_P \sim_\pi N(0, g(D)D^{-1}) \tag{145}$$

This implies following a priori expected maximal squared Sharpe ratio:

$$\mathbb{E}_\pi[||SR_P||_2^2] = tr(D^{-1}\mathbb{E}_\pi[\mu_P \mu_P']) = tr(D^{-1}g(D)) = \sum_{i=1}^{N} \frac{g(d_i)}{d_i} \tag{146}$$

Similarly, the a priori expected squared norm of the Markowitz portfolio is:

$$E_\pi[||b||_2^2] = tr(\Sigma^{-2}\mathbb{E}_\pi[\mu\mu'])) = tr(\Sigma^{-2}g(\Sigma)) = tr(D^{-2}g(D)) = \sum_{i=1}^{N} \frac{g(d_i)}{d_i^2} \tag{147}$$

As a consequence, g implicitly models the contribution of individual covariance matrix eigenvalues to the expected economy's maximal squared Sharpe ratio and Markowitz portfolio norm.

Given hyper-parameters $\mu_0 = 0$ and $\Sigma_0 = g(\Sigma)$ in prior (144) for expected returns, posterior expected returns are given by:

$$\begin{aligned}\mathbb{E}[\mu|F_1,\ldots,F_T] &= (\Sigma_0^{-1} + T\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + T\Sigma^{-1}\bar{E}[F]) \\ &= (\frac{1}{T}\Sigma\Sigma_0^{-1} + I)^{-1}\bar{E}[F] \ .\end{aligned}$$

Therefore, the posterior expected Markowitz weight vector is:

$$\mathbb{E}[b|F_1,\ldots,F_T] = (\frac{1}{T}\Sigma[g(\Sigma)]^{-1}\Sigma + \Sigma)^{-1}\bar{E}[F] \ . \tag{148}$$

This last equation suggests following factorization of g, using a strictly positive matrix function h_1:[39]

$$g(\Sigma) = \frac{1}{N}\Sigma h_1(\Sigma)\Sigma \ , \tag{150}$$

which yields:

$$\mathbb{E}[b|F_1,\ldots,F_T] = \left(\frac{N}{T}[h_1(\Sigma)]^{-1} + \Sigma\right)^{-1}\bar{E}[F] =: f_1(\Sigma)\bar{E}[F] \ . \tag{151}$$

Here, parameter N/T is a complexity multiplier scaling the contribution of $[h_1(\Sigma)]^{-1}$ to the total shrinkage $f_1(\Sigma)$. This contribution vanishes asymptotically in the standard regime where $N/T \to 0$, but it does not in the complex regime where $N/T \to c > 0$. This last feature reflects the not-vanishing prior uncertainty about the Markowitz portfolio in the complex regime. Whenever $h_1(\Sigma)$ is monotonically decreasing, then

$$f_1(\Sigma) = \left(\frac{N}{T}[h_1(\Sigma)]^{-1} + \Sigma\right)^{-1} \tag{152}$$

[39] An alternative natural factorization with scaling factor $tr(\Sigma)$ is:

$$g(\Sigma) = \frac{1}{tr(\Sigma)}\Sigma h_1(\Sigma)\Sigma \ , \tag{149}$$

is also monotonically decreasing and can be directly approximated by CUPSA using the Ridge ensemble. If h_1 is monotonically increasing and such that $h(d) \to \infty$ as $d \to \infty$, the monotonicity of f_1 is ambiguous, but in such a case, we can still approximate $[h_1(\Sigma)]^{-1}$ directly using CUPSA. Note that the monotonicity properties of h_1 are equivalent to assuming in equation (147) an associated monotonicity for the contribution of different eigenvalues to the a priori expected squared norm of the Markovitz portfolio:

$$E_\pi[b'b] = \sum_{i=1}^{N} h_1(d_i) \tag{153}$$

For instance, $g(\Sigma) = \frac{1}{z}\Sigma^\eta$ for some tuning parameters $z > 0$ and $\eta \geq 0$ yields $h_1(\Sigma) = \frac{1}{z}\Sigma^{\eta-2}$, which is strictly decreasing if and only if $\eta < 2$. In the limit case $\eta = 2$, the contribution to the a priori expected squared norm of the Markovitz portfolio is identical across eigenvalues. In this case, the resulting portfolio shrinkage produces a standard Ridge-type shrinkage:

$$f_1(\Sigma) = \left(\frac{N}{T} z I_N + \Sigma\right)^{-1} . \tag{154}$$

A natural alternative factorization to (150) is:

$$g(\Sigma) = \frac{1}{N}\Sigma^{1/2} h_2(\Sigma) \Sigma^{1/2} , \tag{155}$$

for some strictly positive matrix function h_2. This yields:

$$\mathbb{E}[b|F_1, \ldots, F_T] = \left(\frac{N}{T}[h_2(\Sigma)]^{-1}\Sigma + \Sigma\right)^{-1} \bar{E}[F] =: f_2(\Sigma)\bar{E}[F] . \tag{156}$$

Whenever $h_2(\Sigma)$ is monotonically decreasing, then

$$f_2(\Sigma) = \left(\frac{N}{T}[h_2(\Sigma)]^{-1}\Sigma + \Sigma\right)^{-1} \tag{157}$$

is again monotonically decreasing and can be approximated by CUPSA using the Ridge ensemble. Here, the monotonicity properties of h_2 completely pin down in equation (146) those of the contribution across eigenvalues to the a priori expected squared maximal Sharpe

ratio:

$$\mathbb{E}_\pi[||SR_P||_2^2] = \sum_{i=1}^{N} h_2(d_i) \tag{158}$$

For instance, $g(\Sigma) = \frac{1}{z}\Sigma^\eta$ for some tuning parameters $z > 0$ and $\eta \geq 0$ yields $h_2(\Sigma) = \Sigma^{\eta-1}$, which is strictly decreasing if and only if $\eta < 1$. In the limit case $\eta = 1$, the contribution to the a priori expected squared maximal Sharpe ratio is identical across eigenvalues, and the resulting portfolio shrinkage produces a simple scaling of the inverse covariance matrix of returns:

$$f_2(\Sigma) = \frac{1}{1 + \frac{N}{T}z}\Sigma^{-1} \ . \tag{159}$$

Finally, the following remarks are in order for the above Bayesian motivated CUPSA-type shrinkages.

1. Function f_1 is computable also when Σ is not full rank, even though the derivation assumes Σ to be full rank.

2. The above Bayesian foundation suggests naturally interpretable scaling procedures for the final shrinkages f_1, f_2.

3. We can also directly apply CUPSA to h_1, h_2 and study the performance of the associated f_1, f_2. This can give insight on the role of different economically founded prior assumptions about Markowitz portfolio and Sharpe ratio shapes for producing an improved performance out of sample. It may also help to clarify the discussion about whether the optimal shrinkage can be sparse.