Universal Portfolio Shrinkage

Bryan Kelly, Semyon Malamud, Mohammad Pourmohammadi, and Fabio Trojani

December 15, 2023

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 existing shrinkage methods. It exhibits an explicit two-fund separation, complementing the Markowitz portfolio with an optimal *complexity correction.* Importantly, UPSA does not annihilate the low-variance principal components (PCs) of returns but weights them optimally. Contrary to conventional wisdom, we find that low variance in-sample PCs are key to out-of-sample portfolio performance. By optimally balancing them in the portfolio construction, UPSA produces a stochastic discount factor that substantially improves on its PC-sparse counterparts, showing that PC sparsity is highly costly once SDFs are optimally shrunk.

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]

[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.

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 severely contaminated 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 favorable risk-return tradeoff, statistical limits to arbitrage (Da et al., 2022) and limits to learning (Didisheim et al., 2023) 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

[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.

generality, we start from a broad family of *spectral shrinkage* estimators, which transform the sample variances of estimated PCs, or equivalently the eigenvalues of the sample covariance matrix of returns, with some potentially non-linear function f. De facto, these estimators adjust individual PC weights in the Markowitz portfolio with a single transformation of PC variances, thus effectively re-weighting all individual PC estimated Sharpe ratios with a flexible scheme. 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 devices can be efficiently spanned using a tractable Universal Portfolio Shrinkage Approximator (UPSA), which is built just from a basis of Ridge-penalized portfolios depending on a corresponding 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 essentially 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 estimated SDFs in which the order of the risks of each estimated PC is maintained as the one before shrinking. It also means that sample covariance matrices associated with a larger estimated risk for all portfolios will still 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 de facto implicitly assign an infinite risk to some low-variance PCs.

By construction, the CUPSA portfolio is equivalent to an optimal fund portfolio, which allocates wealth across a family of funds, given by distinct ridge-shrunk portfolios, and is subject to a short-selling constraint on each of the funds. Therefore, CUPSA also has the interpretation of a Bayesian optimal portfolio, which extends the standard interpretation

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

of ridge-shrunk portfolios in the literature. While ridge-shrunk portfolios are the optimal Bayesian portfolios of an investor having a Gaussian prior with scalar covariance matrix on expected returns, the CUPSA portfolio is the optimal Bayesian portfolio of an investor having a mixture of Gaussian prior on expected returns. In this interpretation, the CUPSA portfolio weights allocated to individual ridge-shrunk portfolios equal the weights in the mixture of Gaussian 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 capture better and 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 performance. 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). In all experiments we perform, we find that CUPSA achieves a higher OOS Sharpe ratio than the benchmark methods, i.e., the associated SDF leads to lower out-of-sample pricing errors (see, again, (Didisheim et al., 2023)). When stratifying into 13 anomaly themes the OOS pricing performance of the CUPSA-SDF across the 150 anomalies in (Jensen et al., 2023), we further find that the CUPSA-SDF achieves significantly lower pricing errors for every single theme. These effects are strongest for the two "hardest-to-price" anomaly themes: Momentum and low risk. Indeed, while benchmark SDFs have a hard time pricing these, the CUPSA-SDF achieves comparably low pricing errors for momentum and low risk as it does for other themes.

To understand the origins of the improved out-of-sample performance provided by the UPSA portfolio, we make use of random matrix theory and derive an explicit link between the latter, the standard estimator of the maximum Sharpe ratio portfolio, the unknown covariance matrix of returns and the degree of model complexity $c = N/T$, which is given by the ratio between the number N of assets in a portfolio problem and the number T of available time series observations for estimation. High complexity $c > 0$ causes a breakdown of the law of large numbers. This creates a wedge between in-sample and out-of-sample moments that does not disappear asymptotically. Intuitively, a larger model complexity is responsible for a larger wedge between the out-of-sample performance of CUPSA and Markowitz portfolios. We find that such a wedge is particularly large when using monthly data and a number of SDF factors, as in the recent literature. Perhaps surprisingly, we

further find that the complexity wedge remains significant even when c is relatively small, e.g., $c \sim 0.1$, when using $N = 153$ factors and $T = 1500$ daily observations. This finding can be partly explained by the covariance structure of the asset pricing factors, which exhibits many factors with high correlations. This feature further lowers the effective sample size and makes the small eigenvalues hard to estimate even with large training samples.

We study the behavior of CUPSA for various rolling training windows, from very short (a few weeks) to very long (seven years of daily data). In this way, we better account for a potential non-stationarity in the data. We find evidence for this non-stationarity. In particular, the OOS performance of the different benchmark portfolios is non-monotonic in the rolling window size: Models based on shorter windows suffer more from the lack of observations, but they adapt faster to changing economic conditions, and vice versa. Our empirical results show that CUPSA can better adjust to this non-stationarity than its feasible "optimal ridge" counterpart. It does so by diversifying and smoothing out the choice of the optimal shrinkage parameter across the multiple ridge-shrunk portfolios in the UPSA basis. Strikingly, we find that the average CUPSA weights across ridge penalties closely emulate 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 and generating a huge turnover. In contrast, CUPSA shrinkage better adjusts to non-stationarity by diversifying and smoothing across multiple ridge penalties to construct the optimal portfolio shrinkage. In doing so, it also generates a substantially lower turnover.

We finally explore the shape of the optimal CUPSA shrinkage, particularly whether it gives rise to SDFs with nontrivial exposure to low-variance PCs. Therefore, we compare the behavior of all portfolio shrinkage estimators under scrutiny on the set of estimated PCs. We find that the out-of-sample Sharpe ratio of the CUPSA SDF monotonically increases with the number of PCs it incorporates. Strikingly, this monotonic pattern persists with the inclusion of the lowest-variance PCs. The underlying mechanism is as follows. First, low-variance PCs add diversification benefits. Second, as for the other PCs, they contribute non-trivially to the overall out-of-sample portfolio performance. This second feature is intrinsically related to the fact that CUPSA can better capitalize on the complexity of the PC space by means of a more flexible shrinkage able to form a fund portfolio of high- and low-variance PCs with improved out-of-sample economic value. The importance of a 'flexible portfolio shrinkage" for optimal portfolio and SDF estimation is a subtle and novel phenomenon in connection with complexity. In stark contrast with CUPSA, the out-of-sample Sharpe ratio of the standard ridge portfolio is inverse U-shaped with respect to the number of included top PCs and saturates quite early, at about 20 top PCs. Therefore, the restrictive shape of the family

of portfolio shrinkages admitted by the standard ridge estimator spuriously suggests sparse optimal SDFs that load exclusively on the top estimated PCs. These findings spotlight a novel form of *shrinkage-based virtue of complexity*, consistent with earlier findings in (Kelly et al., 2022; Didisheim et al., 2023). It also indicates that the conventional APT wisdom, advocating SDF–sparsity in the PC space, is misplaced when building optimally shrunk portfolios and SDFs in complex environments. The opposite actually holds: SDF–sparsity is highly costly once SDFs are optimally shrunk.

2 Literature Review

Covariance estimation: Our research contributes to the field of covariance matrix estimation. Within this domain, the predominant focus of existing research has been minimizing the Frobenius norm of the difference between the true and the estimated covariance matrices.[4] Pioneering work in this area includes simple, linear shrinkage estimators introduced in (Ledoit and Wolf, 2004b) and their applications to minimum variance portfolios (Ledoit and Wolf, 2003) and portfolio tracking (Ledoit and Wolf, 2004a).

The idea of spectral shrinkage of covariance matrices was introduced in the influential lecture notes by (Stein, 1986). In a sequence of path-breaking papers, (Ledoit and Wolf, 2012), (Ledoit and Wolf, 2015), and (Ledoit and Wolf, 2020) developed novel, non-linear spectral shrinkage estimators based on random matrix theory techniques introduced in (Ledoit and Péché, 2011). These estimators can be computed analytically and are asymptotically optimal with respect to the Frobenius norm in the limit as $N, T \to \infty,\ N/T \to c$. In (Ledoit and Wolf, 2017), the authors apply the same methodology to compute the asymptotically optimal non-linear shrinkage, minimizing the variance of the portfolio return.

To our knowledge, our paper is the first to directly compute the optimal, non-linear spectral shrinkage for the generic (Markowitz, 1952) problem. In stark contrast to the above papers, our shrinkage operator directly optimizes the out-of-sample utility and is explicitly designed for portfolio optimization and asset pricing.[5] While our estimator is a spectral shrinkage estimator (it only transforms the eigenvalues of the sample covariance matrix), the shrinkage itself depends explicitly and non-linearly on the expected asset returns. This form of shrinkage (shrinking Σ based on μ) is novel and has never been studied in the literature before.

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

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

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 representation of the SDF.[6] 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 150 factors from (Jensen et al., 2023)), we find that low-variance estimated PCs are important contributors to SDF performance. In fact, the out-of-sample Sharpe ratio strictly 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 captures heterogeneous beliefs of investors with varying degrees of prior uncertainty. Our findings suggest that the intuition of (Kozak et al., 2020) is misled by a sub-optimal shrinkage procedure: When using a simple ridge for shrink, optimal SDF is indeed PC-sparse. However, the optimal non-linear shrinkage allows us to a form of "goldilocks" shrinkage, whereby each PC is shrunk "just right," according to its estimated risk-return tradeoff.

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 [7]. To deal with the weak factor problem, (Lettau and Pelger, 2020) develop a novel covariance shrinkage methodology they name 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 the PCs towards the vector of their sample means. In particular, their estimator does not belong to the spectral family. The RP-PCA's goal is to correct the bias (induced by complexity $c = N/T > 0$) in the estimation of 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.

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

[7] see, (Bryzgalova et al., 2023a), (Preite et al., 2022)

These low-variance PCs matter precisely because they capture the exposures to the *weak factors*. Shrinkage corrects some of the bias in estimating these weak factor risk premia and generates the virtue of complexity, whereby the inclusion of these (properly re-weighted) weak factors boosts OOS performance.

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) emphasize the importance of employing both shrinkage techniques and out-of-sample (OOS) testing in Bayesian high-dimensional models. In a similar vein, (Kelly et al., 2022) and (Didisheim et al., 2023) highlight both theoretically and empirically the advantages of complex models in asset pricing for achieving superior out-of-sample performance. This holds true 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[8]

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

of portfolio returns

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

[8]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 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)

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

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 that completely ignores the covariance matrix and

invests proportionally to in-sample mean returns:

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

By varying z between 0 and ∞, we effectively interpolate between the no-shrinkage and full shrinkage regimes.

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

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

be the *f-spectral shrinkage estimator of the infeasible efficient portfolio* $E[FF']^{-1}E[F]$. Then, we can rewrite portfolio returns $R_f(f) \;=\; R_t^{\bar{\pi}(f)}$ of this portfolio 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 eigenvalues of the empirical second-moment matrix, the shrinkage operator, f, has, in fact, $NT+1$ arguments.[9] With i.i.d. data, in the low complexity regime when $N/T \to 0$, Lemma 6 implies that shrinkage is sub-optimal and, hence, $f(\lambda; F_{IS}) = \lambda$ is independent of the in-sample data. However, as we show below, even when N/T is only slightly different from zero (e.g., when $N/T \sim 0.1$), the benefits of shrinkage are significant.

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 moment 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.

[9] (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).

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:

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

where, $t > T$. The formula (24) motivates the following feasible version of the infeasible problem (19).[10]

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

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

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 start our analysis in this section by investigating the behavior of the simple ridge shrinkage operator, corresponding to $f_z(\lambda) \;=\; \frac{1}{z+\lambda}$. The following formula plays an important role in our analysis.

[10] An important theoretical question is whether the estimator $U^{OOS}_{LOO}(f)$ is consistent: Is it true that, as $T \to \infty$, $U^{OOS}_{LOO}(f) \to E[U(R_T(f))]$ in probability. Such a result would imply that maximizing $U^{OOS}_{LOO}(f)$ 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.

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

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

be the (21) estimator for $f_z(\lambda) = \frac{1}{z+\lambda}$ and $R_{T,\tau}(f_z)$ the corresponding portfolio return (22).

Lemma 2 (LOO ridge Performance) *Then 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,

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

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. 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.[11] One can derive the following bound for ψ_τ. This bound links the magnitude of overfitting to model complexity. Indeed, Lemma 3 shows that for low values of complexity, there is no overfit, and one can use IS performance.

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

[11] 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 stay significant, proportional to 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 expected return than the OOS $R_{T,\tau}(f_z)$ because of the overfit term in (27). Second, the in-sample return underestimates 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.

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.*

$$f_{Z,W}(\lambda) \;=\; \sum_{i=1}^{L} \underbrace{(z_i + \lambda)^{-1}}_{ridge}\, \underbrace{w_i}_{weight} \tag{32}$$

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 $\mathcal{S}_+^L = \{W \in \mathbb{R}_+^L,\; \sum_{i=1}^N w_i = 1\}$ is the L-dimensional simplex.

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*

$$\begin{aligned}\bar{\mu}(Z) &= \left(\frac{1}{T}\sum_{t=1}^{T} R_{T,t}(f_{z_i})\right)_{i=1}^{L} \in \mathbb{R}^L \\ \bar{\Sigma}(Z) &= \left(\frac{1}{T}\sum_{t=1}^{T} R_{T,t}(f_{z_i})R_{T,t}(f_{z_j})\right)_{i,j=1}^{L} \in \mathbb{R}^{L\times L}\end{aligned} \tag{33}$$

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

$$R_{T,\tau}(f_{Z,W}) = \sum_{i=1}^{L} w_i R_{T,\tau}(f_z). \tag{34}$$

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

$$U_{LOO}^{OOS}(f_{Z,W}) = W'\bar{\mu}(Z) - 0.5W'\,\bar{\Sigma}(Z)\,W\,. \tag{35}$$

Lemma 4 shows how the OOS utility can be computed explicitly in terms of the estimated OOS moments (33). As a result, Lemma 4 implies that finding the optimal spectral shrinkage inside the ridge ensemble amounts to solving the OOS (based on Leave-One-Out) 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):

$$\begin{aligned}f_{UPSA} &= \arg\max_{f\in\mathcal{F}(Z)} U_{LOO}^{OOS}(f) \\ f_{CUPSA} &= \arg\max_{f\in\mathcal{F}_C(Z)} U_{LOO}^{OOS}(f)\end{aligned} \tag{36}$$

The key implication of the above discussion is that the non-parametric *optimization over functions* in (36) is equivalent to a closed form, explicit *optimization over weight vectors* W.

Theorem 1 (UPSA and CUPSA) *We have*

$$\begin{aligned} f_{UPSA}(\lambda) &= f_{Z,W_{UPSA}}(\lambda), \\ f_{CUPSA}(\lambda) &= f_{Z,W_{CUPSA}}(\lambda) \end{aligned} \tag{37}$$

with

$$\begin{aligned} W_{UPSA} &= \bar{\Sigma}(Z)^{-1}\bar{\mu}(Z) \\ W_{CUPSA} &= \arg\max_{W\in\mathcal{S}_+^L}(W'\bar{\mu}(Z) - 0.5W'\bar{\Sigma}(Z)W). \end{aligned} \tag{38}$$

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, any matrix monotone-decreasing function[12] *$f(\lambda)$ satisfying the normalization $\lim_{\lambda\to\infty} f(\lambda)\lambda = 1$ can be uniformly approximated by a function $f \in \mathcal{F}_C(Z)$ 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*

$$\begin{aligned} f^*(\lambda) &= \arg\max\{U_{LOO}^{OOS}(f): \ f \text{ is continuous}\} \\ f_C^*(\lambda) &= \arg\max\{U_{LOO}^{OOS}(f): \ f \text{ is matrix monotone and } \lim_{\lambda\to\infty} f(\lambda)\lambda = 1\} \end{aligned} \tag{39}$$

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

$$\begin{aligned} U_{LOO}^{OOS}(f_{UPSA}) &\geq U_{LOO}^{OOS}(f^*) - \varepsilon \\ U_{LOO}^{OOS}(f_{CUPSA}) &\geq U_{LOO}^{OOS}(f_C^*) - \varepsilon \end{aligned} \tag{40}$$

[12] 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.

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 $\sum_i w_i = 1$ ensures that CUPSA shrinks eigenvalues by building a convex combination of simple ridge shrinkages. Furthermore, the condition $\lim_{\lambda\to\infty} f(\lambda)\lambda = 1$ of Lemma 5 implies that, for very large λ, $f(\lambda)$ behaves like $1/\lambda$. This ensures that very large eigenvalues are not shrunk too much. 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. We will discuss these issues in the next subsection.

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 (38). To understand these benefits, we use the formula (27) and derive the *overfit* for the estimated means and covariances in (33). The following is true. Although the overfit $\psi_t(z)$ typically varies with t, it is possible to apply asymptotic principles from Random Matrix Theory to remove 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*

$$\begin{aligned} m(-z;c) &= \lim_{N,T\to\infty,\ N/T\to c} N^{-1}\operatorname{tr}((\bar{E}[FF'] + zI)^{-1}) \\ \psi(z;c) &= \lim_{N,T\to\infty,\ N/T\to c} \psi_t(z) \end{aligned} \tag{41}$$

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

Furthermore, the asymptotic overfit is given by

$$\psi(z;c) \;=\; c(1 - zm(-z;c))\,. \tag{42}$$

4.1.1 Optimality of Non-Linear Shrinkage

Proposition 3 allows us to drastically simplify the calculations of (33) 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) *we have*

$$\sup_{z>0} U_{LOO}^{OOS}(f_z) \;>\; U_{LOO}^{OOS}(f_0)\,, \tag{43}$$

and the supremum is always achieved for some $z_* > 0$.

Suppose now that, under the hypothesis of Proposition 3, we multiply the vector of risk premia λ by a constant $\alpha > 0$. Then, $z_*(\alpha)$ is monotone decreasing in α. Thus, larger shrinkage is needed when the size of λ is smaller.

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.

The second part of the Lemma emphasizes a subtle link between optimal shrinkage and the size of factor risk premia.[13] When factor risk premia are large, the complexity-driven estimation noise and the inherent overfit have a marginally negative impact on performance. By contrast, when factor risk premia are low, estimated portfolio weights are dominated by noise, making efficient shrinkage vital for OOS portfolio performance.

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 $z_0 = z$ (so that ridgeless, (14), is the first element of the ridge ensemble). Let $\psi(Z) =$*

[13] See, also, Kelly et al. (2022) who show that the optimal shrinkage is inversely proportional to the signal-to-noise ratio in a regression setting.

$(\psi(z_i))_{i=1}^L$ *be the vector of overfits 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* $\alpha,\ \beta > 0$ *:*

$$W_{UPSA} = \alpha\,\delta_{z_0} + \beta\, D(Z)^{-1}\Sigma_{IS}(Z)^{-1}\psi(Z), \tag{44}$$

so that the UPSA efficient portfolio return is given by

$$\bar{\pi}(f_{UPSA}) = \alpha \underbrace{\bar{\pi}(f_0)}_{Markowitz} + \beta \underbrace{\bar{\pi}^{\psi}}_{complexity\ \ correction}, \tag{45}$$

where

$$\bar{\pi}^{\psi} = \sum_{z\in Z} \underbrace{\bar{\pi}(f_z)}_{ridge\ \ portfolio\ \ (12)} (D(Z)^{-1}\Sigma_{IS}(Z)^{-1}\psi(Z))(z)\,. \tag{46}$$

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, these overfits depend 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 of 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 (45) 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 (33). 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 $E[F]$. Here, we extend this observation to the case of the $\mathcal{F}_C(Z)$ ensemble.

Lemma 6 *Consider an economic agent who (irrationally) believes that* $\Sigma = \bar{E}[FF'] - \bar{E}[F]\bar{E}[F]' = E[FF'] - E[F]E[F]'$, *and 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,*

$$E[\mu|F_{IS}] \;=\; \bar{w}^{-1} \sum w_i (z_i I + \Sigma)^{-1} \bar{E}[F]\,. \tag{47}$$

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. Alternatively, CUPSA can also be interpreted as an aggregation of beliefs of market participants with diverse degrees of 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 might have a more dispersed prior. In equilibrium, (47) 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 daily frequency dataset from (Jensen et al., 2023)[14]. This comprehensive dataset contains daily returns for $N = 153$ factors, which are constructed from publicly traded stocks in the United States, covering the period from 1963 to 2020[15]. Each factor represents a long-short portfolio that is based on a distinctive characteristic[16], such as momentum, value, or reversal. To fulfill the requirement of sample interchangeability necessary for leave-one-out and Lemma 1, we use volatility management in the style of (Moreira and Muir, 2017). Specifically, this means that for every factor, F_i, we compute the preceding

[14]The data is accessible online at jkpfactors.

[15]Our findings are robust across different datasets and size stratifications. For robustness checks, refer to the online appendix.

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

30-day realized volatility[17]:

$$\sigma_{F_i}(t-30,t) = \sqrt{\sum_{j=1}^{30} F_{i,j}^2} \tag{48}$$

and normalize the factor returns in the next period by this amount:

$$\tilde{F}_{i,t+1} = \frac{F_{i,t+1}}{\sigma_{F_i}(t-30,t)}. \tag{49}$$

For the rest of the empirics section, we continue to work with the notation F_t instead of $\tilde{F}_t$ for convenience. By volatility managing factor returns, we aim to eliminate the effects of heteroskedasticity and make our data closer to interchangeable.

5.2 Methodology

We estimate portfolio weights using a rolling window of T days and rebalance every 30 days. Thus, we end up constructing portfolios and SDFs for pricing monthly stock returns.[18] We fix the grid of ridge penalties,[19] $z = [10^i : i \in \{3, 2, \cdots, -3\}]$ and construct ridge shrunk Markowitz portfolio weights with $Z = (z_i)_{i=1}^L$:

$$\begin{aligned} \bar{\pi}_t(f_Z) &= (\bar{\pi}_t(z_i))_{i=1}^L, \\ \bar{\pi}_t(f_{z_i}) &= (\bar{E}[FF'](t-T,t) + z_i I)^{-1} \bar{E}[F](t-T,t), \end{aligned} \tag{50}$$

where $\bar{E}[F](t-T,t), \bar{E}[FF'](t-T,t)$ are sample means (5) estimated with the rolling window $[t-T,t]$. Given a vector of weights $W = (w_i)_{i=1}^L$, we construct the ridge shrinkage approximator from Lemma 4:

$$\bar{\pi}_t(f_{Z,W}) = \sum_{i=1}^{L} w_i \bar{\pi}_t(f_{z_i}). \tag{51}$$

With the estimates $\bar{E}[F](t-T,t), \bar{E}[FF'](t-T,t)$ in our hands, we compute leave-one-out returns using formula (27), and use these LOO returns to compute $\bar{\mu}(Z)(t-T,t)$, $\bar{\Sigma}(Z)(t-T,t)$ from Lemma 4, and then recover the optimal weight vectors $W_{\text{CUPSA}}(t-T,t)$ and

[17] results are very similar if we instead use standard deviations.

[18] Since most of the characteristics underlying our factors are updated at a monthly frequency, it is natural to use them for pricing stock returns at the monthly horizon. This is the conventional approach to asset pricing. The only reason we use daily data is to get better estimates of factor covariance matrices. (Kozak et al., 2020) follow the same approach.

[19] Our results are robust to the choice of the grid Z. Results for alternative choices of the grid are available upon request.

$W_{\text{UPSA}}(t-T,t)$ using Theorem 1. Everywhere in the sequel, we focus exclusively on CUPSA. While UPSA also achieves performance superior to that of classical shrinkage methodologies (such as those of (Ledoit and Wolf, 2020), as well as the simple ridge), we find that CUPSA strictly dominates UPSA in almost every experiment we run.[20] 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 "minimal shrinkage for large eigenvalues" for the shrunk covariance matrix. See the discussion after Corollary 2.

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

$$R_{t+1}(f_{Z,W_{CUPSA}(t-T,t)}) \;=\; \bar{\pi}_t(f_{Z,W_{CUPSA}(t-T,t)})'F_{t+1}. \tag{52}$$

By construction, $\bar{\pi}_t(f_{Z,W_{CUPSA}(t-T,t)})$ only depends on factor returns during the $[t-T,t]$ time interval and, hence, portfolio returns (52) are indeed OOS. We compare the performance of (52) with that of three main benchmarks

- **Best z**: $R_{t+1}(f_{z_*(t-T,t)})$, the best LOO-based ridge shrinkage utilizing the optimal penalty $z_*(t-T,t)$ computed in Corollary 4 computed using the data in the $[t-T,t]$ time interval.[21]

- **LW** (Ledoit-Wolf): $R_{t+1}(f_{LW}(t-T,t))$, where $f_{LW}(t-T,t)$ is the optimal non-parametric non-linear shrinkage from (Ledoit and Wolf, 2020), computed using the data in the $[t-T,t]$ time interval.[22]

- **Ridgeless**: $R_{t+1}(f_0)$, where $f_0(\lambda)=\lambda$. While the ridgeless limit is commonly defined as $\lim_{z\to 0}(zI+\bar{E}[FF'](t-T,t))^{-1}$, in practice, we just use the smallest value of the z-grid Z.[23]

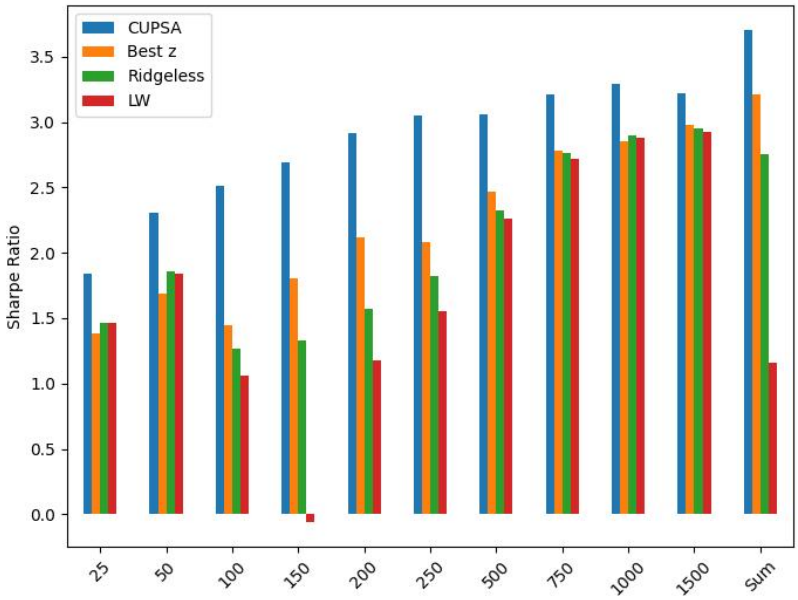

Figure 1: The plot compares the out-of-sample Sharpe ratio of our different benchmarks across different rolling windows T. Annualized Sharpe ratios are computed with monthly rebalancing, for the period 1977-11-22 to 2022-12-30. "Sum" reports the Sharpe ratios of summed returns across all different rolling windows. E.g., for CUPSA it is $\sum_{T\in\{25,50,\ldots,1500\}} R_{t+1}(f_{Z,W_{CUPSA}(t-T,t)})$.

5.3 The Performance of CUPSA

Figure 1 depicts the Sharpe ratios of the different benchmarks, for various rolling windows, starting from one month (25 days) to six years (1500 days). We see that CUPSA beats all other benchmarks by a significant margin, for every single rolling window. The gains are larger for shorter windows, consistent with the *complexity-based* interpretation advocated in Section 4.1: When T is small, complexity corrections in Lemma 2 need to be accounted for, and CUPSA does so successfully.

Interestingly, the percentage gain in the Sharpe ratio from using CUPSA is inverse-U-shaped with respect to the size of the training window, achieving its maximum for T around one year. The picture shows a clear hierarchy in performance,

$$LW \ < \ ridgeless \ < \ Best\ z \ < \ CUPSA\,. \tag{53}$$

This is particularly surprising given that the LW non-parametric shrinkage is a close relative of CUPSA. The underperformance of LW supports our intuition underlying the construction of CUPSA: The objective used for selecting the optimal shrinkage matters a lot, and simply minimizing the distance to the true covariance matrix (as does the LW estimator) might lead to highly sub-optimal shrinkage estimators. If we want to maximize the Sharpe ratio, we should pick shrinkage that maximizes the Sharpe ratio.

Figure 1 naturally raises two important questions:

(1) Why do we even need shorter rolling windows if the Sharpe ratio is monotone increasing in T?

(2) What is the statistical significance of the gain from using CUPSA?

Indeed, if the data is stationary, then using the possible longest window should be the preferred choice. However, the non-monotonic performance of Best z with respect to T suggests that there might be significant non-stationarity in the data. In this case, shorter rolling windows might capture some fast-changing market regimes at the cost of higher statistical errors produced by higher complexity (indeed, with $N = 153$, we have $c = 153/T$ is high for shorter rolling windows). To test the potential gains of exploiting

[20] Results for UPSA are available upon request.

[21] In practice, we just directly compute $z_* = \arg\max_{z\in Z} U^{OOS}_{LOO}(f_z)$.

[22] This algorithm minimizes the distance between the true and empirical covariance matrix. Distance is defined by using the Frobenius norm. We use the analytical version of this algorithm, available in (Ledoit and Wolf, 2020). (Ledoit and Wolf, 2017) discuss an application of a similar algorithm for minimum variance portfolios of stock returns.

[23] The ridgeless limit is theoretically optimal in the frictionless, zero complexity limit as $N/T \to 0$. in reality, $\bar{E}[FF']$ is highly degenerate, and using very small z often leads to instabilities.

shorter rolling windows, we report in the "Sum" column of Figure 1 the Sharpe ratio of $\sum_{T\in\{25,50,\ldots,1500\}} R_{t+1}(f_{Z,W_{CUPSA}(t-T,t)})$ as well as the corresponding Sharpe ratios for other three shrinkage methodologies. As one can see, the gains from using shorter rolling windows are significant (the Sharpe ratio increases from 3.2 to 3.7), suggesting that CUPSA based on shorter rolling windows is indeed able to capture some non-stationary patterns in the data, leading to superior performance.

To answer the question (2) above and evaluate the statistical significance of non-linear shrinkage, we run the following regression

$$\begin{aligned} R_{t+1}(f_{Z,W_{CUPSA}(t-T,t)}) \;=\; & \alpha + \beta_{z_*} R_{t+1}(f_{z_*(t-T,t)}) + \beta_{z_0} R_{t+1}(f_0(t-T,t)) + \beta_{z_\infty} R_{t+1}(f_\infty(t-T,t)) \\ & + \; R_{t+1}(f_{LW}(t-T,t)) + \beta_{MKT} MKT_{t+1} + \beta_{SMB} SMB_{t+1} + \beta_{HML} HML_{t+1} \\ & + \beta_{CMA} CMA_{t+1} + \beta_{RMA} RMA_{t+1} + \beta_{MOM} MOM_{t+1} + \varepsilon_{t+1}, \end{aligned} \tag{54}$$

where $R_{t+1}(f_{z_*})$ is the "Best z" portfolio return, $R_{t+1}(f_0)$ is the "ridgeless" portfolio return, and $R_{t+1}(f_{LW}(t-T,t))$ is the LW (Ledoit and Wolf, 2020) shrinkage portfolio return. In addition, we use the returns of the five Fama-French factors, (Fama and French, 2015), and momentum, (Jegadeesh and Titman, 1993), as controls.[24] 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. Finally, we add one more control, $R_{t+1}(f_\infty)$ corresponding to the limiting ridge portfolio as $z \to \infty$. By (15), this is a simple factor momentum portfolio (Arnott et al., 2023), (Gupta and Kelly, 2019), investing proportionally to realized mean returns of the factors.

Figure 2 reports the (heteroskedasticity-adjusted) t-statistics of α in (54). As one can clearly see, the t-statistics are large and significant for every single T, as well as for the sum across T's. We observe the same inverse-U-shaped pattern in the alpha t-statistic as for the gain in the Sharpe ratio (Figure 1): For Short windows, statistical estimation errors due to complexity are so large that even CUPSA has a hard time reducing them. Around $T = 250$ (corresponding to complexity $c = 153/250$,) the gains from CUPSA saturate (enough observations to estimate optimal shrinkage; enough complexity for shrinkage to be valuable). Finally, summing up across all windows once again produces very large gains, suggesting significant non-stationarity in the data.

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

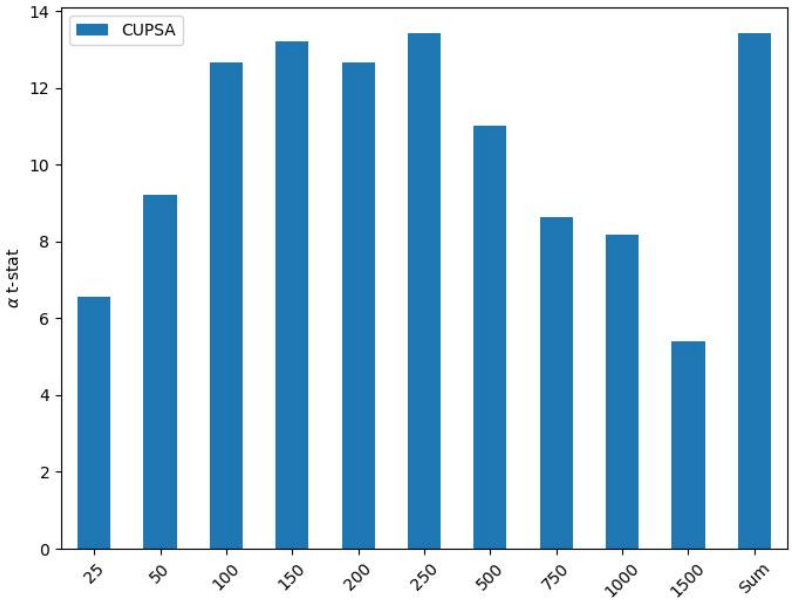

Figure 2: Heteroskedasticity-adjusted (with five lags) t-statistics of α from the regression (54) for different rolling windows. t-stats are computed for the period 1977-11-22 to 2022-12-30. "Sum" corresponds to summed returns across all different rolling windows. E.g., for CUPSA it is $\sum_{T\in\{25,50,\ldots,1500\}} R_{t+1}(f_{Z,W_{CUPSA}(t-T,t)})$.

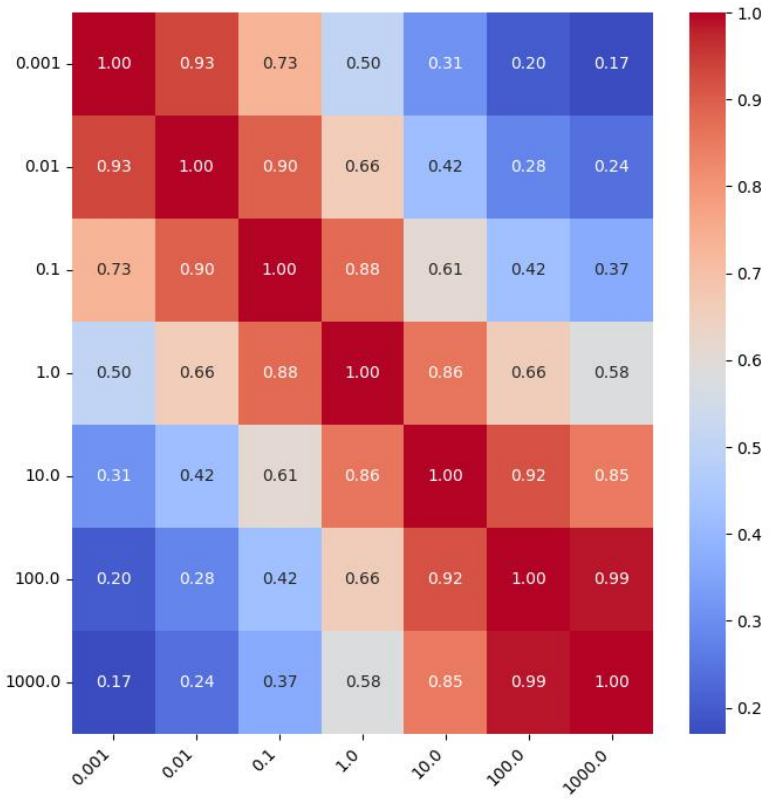

Figure 3: Correlation of out-of-sample returns of efficient portfolios with different levels of ridge shrinkage, $Corr(R_{t+1}(f_Z), R_{t+1}(f_Z))$. Correlations are computed over the period 1972-12-08 to 2022-12-30. Portfolios are estimated with a rolling window of $T = 250$ days, and re-balanced monthly.

5.4 Uderstanding the Dynamics of CUPSA

Everywhere in the sequel, we focus on $T = 250$[25]. Drawing on the insights from Theorem 1, we know that $CUPSA$ generates alpha by efficiently combining various "simple ridge" portfolios. Thus, the superior performance of CUPSA documented in the previous section suggests significant diversification gains (i.e., low correlations) across ridge portfolios. Figure 3 confirms this intuition: average correlations between low-z (i.e., Markowitz) and high-z (i.e., factor momentum (15)) are indeed low.

How does CUPSA achieve its performance? How does it select the optimal weights

[25]Results are very similar for different rolling windows

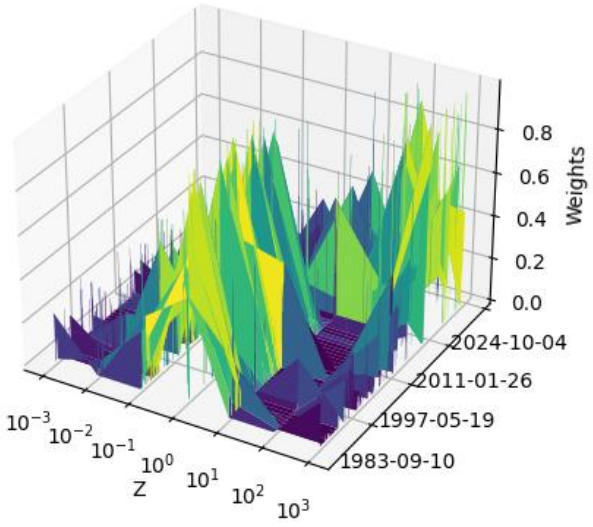

Figure 4: Weights for each ridge portfolio associated with the CUPSA strategy, $W_{CUPSA}(t-T,t)$. The weights are determined using Theorem 1 and computed over the period 1972-12-08 to 2022-12-30. Portfolios are estimated with a rolling window of $T = 250$ days, and re-balanced monthly.

W_{CUPSA}? To answer these questions, we report in Figures 4 and 5 the dynamics of the weight vector, $W_{CUPSA}(t-T,t)$ and the optimal ridge penalty $z_*(t-T,t)$ over time.

Both figures reveal consistent patterns. First, there was a clear regime shift around 2008 (after the Great Financial Crisis). Prior to 2008, both z_* and the shrinkage values $z \in Z$ with non-zero W_{CUPSA} weights are lower and relatively stable over time: While z_* oscillated between 1 and 0.1, CUPSA selected an optimal convex combination of these two 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 two together. These effects become particularly apparent in Figure 7 showing the histogram of the number of times a given ridge penalty has been chosen over the whole period. As one can see, this histogram is almost identical to the CUPSA weight box plot of 6 prior to 2008, 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.

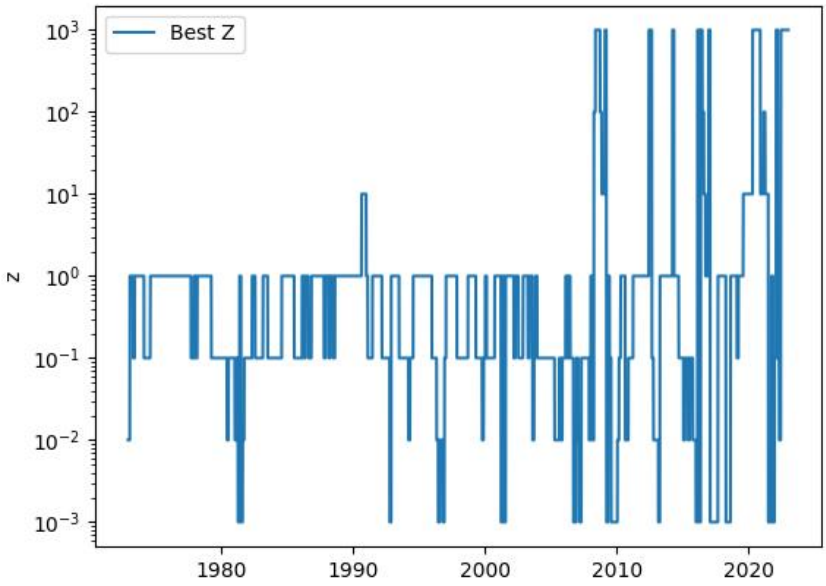

Figure 5: The figure shows the time series of optimal ridge shrinkage, $z_*(t-T,t)$. Optimal ridge shrinkage is chosen using Corollary 4 and computed over the period 1972-12-08 to 2022-12-30. Portfolios are estimated with a rolling window of $T = 250$ days, and re-balanced monthly.

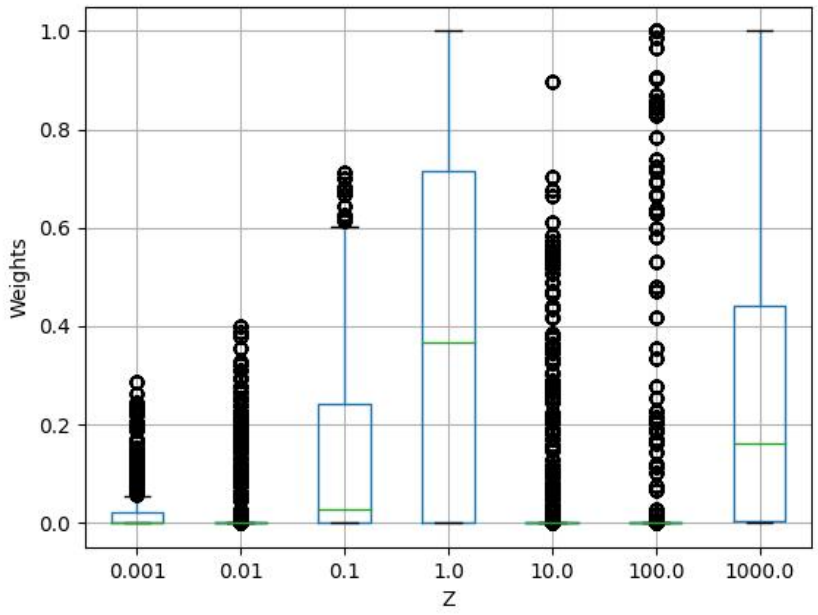

Figure 6: Box plot of ridge weights for the CUPSA strategy, $W_{CUPSA}(t-T,t)$. The weights are determined using Theorem 1 and computed over the period 1972-12-08 to 2022-12-30. Portfolios are estimated with a rolling window of $T = 250$ days, and re-balanced monthly.

Second, post-2008, the behavior of both W_{CUPSA} and z_* changed drastically. First, both objects became extremely unstable. Second, z_* fluctuates over a much wider range, often jumping all the way to 1000, and W_{CUPSA} behaves similarly, assigning significant weights to very large shrinkage levels in that period. See also Figure 6 that shows the distribution of CUPSA weights assigned to different ridge penalties across time: The three values, $0.1, 1.$, and 1000, clearly stand out.

Why is such large shrinkage necessary post-2008? As we show in Corollary 4, larger shrinkage is necessary when the size of factor risk premia drops and, hence, the optimality of large z is perfectly consistent with the significant drop in factor performance post-2008, documented, for example, in (Chordia et al., 2014).

Having understood the average behavior of CUPSA, we now turn to its dynamic properties. One of the most common arguments against the Markowitz portfolio (the ridgeless portfolio in our notation) is its instability. Small changes in the data often lead to huge jumps in portfolio weights, making it impractical to use in real-world applications. One way

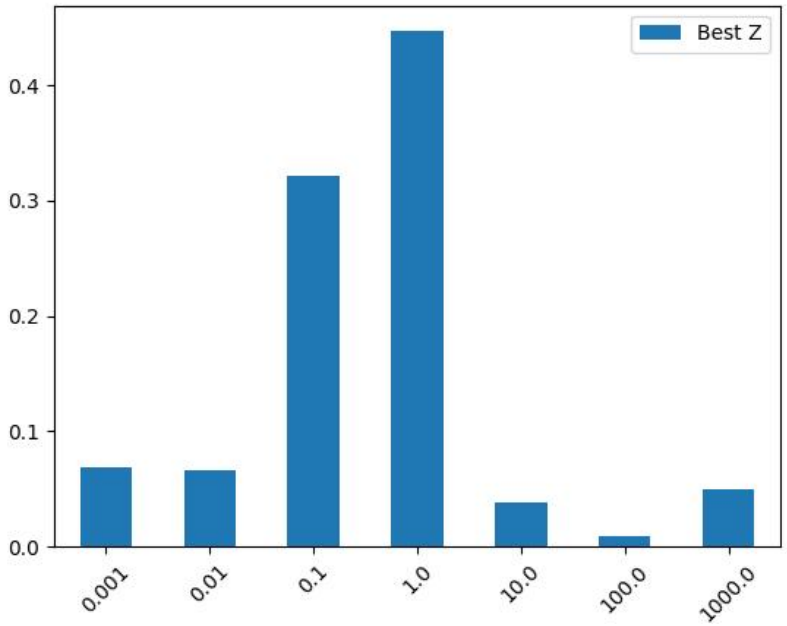

Figure 7: Histogram of the number of times a given ridge penalty has been chosen as $z_*(t-T,t)$. Optimal ridge is chosen using Corollary 4 and computed over the period 1972-12-08 to 2022-12-30. Portfolios are estimated with a rolling window of $T = 250$ days, and re-balanced monthly.

to measure these instabilities is by looking at portfolio turnover, defined as

$$turn_t = \bar{E}\left[\frac{|\bar{\pi}_{t,i} - \bar{\pi}_{t-1,i}|}{|\bar{\pi}_{t-1,i}|}\right], \tag{55}$$

where $\bar{\pi}t, i$ represents the portfolio weight allocated to factor i at time t.

We start by comparing average portfolio weights $\bar{\pi}_i$ in 8. As one can see, CUPSA portfolio weights are significantly less extreme than those of Best z : Diversification across multiple z smoothes the portfolio weight distribution.

The regularization effects of CUPSA are even more striking when we look at the turnover depicted in Figure 9. We see a remarkable difference between CUPSA and "Best z" approaches, whereby CUPSA reduced factor portfolio turnover by a factor of 5 to 8 *for every single factor theme*.[26] This drastic improvement in weight stability has major implications for the applicability of the general Markowitz methodology and the negative sentiment towards it in the finance profession. CUPSA results show that the Markowitz portfolio *can be made stable, useful, robust, and effective out-of-sample* when an optimal (albeit sophisticated) non-linear shrinkage methodology is applied.

5.5 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

$$M^*_{t+1} \;=\; 1 \;-\; \pi'_* F_{t+1} \tag{56}$$

satisfying the *zero pricing errors condition*

$$E[F_{i,t+1}\, M_{t+1}] \;=\; 0,\; i = 1, \cdots, N\,. \tag{57}$$

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

$$\bar{E}[F_{i,t+1}\, M_{t+1}(f_0)] \;=\; 0,\; i = 1, \cdots, N\,. \tag{58}$$

[26] We group 153 factors into 13 themes following the classification in (Jensen et al., 2023). See Section 5.5 for details.

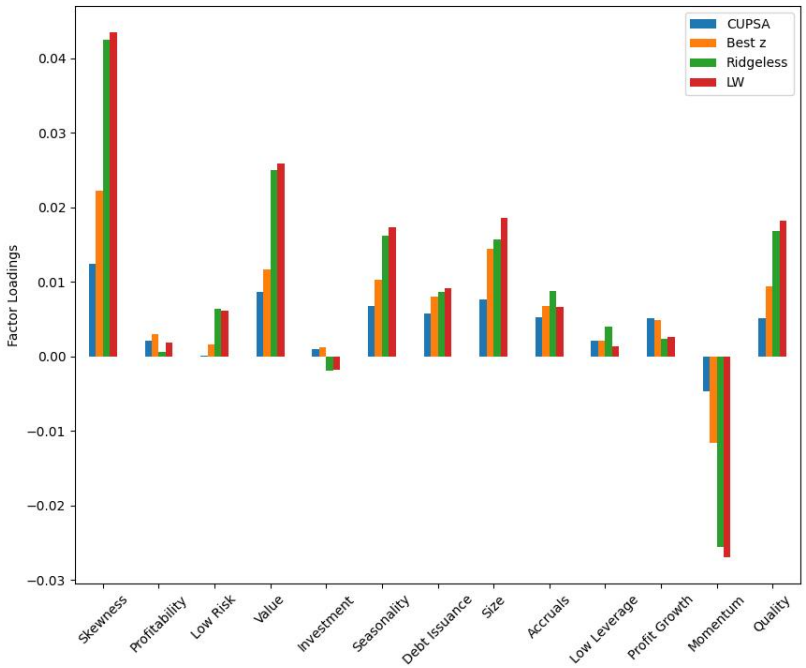

Figure 8: This figure shows average factor loadings, $\bar{E}[\pi]$, for CUPSA (Theorem 1),"Best z" (z_* of Corollary 4), ridgeless ($z \to 0$), and LW (Ledoit and Wolf, 2020) over the period 1972-12-08 to 2022-12-30. For any given factor theme (see (Jensen et al., 2023)), we take the average of loading across all factors in that theme. Portfolios are estimated with a rolling window of $T = 250$ days, and re-balanced monthly.

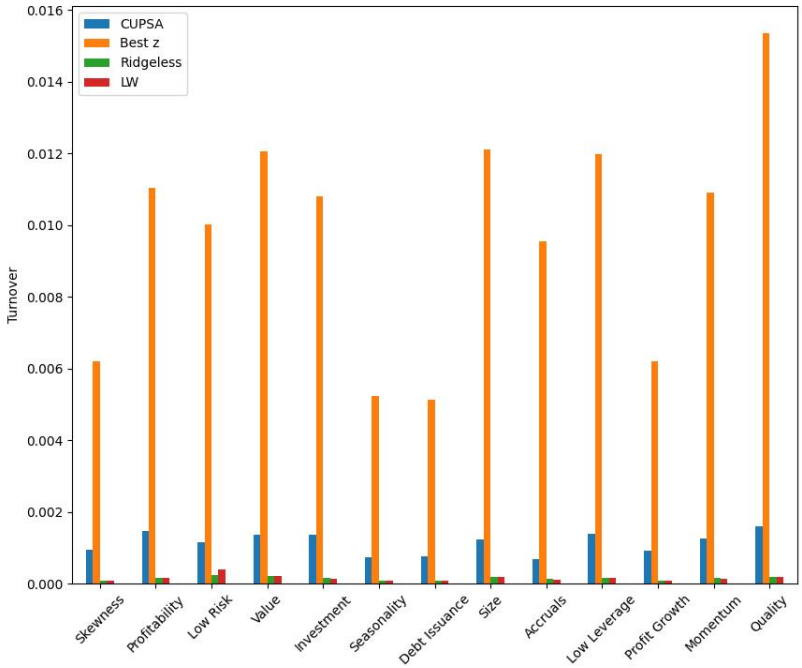

Figure 9: This figure shows average turnover, (55), for CUPSA (Theorem 1),"Best z" (z_* of Corollary 4), ridgeless ($z \to 0$), and LW (Ledoit and Wolf, 2020) over the period 1972-12-08 to 2022-12-30. Turnover is computed per theme, whereby, for any given factor theme (see (Jensen et al., 2023)), we take the average of turnover (55) across all factors in that theme. Portfolios are estimated with a rolling window of $T = 250$ days, and re-balanced monthly.

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:

$$M_{t+1}(CUPSA) \;=\; 1 \;-\; \bar{\pi}(f_{CUPSA})' F_{t+1}\,. \tag{59}$$

Intuitively, we expect CUPSA to produce small OOS pricing errors (while at the same time severely failing the IS pricing equation (58)). 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[27] of size T_{OOS}, we define

$$\bar{E}_{OOS}[X] \;=\; \frac{1}{T_{OOS}} \sum_{t=T+1}^{T_{OOS}} X_t\,, \tag{60}$$

and then the OOS pricing errors are defined as

$$PE_{OOS}(i) \;=\; \bar{E}_{OOS}[F_{i,t+1} M_{t+1}]\;,\; PE_{OOS} \;=\; (PE_{OOS}(i))_{i=1}^{N}\,. \tag{61}$$

The Hansen-Jagannathan distance is then defined using a weight matrix A (judiciously chosen by the researcher) as

$$D^{HJ}_{OOS}(A) \;=\; (PE_{OOS})' A (PE_{OOS})\,. \tag{62}$$

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.

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

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

$$M_{t+1}(f(t-T,t)) = 1 - \alpha_f R_{t+1}(f(t-T,t)), \tag{63}$$

where $f \in \{CUPSA,\ Best\ z,\ ridgeless,\ LW\}$ and the optimal scaling[28] α_f is

$$\alpha_f \ = \ \frac{\bar{E}_{OOS}[R_{t+1}(f(t-T,t))]}{\bar{E}_{OOS}[R_{t+1}(f(t-T,t))^2]}. \tag{64}$$

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

$$PE_j(f) \ = \ (\bar{E}_{OOS}[F_{i,t+1}M_{t+1}(f(t-T,t))])_{i \in theme_j}, \tag{65}$$

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

$$D^{HJ}_{theme_j}(f) \ = \ PE_j(f)' \bar{E}_{OOS}[F_{theme_j} F'_{theme_j}]^{-1} PE_j(f), \tag{66}$$

where,

$$F_{theme_j} \ = \ (F_i)_{i \in theme_j}. \tag{67}$$

Figure 10 reports these errors for all themes. We make several observations. First, the pricing error reductions achieved by shrinkage are very large: For most themes, using either z_* from Corollary 4 or CUPSA attains a very large reduction in pricing errors, sometimes by an order of magnitude. While the performance of the ridgeless portfolio is unsurprisingly dreadful, the magnitude of the gain from shrinkage is still surprising. While CUPSA dominates all alternatives for every single theme, shrinking with the "Best z" does

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

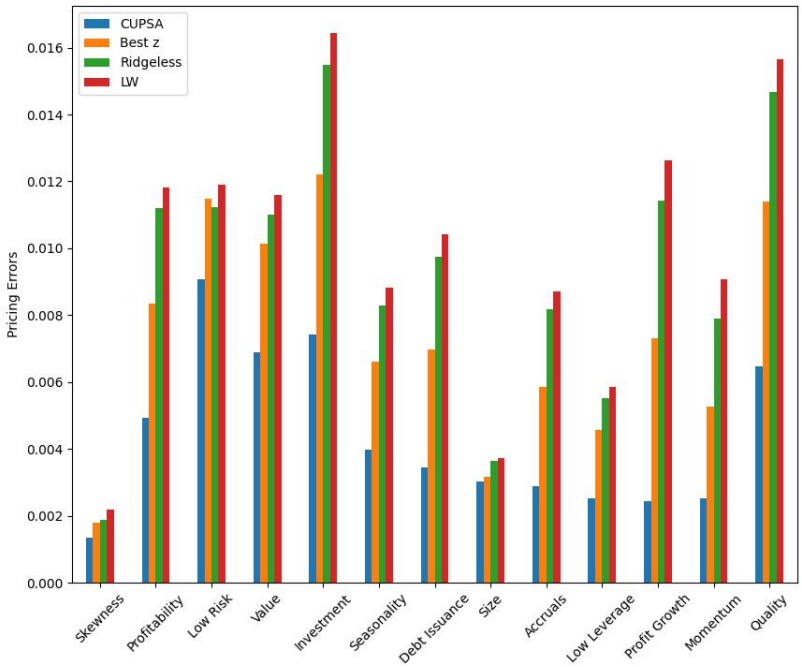

Figure 10: This figure shows the OOS HJ distance, (66), using SDFs from (63), for CUPSA (Theorem 1), "Best z" (z_* of Corollary 4), ridgeless ($z \to 0$), and LW (Ledoit and Wolf, 2020) over the period 1972-12-08 to 2022-12-30. Pricing errors are aggregated over themes as in (Jensen et al., 2023). Portfolios are estimated with a rolling window of $T = 250$ days, and re-balanced monthly.

achieve decent pricing errors for all themes, and its performance is comparable with that of CUPSA for all themes except profit growth, momentum, and quality. Given the discussion above, these findings suggest that factor risk premia for these three themes fluctuate a lot over time. For the latter two themes, there is indeed strong evidence for very large fluctuations in the corresponding risk premia. See, (Daniel and Moskowitz, 2016) and (Asness et al., 2019). We finally note that the low-risk and momentum factors are commonly viewed as being difficult to price because they are "anomalies" and do not reflect compensation for risk. Figure 10 suggests that these results might be driven by inefficient shrinkage: With optimal shrinkage, pricing errors for momentum and low-risk "anomalies" are comparable to those of other themes.

5.6 Non-Linearly Shrinking The Cross Section

The emergence of the factor zoo (Cochrane, 2011), (Harvey et al., 2016) and the failure of the attempts to find a characteristics-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 is necessary to construct efficient SDFs. In this section, we provide evidence of a significant *virtue of complexity* of the CUPSA-SDF in the space of PCs: The OOS performance of the CUPSA-SDF is monotone increasing in the number of PCs and keeps increasing even when we add very low-variance PCs. By contrast, Best z-SDF is indeed PC-sparse. Based on these surprising findings, we argue that the existing evidence for PC-sparse SDFs is likely an artifact of inefficient shrinkage.

Following the above approach, we compute PCs by decomposing factor returns covariance matrix $\bar{E}[FF'](t-T,t)$:[29]

$$\bar{E}[FF'](t-T,t) = \bar{U}(t-T,t)\,\mathrm{diag}(\bar{\lambda}(t-T,t))\bar{U}(t-T,t)'\,, \tag{68}$$

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

$$R^{PC}_{i,\tau}(t-T,t) \;=\; \bar{U}_i(t-T,t)'F_\tau\,. \tag{69}$$

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 $R^{PC}_\tau(I) \;=\; (R^{PC}_{i,\tau}(t-T,t))^I_{i=1}$ compute the in-sample ridge portfolios based on the in-sample covariance matrix $R^{PC}_\tau(I)$ (computed using $\tau \in [t-T,t]$).[30] We then apply all our shrinkage methodologies (CUPSA, ridgeless, Best z, and LW) to these returns and study their out-of-sample behavior, defined as

$$R^{PC}_{t+1}(I, f(t-T,t)) \;=\; \pi^{PC}(I)(f)'R^{PC}_\tau(I) \tag{70}$$

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

Figure 11 illustrates the resulting OOS Sharpe ratios. Our first observation is that, on our

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

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

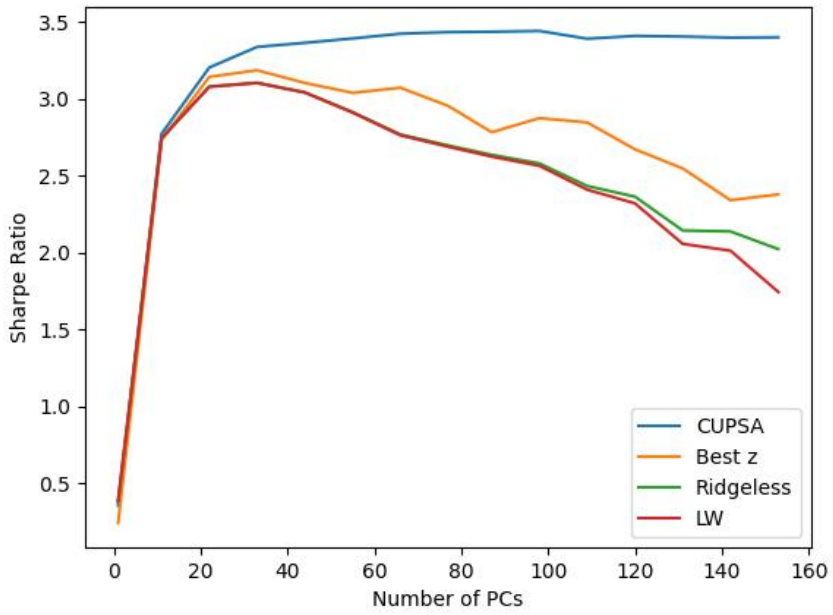

Figure 11: The figures show the Sharpe Ratio of PCs for CUPSA (Theorem 1), "Best z" (z_* of Corollary 4), ridgeless ($z \to 0$), and LW (Ledoit and Wolf, 2020) over the period 1972-12-08 to 2022-12-30, as the number of PCs grows. This is done by using (69). Portfolios are estimated with a rolling window of $T = 250$ days and re-balanced monthly.

dataset, all three "simpler" shrinkage methods (Best z, ridgeless, LW) saturate at around 20 PCs [31]. This implies that, even from the point of view of these shrinkage methods, there are at least 20 "factors" important for the cross-section of returns. This number is much higher than that for the SDF constructed in (Kozak et al., 2020), who argue that 5-10 PCs are sufficient to span the SDF.[32]

The most important implication of Figure 11 is the remarkable ability of CUPSA to capitalize on the *virtue of complexity*:[33] The fact that the OOS performance of CUPSA is monotonically increasing as we keep adding low-variance PCs. Another surprising implication of Figure 11 is the divergence between the performance of the simpler shrinkage estimators (LW, Best z, and ridgeless) that happens after the inclusion of top 10 PCs. This

[31] Figure 14 in the appendix illustrates that, across all rolling windows, this saturation point for Best z remains consistently lower than what is nominally observed with CUPSA Shrinkage

[32] (Kozak et al., 2020) use a different (much smaller) set of factors, but also get much lower Sharpe ratios. This suggests that our paper's larger set of factors from (Jensen et al., 2023) spans quantitatively important, additional risk premia.

[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.*

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. This optimal weighting allows CUPSA to exploit the diversification benefits in these PCs. By contrast, neither the Best z (which is too rigid and shrinks all low eigenvalues proportionally) nor the LW shrinkage (that completely ignores the risk-return tradeoffs and simply tries to minimize the estimation error of the covariance) are able to benefit from the low-variance PCs.

In the language of (Ledoit and Wolf, 2017), CUPSA follows the "Goldilocks rule," shrinking each principal component "just right," based on its unique ability to effectively identify and leverage the risk premiums embedded within the low-variance PCs. 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). Namely, Figure 11 suggests that the SDF might be driven by many orthogonal PCs, with risk premia that spread much more uniformly than conventional wisdom suggests. This uniform distribution of risk premia is a hallmark of the CUPSA approach.

To demonstrate this, we analyze the risk-return tradeoffs implied by each shrinkage method, as depicted in equation (18). Figure 12 illustrates the time series averages of these tradeoffs:

$$\frac{1}{T_{OOS}} \sum_{t=T+1}^{T+T_{OOS}} \bar{R}_i^{PC}(t-T,t)\, f(\lambda_i)(t-T,t)\,. \tag{71}$$

These quantities are crucial for evaluating how different shrinkage methods value each PC. Notably, the no-shrinkage strategy disproportionately favors smaller PC factors due to their minimal variance by choosing naive weights

$$\frac{\bar{R}_i^{PC}}{\lambda_i}\,, \tag{72}$$

effectively treating them as near-arbitrage opportunities. In contrast, the Ridge method opts for an average optimal shrinkage level, but this often leads to significant and unpredictable swings in the weights for low-variance PCs. By contrast, CUPSA adopts a more balanced approach, assigning almost equal importance to all PC factors. This equitable allocation aligns perfectly with our key intuition: CUPSA views all PC factors as having similar risk-return tradeoffs, emphasizing the importance of complexity where each PC is important. It is this uniform weighting of PCs that makes CUPSA an ideal shrinkage methodology for exploiting the high complexity of the SDF.

To strengthen our main message, we repeat the regression exercise (54) for the PC-based

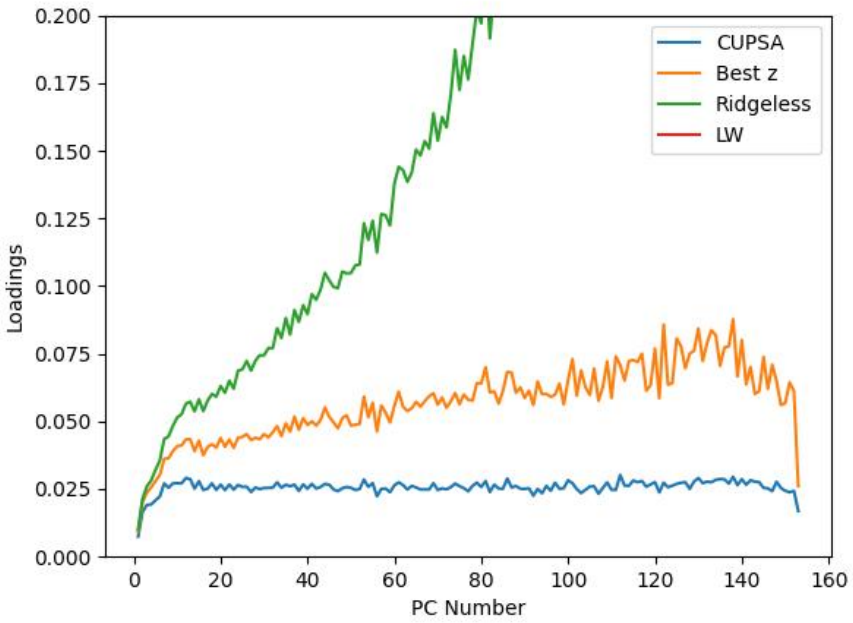

Figure 12: The figures show the average risk-return tradeoffs (factor loadings) for CUPSA (Theorem 1), "Best z" (z_* of Corollary 4), ridgeless ($z \to 0$), and LW (Ledoit and Wolf, 2020) over the period 1972-12-08 to 2022-12-30, for different PC factors. This is done by using (18). Portfolios are estimated with a rolling window of $T = 250$ days and re-balanced monthly.

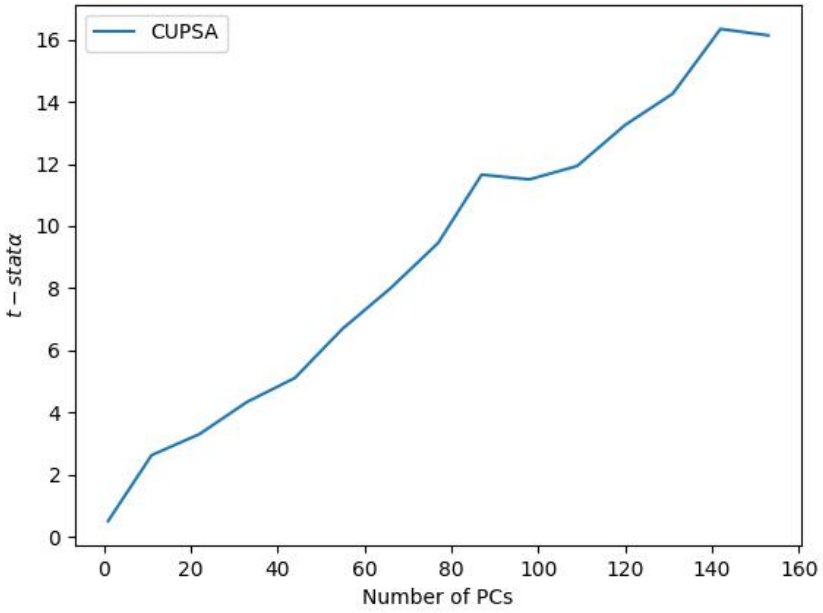

Figure 13: We plot the Heteroskedasticity-adjusted (with five lags) t-statistics of α of PC portfolios as the number of PCs grows. The regression is done using (54). PC portfolio returns are derived from (69). Portfolios are estimated with a rolling window of $T = 250$ days and rebalanced monthly.Please note that since LW and Ridgless exhibit similar risk-return tradeoffs, they are superimposed in the plot.

CUPSA portfolio returns (see (69)), gradually increasing the number of PCs used [34]. as set out in Equation (69). The results of this regression are reported in Figure 13. The t-statistics of alpha reveal a clear virtue of complexity and keep increasing as we add PCs, even beyond the PC number 100. This striking statistical pattern is consistent with our theoretical results, suggesting that the benefits of non-linear shrinkage are particularly large in high-dimensional settings when complexity corrections (Theorem 5) become particularly significant.

6 Conclusions

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:

[34]Figure 15 in the appendix demonstrates that, throughout various rolling windows, the inclusion of the saturation point for Best z as a control variable in regression (54) does not diminish the notable performance effectiveness of CUPSA Shrinkage.

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, taking into account 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 very large *virtue of complexity*, with its performance 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 optimally, adjusting to their risk-return tradeoffs. While standard shrinkage estimators (e.g., ridge) suggest that the optimal SDF should be PC-sparse, our results imply that sparsity is an artifact of inefficient shrinkage.

References

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

Asness, Clifford S, Andrea Frazzini, and Lasse Heje Pedersen, "Quality minus junk," *Review of Accounting Studies*, 2019, *24* (1), 34–112.

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.

Chordia, Tarun, Avanidhar Subrahmanyam, and Qing Tong, "Have capital market anomalies attenuated in the recent era of high liquidity and trading activity?," *Journal of Accounting and Economics*, 2014, *58* (1), 41–58.

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.

Daniel, Kent and Tobias J Moskowitz, "Momentum crashes," *Journal of Financial economics*, 2016, *122* (2), 221–247.

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.

Moreira, Alan and Tyler Muir, "Volatility-managed portfolios," *The Journal of Finance*, 2017, *72* (4), 1611–1644.

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.

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

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 relies on an application of the Stone-Weierstrass Theorem (Rudin, 1976). Consider the algebra of functions generated by the ridge ensemble $\{\Theta_z : z > 0\}$. 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 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 $[a, b]$ and vanishes nowhere. As a consequence, the algebra generated by the ridge ensemble is dense in $C(a, b)$ – by the Stone-Weierstrass Theorem – and the claim follows.

The next part of the proof is for the matrix monotone decreasing functions f. Suppose first

$$f(\lambda) = \sum w_i(\lambda + z_i)^{-1},\ w_i \geq 0,\ \sum_i w_i = 1\,.$$

Then, $f(\lambda)$ is matrix monotone decreasing by Löwner's theorem (Löwner, 1934), and

$$\lim_{\lambda\to\infty} \lambda f(\lambda) = \lim_{\lambda\to\infty} \sum w_i\lambda(\lambda + z_i)^{-1} = \sum w_i = 1\,.$$

Conversely, let f be a matrix monotone increasing function satisfying the technical condition

$f(\lambda)\lambda \to 1$ as $\lambda \to \infty$. Our goal is to show that there exists a sequence of functions

$$f_j(\lambda) \;=\; \sum_{i=1}^{n_j} w_{i,j}(\lambda + z_i)^{-1}$$

that uniformly converges to $f(\lambda)$ on compact intervals.

We define $g = -f$. Now g is matrix monotone increasing. By Löwner's theorem (Löwner, 1934), any *matrix monotone* increasing function on $(0, \infty)$ can be written as

$$\begin{aligned} g(\lambda) \;&=\; a\lambda + b + \int_0^\infty \frac{\lambda}{\lambda + z} d\mu(z) \\ &=\; a\lambda + b + \int_0^\infty \frac{\lambda + z - z}{\lambda + z} d\mu(z) \\ &=\; a\lambda + b + \int_0^\infty d\mu(z) - \int_0^\infty \frac{z}{\lambda + z} d\mu(z). \end{aligned} \tag{84}$$

for some $a \in \mathbb{R}_+$, $b \in \mathbb{R}$ and some positive, finite measure μ. By assumption, we have that $\lim_{\lambda\to\infty} g(\lambda)\lambda = -1$. In other words,

$$\lim_{\lambda\to\infty} g(\lambda)\lambda \;=\; \lim_{\lambda\to\infty} a\lambda^2 + \lim_{\lambda\to\infty} (b + \int_0^\infty d\mu(z))\lambda - \lim_{\lambda\to\infty} \int_0^\infty \frac{z\lambda}{\lambda + z} d\mu(z). \tag{85}$$

Since μ is a finite measure, we have

$$\lim_{\lambda\to\infty} \int_0^\infty \frac{z\lambda}{\lambda + z} d\mu(z) \;=\; \int_0^\infty z d\mu(z)\,.$$

If $a \neq 0$ or $b + \int_0^\infty d\mu(z) \neq 0$ then $\lim_{\lambda\to\infty} g(\lambda)\lambda$ will not converge. Therefore,

$$\begin{aligned} &a = 0 \\ &b + \int_0^\infty d\mu(z) \;=\; 0, \end{aligned} \tag{86}$$

and

$$f(\lambda) \;=\; \int_0^\infty \frac{z}{\lambda + z} d\mu(z). \tag{87}$$

with

$$\int_0^\infty z d\mu(z) \;=\; 1\,.$$

Let us define $d\tilde{\mu}(z) \;=\; z d\mu(z)$ so that

$$f(\lambda) \;=\; \int_0^\infty (\lambda + z)^{-1} d\tilde{\mu}(z)\,.$$

Let $\{\Delta_j\}_{j=1}^{n_j}$ be a partition of $[0,\infty)$ such that the Rieman-Stieltjes sums converges to the integral:

$$f(\lambda) \;=\; \lim_{n_j \to \infty} (\lambda + z_j)^{-1} w_j,\; w_j \;=\; \tilde{\mu}(\Delta_j)\,. \tag{88}$$

Clearly, the weights w_j sum up to one, and the proof is complete.

□

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

$$\bar{E}[R(f_Z)R(f_Z)'] \;=\; \bar{E}[F]'(Z_1 I + \bar{E}[FF'])^{-1}\bar{E}[FF'](Z_2 I + \bar{E}[FF'])^{-1}\bar{E}[F] \tag{89}$$

Thus, if $Z_0 = 0$, we have

$$\begin{aligned}(\bar{E}[R(f_Z)R(f_Z)']e_0)_i \;&=\; \bar{E}[F]'(z_i I + \bar{E}[FF'])^{-1}\bar{E}[FF'](0I + \bar{E}[FF'])^{-1}\bar{E}[F] \\ &=\; \bar{E}[F]'(z_i I + \bar{E}[FF'])^{-1}\bar{E}[F] = \bar{\mu}_{IS}(z_i),\end{aligned} \tag{90}$$

implying that

$$\bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(Z) \;=\; e_0, \tag{91}$$

so that no shrinkage is optimal.

$$\begin{aligned}\bar{\Sigma}(Z)^{-1}\bar{\mu}(Z) \;&=\; D(Z)^{-1}\big(\bar{\Sigma}_{IS}(Z) \;+\; \psi(Z)\psi(Z)' \;-\; (\bar{\mu}_{IS}(z)\psi(Z)' \;+\; \psi(Z)\bar{\mu}_{IS}(z'))\big)^{-1}(\bar{\mu}_{IS}(z) - \psi(Z)) \\ &=\; D(Z)^{-1}\big(\bar{\Sigma}_{IS}(Z) \;+\; \psi(Z)\psi(Z)' \;-\; (\bar{\mu}_{IS}(z)\psi(Z)' \;+\; \psi(Z)\bar{\mu}_{IS}(z'))\big)^{-1}\bar{\mu}_{IS}(z) \\ &-\; D(Z)^{-1}\big(\bar{\Sigma}_{IS}(Z) \;+\; \psi(Z)\psi(Z)' \;-\; (\bar{\mu}_{IS}(z)\psi(Z)' \;+\; \psi(Z)\bar{\mu}_{IS}(z'))\big)^{-1}\psi(Z) \\ &=\; D(Z)^{-1}term1 \;-\; D(Z)^{-1}term2\end{aligned} \tag{92}$$

For term2 we have

$$\begin{aligned} term2 &= \big(\bar{\Sigma}_{IS}(Z) + \psi(Z)\psi(Z)' - (\bar{\mu}_{IS}(z)\psi(Z)' + \psi(Z)\bar{\mu}_{IS}(z'))\big)^{-1}\psi(Z) \\ &\underbrace{=}_{Lemma\ 8} \frac{\big(\bar{\Sigma}_{IS}(Z) - (\bar{\mu}_{IS}(z)\psi(Z)' + \psi(Z)\bar{\mu}_{IS}(z'))\big)^{-1}\psi(Z)}{1+\psi(Z)'\big(\bar{\Sigma}_{IS}(Z) - (\bar{\mu}_{IS}(z)\psi(Z)' + \psi(Z)\bar{\mu}_{IS}(z'))\big)^{-1}\psi(Z)} \\ &= \frac{1}{d_{2,1}}\big(\bar{\Sigma}_{IS}(Z) - (\bar{\mu}_{IS}(z)\psi(Z)' + \psi(Z)\bar{\mu}_{IS}(z'))\big)^{-1}\psi(Z). \end{aligned} \tag{93}$$

Then

$$\begin{aligned} &\big(\bar{\Sigma}_{IS}(Z) - (\bar{\mu}_{IS}(z)\psi(Z)' + \psi(Z)\bar{\mu}_{IS}(z'))\big)^{-1}\psi(Z) \\ &\underbrace{=}_{Lemma\ 8} \big(\bar{\Sigma}_{IS}(Z) - \bar{\mu}_{IS}(z)\psi(Z)'\big)^{-1}\psi(Z) \\ &+ \frac{\big(\bar{\Sigma}_{IS}(Z) - \bar{\mu}_{IS}(z)\psi(Z)'\big)^{-1}\psi(Z)\bar{\mu}_{IS}(z)'\big(\bar{\Sigma}_{IS}(Z) - \bar{\mu}_{IS}(z)\psi(Z)'\big)^{-1}\psi(Z)}{1-\mu_{IS}(z)'\big(\bar{\Sigma}_{IS}(Z) - \bar{\mu}_{IS}(z)\psi(Z)'\big)^{-1}\psi(Z)} \\ &= \frac{\big(\bar{\Sigma}_{IS}(Z) - \bar{\mu}_{IS}(z)\psi(Z)'\big)^{-1}\psi(Z)}{1-\mu_{IS}(z)'\big(\bar{\Sigma}_{IS}(Z) - \bar{\mu}_{IS}(z)\psi(Z)'\big)^{-1}\psi(Z)} \\ &= \frac{1}{d_{2,2}}\big(\bar{\Sigma}_{IS}(Z) - \bar{\mu}_{IS}(z)\psi(Z)'\big)^{-1}\psi(Z). \end{aligned} \tag{94}$$

Finally,

$$\begin{aligned} &\big(\bar{\Sigma}_{IS}(Z) - \bar{\mu}_{IS}(z)\psi(Z)'\big)^{-1}\psi(Z) \\ &\underbrace{=}_{Lemma\ 8} \bar{\Sigma}_{IS}(Z)^{-1}\psi(Z) \\ &+ \frac{\bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(z)\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(z)} \\ &= \bar{\Sigma}_{IS}(Z)^{-1}\psi(Z) + \frac{e_0\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(Z)'e_0} \\ &= \bar{\Sigma}_{IS}(Z)^{-1}\psi(Z) + e_0\frac{\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(0)}. \end{aligned} \tag{95}$$

Therefore,

$$\begin{aligned} term2 &= \frac{1}{d_{2,1}d_{2,2}}\left(\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z) + e_0\frac{\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(0)}\right) \\ &= \alpha_2 e_0 + \beta_2\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z) \end{aligned} \tag{96}$$

For term1 the calculations will be slightly more involved:

$$
\begin{aligned}
term1 \;&=\; \big(\bar{\Sigma}_{IS}(Z) \;+\; \psi(Z)\psi(Z)' \;-\; (\bar{\mu}_{IS}(z)\psi(Z)' \;+\; \psi(Z)\bar{\mu}_{IS}(z)')\big)^{-1}\bar{\mu}_{IS}(z) \\
&\underbrace{=}_{Lemma\ 8} \big(\bar{\Sigma}_{IS}(Z) \;+\; \psi(Z)\psi(Z)' \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\bar{\mu}_{IS}(z) \\
&+\; \frac{\big(\bar{\Sigma}_{IS}(Z)+\psi(Z)\psi(Z)'-\psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\bar{\mu}_{IS}(z)\psi(Z)'\big(\bar{\Sigma}_{IS}(Z)+\psi(Z)\psi(Z)'-\psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\bar{\mu}_{IS}(z)}{1-\psi(Z)'\big(\bar{\Sigma}_{IS}(Z) \;+\; \psi(Z)\psi(Z)' \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\bar{\mu}_{IS}(z)} \\
&=\; \frac{\big(\bar{\Sigma}_{IS}(Z) \;+\; \psi(Z)\psi(Z)' \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\bar{\mu}_{IS}(z)}{1-\psi(Z)'\big(\bar{\Sigma}_{IS}(Z) \;+\; \psi(Z)\psi(Z)' \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\bar{\mu}_{IS}(z)} \\
&=\; \frac{1}{d_1}\big(\bar{\Sigma}_{IS}(Z) \;+\; \psi(Z)\psi(Z)' \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\bar{\mu}_{IS}(z).
\end{aligned}
\tag{97}
$$

Next

$$
\begin{aligned}
&\big(\bar{\Sigma}_{IS}(Z) \;+\; \psi(Z)\psi(Z)' \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\bar{\mu}_{IS}(z) \\
&\underbrace{=}_{Lemma\ 8} \big(\bar{\Sigma}_{IS}(Z) \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\bar{\mu}_{IS}(z) \\
&-\; \frac{\big(\bar{\Sigma}_{IS}(Z) \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\psi(Z)\psi(Z)'\big(\bar{\Sigma}_{IS}(Z) \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\bar{\mu}_{IS}(z)}{1+\psi(Z)'\big(\bar{\Sigma}_{IS}(Z) \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\psi(Z)} \\
&= term11 \;-\; term12
\end{aligned}
\tag{98}
$$

Note that

$$
\begin{aligned}
\big(\bar{\Sigma}_{IS}(Z) \;-\; \psi(Z)\bar{\mu}_{IS}(z)'\big)^{-1}\psi(Z) \;&\underbrace{=}_{Lemma\ 8}\; \frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(z)} \\
&=\; \frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(Z)'e_0} \\
&=\; \frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(0)}.
\end{aligned}
\tag{99}
$$

Plugging back

$$\begin{aligned} term12 &= \frac{\frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(z)}{(1-\psi(0))^2}}{1+\frac{\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(0)}} \\ &= \frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(z)}{(1-\psi(0))^2+(1-\psi(0))\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)} \\ &= \frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)\psi(Z)'e_0}{(1-\psi(0))^2+(1-\psi(0))\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)} \\ &= \frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)\psi(0)}{(1-\psi(0))^2+(1-\psi(0))\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}. \end{aligned} \tag{100}$$

Furthermore

$$\begin{aligned} term11 &= \left(\bar{\Sigma}_{IS}(Z) - \psi(Z)\bar{\mu}_{IS}(z)'\right)^{-1}\bar{\mu}_{IS}(z) \\ &\underbrace{=}_{Lemma\ 8} \bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(z)+\frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)\bar{\mu}_{IS}(z)'\bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(z)}{1-\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\bar{\mu}_{IS}(z)} \\ &= e_0+\frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(0)}. \end{aligned} \tag{101}$$

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

$$\begin{aligned} d_1 &= 1-\psi(Z)'e_0+\frac{\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(0)}-\frac{\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)\psi(0)}{(1-\psi(0))^2+(1-\psi(0))\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)} \\ &= 1-\psi(0)+\frac{\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(0)}-\frac{\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)\psi(0)}{(1-\psi(0))^2+(1-\psi(0))\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)} \\ &= 1-\psi(0)+\frac{qf}{1-\psi(0)}-\frac{qf\psi(0)}{(1-\psi(0))^2+(1-\psi(0))qf} \\ &= 1-\psi(0)+\frac{qf}{1-\psi(0)}(1-\frac{\psi(0)}{1-\psi(0)+qf}) \\ &= 1-\psi(0)+\frac{qf}{1-\psi(0)}(\frac{1-2\psi(0)+qf}{1-\psi(0)+qf}) \end{aligned} \tag{102}$$

Hence

$$\begin{aligned} term1 &= \frac{1}{d_1}\left(e_0+\frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}{1-\psi(0)}-\frac{\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)\psi(0)}{(1-\psi(0))^2+(1-\psi(0))\psi(Z)'\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z)}\right) \\ &= \alpha_1 e_0+\beta_1\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z) \end{aligned} \tag{103}$$

$$\begin{aligned}\bar{\Sigma}(Z)^{-1}\bar{\mu}(Z) &= D(Z)^{-1}(term1 \ - \ term2) \\ &= D(Z)^{-1}(\alpha e_0 + \beta\bar{\Sigma}_{IS}(Z)^{-1}\psi(Z))\end{aligned} \tag{104}$$

This concludes the proof of Theorem 5 □

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

$$\begin{aligned}\bar{\mu}(Z) &= D(Z)(\bar{\mu}_{IS}(z) - \underbrace{\psi(Z)}_{overfit}) \\ \bar{\Sigma}(Z) &= D(Z)\big(\bar{\Sigma}_{IS}(Z) \ + \ \underbrace{\psi(Z)\psi(Z)'}_{variance \ overfit} \ - \ \underbrace{(\bar{\mu}_{IS}(z)\psi(Z)' \ + \ \psi(Z)\bar{\mu}_{IS}(z)')}_{mean \ overfit}\big)D(Z)\end{aligned} \tag{105}$$

where,

$$\begin{aligned}\bar{\mu}_{IS}(z) &= \bar{E}[R(f_Z)], \\ \bar{\Sigma}_{IS}(Z) &= \bar{E}[R(f_Z)R(f_Z)']\end{aligned} \tag{106}$$

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, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} R_{T,t}(f_z) &= \frac{1}{T}\sum_{t=1}^{T} \frac{1}{1-\psi_t(z)}\big(R_t(f_z) \ - \ \psi_t(z)\big) \\ &= \frac{1}{1-\psi(z)}\big(\frac{1}{T}\sum_{t=1}^{T} R_t(f_z) \ - \ \psi(z)\big) \\ &= \frac{1}{1-\psi(z)}\big(\bar{\mu}_{IS}(z) \ - \ \psi(z)\big).\end{aligned} \tag{107}$$

In matrix form

$$\bar{\mu}(Z) = \text{diag}(\frac{1}{1-\psi(Z)})\big(\bar{\mu}_{IS}(z) \ - \ \psi(Z)\big) \tag{108}$$

Similarly, for the second moment matrix

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} R_{T,t}(f_{z_i})R_{T,t}(f_{z_j}) &= \frac{1}{T}\sum_{t=1}^{T} \frac{1}{1-\psi_t(z_i)}\big(R_t(f_{z_i}) - \psi_t(z_i)\big)\frac{1}{1-\psi_t(z_j)}\big(R_t(f_{z_j}) - \psi_t(z_j)\big) \\
&= \frac{1}{T}\sum_{t=1}^{T} \frac{1}{1-\psi(z_i)}\big(R_t(f_{z_i}) - \psi(z_i)\big)\frac{1}{1-\psi(z_j)}\big(R_t(f_{z_j}) - \psi(z_j)\big) \\
&= \frac{1}{(1-\psi(z_j))(1-\psi(z_i))}\frac{1}{T}\sum_{t=1}^{T}\big(R_t(f_{z_i})R_t(f_{z_j}) + \psi(z_i)\psi(z_j)\big) \\
&- \frac{1}{(1-\psi(z_j))(1-\psi(z_i))}\frac{1}{T}\sum_{t=1}^{T}\big(R_t(f_{z_j})\psi(z_i) + R_t(f_{z_i})\psi(z_j)\big) \\
&= \frac{1}{(1-\psi(z_j))(1-\psi(z_i))}\big(\bar{\Sigma}_{IS}(z_i,z_j) + \psi(z_i)\psi(z_j)\big) \\
&- \frac{1}{(1-\psi(z_j))(1-\psi(z_i))}\big(\bar{\mu}_{IS}(z_i)\psi(z_i) + \bar{\mu}_{IS}(z_i)\psi(z_j)\big).
\end{aligned} \tag{109}$$

In matrix form

$$\bar{\Sigma}(Z) = \mathrm{diag}(\frac{1}{1-\psi(Z)})\big(\bar{\Sigma}_{IS}(Z) + \psi(Z)\psi(Z)' - \bar{\mu}_{IS}(z)\psi(Z)' - \psi(Z)\bar{\mu}_{IS}(z)'\big)\,\mathrm{diag}(\frac{1}{1-\psi(Z)}) \tag{110}$$

This concludes the proof of Lemma 7. □

Proposition 6 (In-Sample Shrinkage is Not Optimal) *Assume $R_\tau(f_z)$ is the in-sample return at time τ,*

$$R_\tau(f_z) = \bar{\pi}(f_z)'F_\tau,\ \tau \le T. \tag{111}$$

Suppose

$$\bar{\mu}_{IS}(Z) = (\bar{E}[R_\tau(f_{z_i})])_{i=1}^{L} \quad \bar{\Sigma}_{IS}(Z) = (\bar{E}[R_\tau(f_{z_i})R_\tau(f_{z_i})'])_{i=1}^{L} \tag{112}$$

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

$$a(z) = \bar{\mu}_{IS}(z) - 0.5\bar{\Sigma}_{IS}(z) - 0.5 \tag{113}$$

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 Lemma 6. Let

$$\begin{aligned} a(z) &= \bar{\mu}_{IS}(z) - 0.5\bar{\Sigma}_{IS}(z) - 0.5 \\ &= \bar{E}[F]'(zI + \bar{E}[FF'])^{-1}\bar{E}[F] - 0.5 \\ &- 0.5\bar{E}[F]'(zI + \bar{E}[FF'])^{-1}\bar{E}[FF'](zI + \bar{E}[FF'])^{-1}\bar{E}[F] - 0.5 \\ &= \bar{E}[F]'(zI + \bar{E}[FF'])^{-2}(zI + \bar{E}[FF'] - 0.5\bar{E}[FF'])\bar{E}[F] - 0.5 \\ &= \bar{E}[F]'(zI + \bar{E}[FF'])^{-2}(zI + 0.5\bar{E}[FF'])\bar{E}[F] - 0.5. \end{aligned} \tag{114}$$

We compute the derivative of $a(z)$

$$\begin{aligned} a'(z) &= \bar{\mu}'_{IS}(z) - 0.5\bar{\Sigma}'_{IS}(z) \\ &= -z\bar{E}[F]'(zI + \bar{E}[FF'])^{-2}\bar{E}[F] \\ &+ z\bar{E}[F]'(zI + \bar{E}[FF'])^{-3}\bar{E}[FF']\bar{E}[F] \\ &= z\bar{E}[F]'(zI + \bar{E}[FF'])^{-3}(\bar{E}[FF'] - (zI + \bar{E}[FF']))\bar{E}[F] \\ &= -z^2\bar{E}[F]'(zI + \bar{E}[FF'])^{-3}\bar{E}[F]. \end{aligned} \tag{115}$$

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

□

Proof of Corollary 4. We have

$$\begin{aligned} \frac{d}{dz}(\bar{\mu}(z) - 0.5\bar{\Sigma}(z)) &= \frac{d}{dz}(\frac{\bar{\mu}_{IS}(z) - \psi(z)}{1 - \psi(z)} - 0.5\frac{\bar{\Sigma}_{IS}(z) + \psi^2(z) - 2\bar{\mu}_{IS}(z)\psi(z)}{(1 - \psi(z))^2}) \\ &= \frac{d}{dz}(\frac{\bar{\mu}_{IS}(z) - \psi(z) - \bar{\mu}_{IS}(z)\psi(z) + \psi(z)^2}{(1 - \psi(z))^2} - 0.5\frac{\bar{\Sigma}_{IS}(z) + \psi^2(z) - 2\bar{\mu}_{IS}(z)\psi(z)}{(1 - \psi(z))^2}) \\ &= \frac{d}{dz}(\frac{\bar{\mu}_{IS}(z) - \psi(z)}{(1 - \psi(z))^2} - 0.5\frac{\bar{\Sigma}_{IS}(z) - \psi^2(z)}{(1 - \psi(z))^2}) \\ &= \frac{d}{dz}(\frac{\bar{\mu}_{IS}(z) - 0.5\bar{\Sigma}_{IS}(z) - \psi(z) + 0.5\psi^2(z)}{(1 - \psi(z))^2}) \\ &= \frac{d}{dz}(\frac{\bar{\mu}_{IS}(z) - 0.5\bar{\Sigma}_{IS}(z) + 0.5(1 - \psi(z))^2 - 0.5}{(1 - \psi(z))^2}) \\ &= \frac{d}{dz}(\frac{\bar{\mu}_{IS}(z) - 0.5\bar{\Sigma}_{IS}(z) - 0.5}{(1 - \psi(z))^2}) \end{aligned} \tag{116}$$

Note that from Lemma 6, $a(z)$ is decreasing in z and $a(z) < 0$. Let also $b(z) = (1-\psi(z))^2$.

Therefore the OOS quadratic utility objective can be written as

$$(a(z)/b(z))' = \frac{a'(z)b(z) - a(z)b(z)'}{b(z)^2} \tag{117}$$

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

$$b'(0) = -2\psi'(z)(1 - \psi(z)) \tag{118}$$

Lemma 8 will ensure that $\psi(z) < 1$. On the other hand, from the definition of $\psi(z)$

$$\psi(z) = \frac{1}{T}F_\tau'(zI + \bar{E}[FF'])^{-1}F_\tau \tag{119}$$

hence increasing z will decrease $\psi(z)$. This concludes the Proof of Corollary 4.

□

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,*

$$(A + uv')^{-1} = A^{-1} - \frac{A^{-1}uv'A^{-1}}{1 + v'A^{-1}u} \tag{120}$$

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

$$\bar{\pi}(z) = (\bar{E}[FF'])^{-1}\bar{E}[F]. \tag{121}$$

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

$$\begin{aligned}\bar{\pi}^{PC}(z) &= (\bar{E}[F^{PC}F^{PC\prime}])^{-1}\bar{E}[F^{PC}] \\ &= (U'\bar{E}[FF']U)^{-1}U'\bar{E}[F] \\ &= (D)^{-1}U'\bar{E}[F]\end{aligned} \tag{122}$$

The OOS performances of these two portfolios are equal

$$\begin{aligned}F_{OOS}'\bar{\pi}(z) &= F_{OOS}^{PC\prime}\bar{\pi}^{PC}(z) \\ &= F_{OOS}'U(D)^{-1}U'\bar{E}[F] \\ &= F_{OOS}'(\bar{E}[FF'])^{-1}\bar{E}[F]\end{aligned} \tag{123}$$

B Additional Results

We evaluate the efficacy of CUPSA against the *best infeasible sparse model with optimal ridge shrinkage*, denoted as Sparse z. The model is defined as follows:

$$\begin{aligned} &R^{SZ}_{t+1}(z_*(t-T,t)) = R^{PC}_{t+1}(I_*, z_*(t-T,t)), \\ &I_* = \underset{I}{\operatorname{argmax}}\, \bar{E}_{OOS}[U(R^{PC}_{t+1}(I, z_*(t-T,t)))]. \end{aligned} \tag{124}$$

Here, $R^{PC}_{t+1}(I, z(t-T,t))$ represents the return on the top I PC factors, as detailed in Equation (70). The term $z_*(t-T,t)$ refers to the Best z shrinkage, derived from Corollary 4. In essence, Sparse z (SZ) is selected to maximize the OOS Sharpe ratio from all possible PC factor SDFs. For instance, if the optimal Sharpe ratio is achieved with the top 20 PC factors, then the portfolio comprising these top 20 PCs will be designated as the efficient portfolio and utilized in constructing the SDF. The infeasibility of the model arises because $I_\star$ in Equation (124) is selected ex-post, affording us the advantage of hindsight in determining the number of Principal Component (PC) factors that will yield the best Out-of-Sample (OOS) performance. From Figure 14, it is evident that CUPSA outperforms all Sparse z models. This implies that even with the advantage of knowing the best combination of PC factors, ridge shrinkage still falls short of the non-linear shrinkage efficiency offered by CUPSA. This observation holds true even when Sparse z is incorporated into regression (54). The t-statistics remain large and statistically significant across all rolling windows. It appears that even with this 'cheating' approach, the elusive dream of sparsity cannot be fully realized, suggesting that embracing the chaos of complexity might indeed be the more prudent strategy

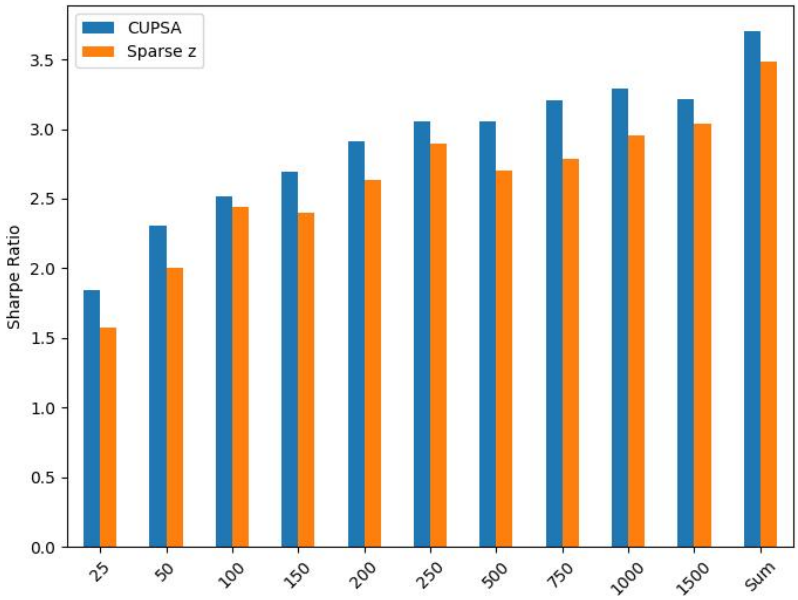

Figure 14: The plot compares the out-of-sample Sharpe ratio of CUPSA (Theorem 1) and Sparse z ((124)) across different rolling windows T. Annualized Sharpe ratios are computed with monthly rebalancing, for the period 1977-11-22 to 2022-12-30. "Sum" reports the Sharpe ratios of summed returns across all different rolling windows. E.g., for CUPSA it is $\sum_{T\in\{25,50,\ldots,1500\}} R_{t+1}(f_{Z,W_{CUPSA}(t-T,t)})$.

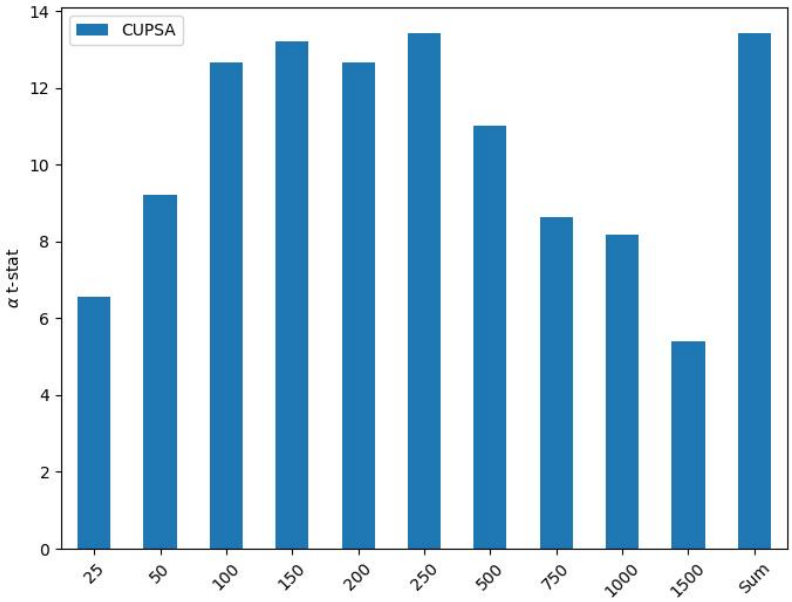

Figure 15: Heteroskedasticity-adjusted (with five lags) t-statistics of α from the regression (54) with the addition of Sparse z ((124)) for different rolling windows. t-stats are computed for the period 1977-11-22 to 2022-12-30. "Sum" corresponds to summed returns across all different rolling windows. E.g., for CUPSA it is $\sum_{T\in\{25,50,\ldots,1500\}} R_{t+1}(f_{Z,W_{CUPSA}(t-T,t)})$.