What Do Large Factor Models Learn? Self-Induced Regularization, Cost of Overfitting, and Self-Adaptivity

Xin Xiong*

March 16, 2026

Abstract

This paper studies what large linear factor models learn when estimating stochastic discount factors (SDFs) from factor libraries that may contain many noisy or weak signals. This question is particularly relevant in empirical asset pricing, where researchers increasingly work with large characteristic-sorted, machine-learned, or otherwise constructed factor sets. We analyze the *all-inclusive ridge estimator*, which uses all candidate factors without ex ante screening or dimension reduction, and derive *non-asymptotic bounds* for its out-of-sample pricing error in the inherently misspecified SDF estimation. Our theory yields three main insights. First, adding many low-variance factors generates self-induced regularization: these factors implicitly increase shrinkage on the high-variance directions most relevant for pricing, thereby introducing bias into the estimated SDF. Second, exact interpolation in low-variance directions has a bounded out-of-sample cost: although the estimator may fit these components perfectly in sample, they behave approximately like zero out of sample. Third, the all-inclusive ridge estimator is self-adaptive: even without prior knowledge of the effective dimension, it behaves like a data-driven principal-component procedure that emphasizes strong directions and downweights weak ones. These results have direct empirical implications. Overparameterization can harm out-of-sample pricing performance mainly through implicit bias rather than variance inflation, so adding noise factors, or even weak but priced factors, need not improve performance. More importantly, because overparameterization itself creates shrinkage, the optimal explicit ridge penalty can be negative. Mildly negative ridge penalties can therefore improve out-of-sample performance by offsetting self-induced regularization. Simulations and U.S. equity evidence based on large random-Fourier-feature factor sets support these predictions.

*MIT Sloan School of Management. Email: xinxiong@mit.edu. I am grateful to Hui Chen, Leonid Kogan, and David Thesmar for helpful comments and suggestions. I also thank participants at the MIT Finance Lunch for valuable feedback. Any remaining errors are my own.

1 Introduction

The proliferation of proposed asset pricing factors has given rise to an important question: How can we construct robust stochastic discount factors (SDFs) when the set of candidate factors is large and potentially noisy? Traditional approaches and conventional wisdom favor parsimonious models based on carefully selected factors (Ross, 1976; Fama and French, 1993, 2015). Recent advances in machine learning, however, suggest that high-dimensional models may still generalize well despite overparameterization if appropriately regularized, or sometimes even in the absence of explicit regularization (Belkin et al., 2019; Bartlett et al., 2020; Hastie et al., 2022). Motivated by these perspectives, this paper builds on recent discussions of large, overparameterized linear factor models for SDF estimation (Kozak et al., 2020; Kelly and Xiu, 2023; Didisheim et al., 2024), focusing on the behavior and performance of the *all-inclusive ridge estimator*. It is a ridge-regularized estimator incorporating all candidate factors without any pre-screening or dimension reduction steps. We establish *non-asymptotic (finite-sample) pricing error bounds* for large factor models and use them to examine our research questions concerning overfitting, model complexity, and regularization in SDF estimation.

First, we study the role of noise factors. Specifically, we consider a K-factor model and investigate the impact of adding independent noise factors (defined as factors with zero risk premium) into the factor space. We ask the following questions: How does the inclusion of noise factors affect the out-of-sample pricing error? Under what conditions are noise factors harmful, and when might they be benign? Second, we examine the role of weak factors. Weak factors are defined as those with small individual risk premia and variances. Analogous to the previous setting, we analyze the consequences of introducing independent weak factors into a K-factor model. In particular, we explore whether, in cases where many weak factors collectively contribute to a substantial aggregate risk premium, large factor models can correctly price them out-of-sample. Third, we investigate the phenomenon of negative ridge regularization. We study whether there exist scenarios in which the optimal ridge penalty is negative, and explore why such scenarios may occur.

1.1 Main Results

Our main theoretical results (see Theorem 1 and its more interpretable Corollary 1) provide three key insights on what the all-inclusive ridge estimator learns in large factor models, under appropriate assumptions. These insights emphasizes the structure of principal component (PC) factors, and we describe them below.

One insight concerns the **self-induced regularization** that arises in the high-variance PC factor space. Including many low-variance PC factors implicitly increases the effective penalty on the high-variance PC factors, shrinking the estimated SDF and biasing performance. Another insight relates to the **bounded cost of overfitting** in the low-variance PC factor space. Despite exact interpolation, the fitted coefficients on low-variance PC factors behave like zero out-of-sample. As a result, overfitting in this space does not amplify pricing error beyond what under-specification would incur. The final insight highlights the idea of **self-adaptivity**. Without knowing which factors are important, the ridge estimator adaptively emphasizes top PC factors and suppresses the rest, effectively mimicking principal component

regression with a data-dependent cutoff, modulo a shift due to self-induced regularization.

We then use these results to answer the research questions previously outlined and provide the following answers.

(i) **Role of noise factors.** Adding unpriced factors *degrades* out-of-sample performance by narrowing the effective tuning range of the ridge penalty and introducing unnecessary bias. The harm comes not from overfitting but from self-induced regularization.

(ii) **Role of weak factors.** Even when all factors are priced, the ridge estimator may *fail* to recover signals from weak factors. The benefit of large factor models lies not in capturing every small signal, but in automatically selecting the right number of factors to recover signal.

(iii) **Negative ridge.** In the presence of self-induced regularization, the optimal penalty may lie below zero. Our results suggest the possibility of using negative ridge to offset self-induced regularization and improve pricing performance.

Lastly, we validate these insights in two complementary ways. First, we run controlled simulations directly in the PC space. Second, we test the predictions in an empirical study using U.S. stock markets data. Our results show that adding noise factors systematically degrades performance through self-induced bias rather than variance inflation. Adding weak yet priced factors can hurt model performance as well. Perhaps surprisingly, we observe that mildly negative ridge penalties can enhance out-of-sample pricing performance.

1.2 Comparison to Prior Work

1.2.1 Benign Overfitting in Overparametrized Linear Regression

Our work is closely related to the recent machine learning literature on *benign overfitting* in overparameterized linear models (Bartlett et al., 2020; Shamir, 2023; Tsigler and Bartlett, 2023; Barzilai and Shamir, 2024; Lecué and Shang, 2024; Gavrilopoulos et al., 2024), which investigates how certain estimators that perfectly fit high-dimensional training data can still generalize well on test data under suitable conditions on the data-generating process (DGP) and noise structure.

However, existing results largely focus on well-specified linear regression with i.i.d. zero-mean noise, i.e., $Y = \boldsymbol{\beta}^\top \boldsymbol{X} + \varepsilon$, with $\mathbb{E}[\varepsilon] = 0$ and $\varepsilon \perp \boldsymbol{X}$ (with the rare exception of Shamir (2023)). In contrast, the SDF estimation problem can be written as $1 = \boldsymbol{\beta}^\top \boldsymbol{F} + M$, where M is the SDF. While this superficially resembles linear regression, it is fundamentally *mis-specified* linear regression: M is not mean-zero and is *dependent* on $\boldsymbol{F}$, violating key assumptions underlying the first steps in most benign overfitting analyses (e.g., equation (2) and Lemma 22 in Tsigler and Bartlett (2023)). Indeed, Shamir (2023) shows that benign overfitting generally *fails* to hold in such misspecified settings.

Our main theoretical contribution is to extend the logic of benign overfitting to this inherently misspecified regime, by leveraging the unique structure of the SDF estimation problem (most notably, the constant outcome $Y \equiv 1$) and additional economic assumptions on the DGP, such as Gaussianity and the absence of asymptotic arbitrage. Our results (see Theorem 1) demonstrate that benign overfitting can still occur under a new set of conditions.

Methodologically, our approach draws on the *comparator-based* viewpoint introduced by Shamir (2023), and probabilistic tools from Lecué and Shang (2024) and Tsigler and Bartlett (2023) to control the norms and spectra of random matrices.

1.2.2 Asymptotic Characterization of Large Factor Models

The asymptotic properties of ridge regression in large factor models were first studied in Didisheim et al. (2024), which characterizes the limiting out-of-sample Sharpe ratio and pricing error. Our non-asymptotic results complement this work in several important ways.

One important difference is the finite-sample validity and convergence rates. Our results provide non-asymptotic guarantees that hold uniformly for all finite P, T (i.e., allowing arbitrary relationships between them). In contrast, asymptotic analyses require $P, T \to \infty$ with $P/T \to c$, and do not offer convergence rates. Our bounds quantify how fast ridge estimators approach their "effective" asymptotic limits, enabling meaningful comparisons in practical sample sizes and dimensions.

A second difference that is central in practice concerns *penalty tuning* and *boundary regimes*. Our framework permits ridge penalties z that vary with P and T, and it covers the ridgeless and near-ridgeless regimes ($z = 0$ and $z \to 0$), as well as the negative-regularization regime ($z < 0$ whenever the relevant matrices remain invertible). This matters because in practice z is usually tuned based on observed (P, T), and finite-sample theory predicts the optimal choice depends on (P, T) rather than being a fixed constant (Zhang, 2023). In contrast, asymptotic analyses like Didisheim et al. (2024) typically fix $z > 0$ independent of sample size and dimension; as a result, they are not designed to track data-dependent tuning rules $z = z(P, T)$ at finite (P, T), nor to fully characterize the boundary regimes $z \to 0$ (and especially $z = 0$), and they do not address negative regularization.[1]

Finally, our approach offers interpretability via factor space decomposition. By decomposing factor space into estimation and overfitting components, our approach yields three new insights (see Section 3.4) that are not visible in asymptotic limits. The notion of *self-induced regularization*, which we quantify explicitly, complements the "implicit shrinkage" phenomenon observed in Didisheim et al. (2024) by attributing shrinkage to specific components of the factor space.

Nevertheless, asymptotic analysis provides exact limiting constants, while our bounds are stated up to universal constants. These perspectives are complementary: we focus on wide applicability across settings and interpretability, while Didisheim et al. (2024) emphasizes precision under asymptotic regimes.

Alongside asymptotic characterizations, there has also been a recent debate on how to interpret "virtue of complexity" evidence when estimation is carried out in very short samples relative to model dimension. Nagel (2025) argues that in regimes with $P \gg T$, random-feature forecasts can behave like a similarity-weighted average of in-sample outcomes, so strong backtest performance may partly reflect mechanical recency- and volatility-driven effects rather

[1] This distinction is substantive from a theoretical point of view: asymptotic analyses that aim to capture the boundary $z \to 0$ (and especially $z = 0$) generally require much stronger assumptions of feature covariance structures (Wu and Xu, 2020; Hastie et al., 2022; Kelly et al., 2024). In particular, covariance structures with substantial near-zero spectral mass can fall outside the scope of asymptotic ridgeless/near-ridgeless characterizations, yet are directly covered by our finite-sample theory.

than stable learning from predictors. Kelly and Malamud (2025) respond with additional discussion and evidence, emphasizing effective complexity and implicit regularization, and they also address related critiques concerning aggregation and benchmarking choices (including intercept restrictions), economic interpretation and the meaning of complexity under shrinkage, and noisy-signal environments and claims of a reversal of "virtue of complexity." These exchanges largely concern return prediction, whereas our paper focuses on SDF estimation. In terms of methodology, our results also underscore the role of feature-map construction, since our theoretical results hinge on the decay pattern of its eigenvalues. This perspective is broadly consistent with the general message in Li et al. (2025) that feature construction can be a first-order determinant of what high-dimensional methods deliver.

Our analysis in this paper is oriented toward introducing mainstream high-dimensional, non-asymptotic tools to the SDF estimation problem. Leveraging theoretical results that allow for finite samples and dimensions and accommodate a broad class of factor covariance structures, we can address a new set of research questions outlined in Section 1 and uncover a set of new insights. Furthermore, our empirical framework builds directly on that of Didisheim et al. (2024), which in turn follows Kelly et al. (2024), enabling consistent benchmarking and validation on real-world data.

1.3 Additional Literature

1.3.1 High-Dimensional SDF Estimation

Our work contributes to a growing literature on high-dimensional SDF estimation and factor modeling. A prominent line of research addresses high-dimensionality by imposing *explicit regularization* or performing *factor selection* via principle component analysis (PCA). Kozak et al. (2020) introduce a robust SDF by applying heavy ridge shrinkage to a rich set of characteristics, allowing large models while mitigating overfitting. They also introduce a PCA-based factor selection method that results in small factor models that perform well. Kelly et al. (2023) develop non-linear shrinkage techniques to deal with high-dimensionality. Kelly et al. (2019) propose Instrumented PCA (IPCA) to recover latent factors as functions of observable characteristics, effectively condensing high-dimensional information into a smaller factor structure. Lettau and Pelger (2020) design a PCA variant that explicitly targets factors explaining both return covariance and expected returns, further refining factor extraction toward economically relevant components.

In contrast to methods that select or synthesize a small number of factors, we focus on *all-inclusive ridge estimators* that incorporate all candidate factors without ex ante screening, and study how the ridge estimator inherently adapts to the intrinsic complexity. Our results show that even without explicit factor selection, ridge regression effectively shrinks weak components and prioritizes strong ones, thereby achieving self-adaptivity. Furthermore, even in the absence of explicit regularization, the structure of the DGP can give rise to self-induced regularization, maintain a bounded cost of overfitting, and in some cases cause the optimal ridge penalty to be negative.

Parallel to the aforementioned econometric innovations, machine-learning methods have increasingly been applied to asset pricing. Prominent examples include tree-based ensembles and deep neural networks (DNNs) (Gu et al., 2020; Bryzgalova et al., 2019; Chen et al., 2024;

Kelly et al., 2025). Despite their impressive empirical performance, theoretical guarantees remain relatively under-explored, particularly on out-of-sample guarantees and economic interpretability.

1.3.2 Role of Weak Factors in Factor Models

The financial econometrics literature has extensively studied the impact of weak factors in small factor models, where the number of factors remains fixed as the number of observations and/or assets grows. This includes both observable and latent factor settings, and covers tasks such as risk premia estimation and SDF estimation; see, for example, Anatolyev and Mikusheva (2022); Lettau and Pelger (2020); Giglio et al. (2025). In contrast, the role of weak factors in large factor models appears to be theoretically unexplored, particularly in the context of the *all-inclusive* ridge estimator. We also note the recent work by Shen and Xiu (2025), which studies weak signal recovery in a well-specified linear regression model. Our setting differs in that the SDF estimation problem is inherently a misspecified linear regression problem.

1.4 Mathematical Notation

For any positive integers $n \leq m$, let $[n : m]$ denote $\{n, n+1, \ldots, m\}$, and let $[n]$ denote $\{1, \ldots, n\}$. Let $\mathbb{N}$ denote the set of natural numbers excluding 0. Let $\mathbb{R}$ denote the set of real numbers. For any $a \in \mathbb{R}$, let $(a)_+ := \max\{a, 0\}$. For any vector $v \in \mathbb{R}^n$, let $\mathrm{diag}(v)$ denote the $n \times n$ diagonal matrix with diagonal entries $v_1, \ldots, v_n$. For any positive semi-definite matrix A, we define $\|\boldsymbol{\beta}\|_A := \sqrt{\boldsymbol{\beta}^\top A \boldsymbol{\beta}}$, and use $\lambda_i(A)$ to denote its i-th largest eigenvalue.

Given two nonnegative functions $f(x)$ and $g(x)$, we write $f(x) \ll g(x)$ or $g(x) \gg f(x)$ if $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$. We write $f(x) \lesssim g(x)$ or $g(x) \gtrsim f(x)$ if there exists a constant $C > 0$ such that $f(x) \leq C\, g(x)$ for all x. We write $f(x) \asymp g(x)$ if there exist constants $0 < c \leq C$ such that $cg(x) \leq f(x) \leq C\, g(x)$ for all x.

2 Model

2.1 Preliminary: Asset Pricing with Characteristics-Based Factors

We begin by outlining the basic asset pricing framework that underpins characteristics-based factor models, which is extensively studied in recent literature (Kozak et al., 2020; Kozak, 2020; Didisheim et al., 2024; Kelly et al., 2023) and exposited in detail in the recent monograph by Kelly and Xiu (2023). Following the setup in Kozak et al. (2020) and adopting the notation in Kelly and Xiu (2023, Chapter 5), we describe this framework in terms of population moments in this subsection, and defer estimation considerations to Section 2.2. Focusing on the population framework at this stage clarifies the economic structure before introducing statistical complexities.

At each time t, let $\boldsymbol{R}_t$ be the $N \times 1$ vector of excess returns of N portfolios. In typical reduced-form factor models, the stochastic discount factor (SDF) is expressed as a linear

function of portfolio excess returns:

$$M_t = 1 - \boldsymbol{w}_{t-1}' \boldsymbol{R}_t, \tag{1}$$

where $\boldsymbol{w}_{t-1}$ is an $N \times 1$ vector of SDF weights that satisfies the *conditional pricing condition* $\mathbb{E}_{t-1}[M_t \boldsymbol{R}_t] = \boldsymbol{0}$. Characteristics-based models impose structure by linking the SDF weights to observable portfolio characteristics:

$$\boldsymbol{w}_{t-1} = S_{t-1}\boldsymbol{\beta}^*, \tag{2}$$

where S_{t-1} is an $N \times P$ matrix of portfolio characteristics, and $\boldsymbol{\beta}^*$ is a time-invariant $P \times 1$ vector of coefficients. This representation is general, as nonlinearities or time variation can be incorporated through appropriately transformations of raw characteristics, which is particularly effective when the S_{t-1} becomes high-dimensional (Kozak et al., 2020; Kozak, 2020; Kelly et al., 2024; Didisheim et al., 2024). Let $\boldsymbol{F}_t = S_{t-1}^\top \boldsymbol{R}_t$ be the factors constructed from characteristics. We will refer to these "characteristic-managed portfolios" simply as "factors" throughout this paper. Given the structure of (2), the SDF representation (1) can be equivalently rewritten as:

$$M_t = 1 - {\boldsymbol{\beta}^*}^\top \boldsymbol{F}_t, \tag{3}$$

which specifies an unconditional, linear (or affine) relationship between the SDF M_t and the factors $\boldsymbol{F}_t$. This motivates a large body of the characteristic-based asset pricing literature to focus on the *unconditional pricing equation*:

$$\mathbb{E}[(1 - {\boldsymbol{\beta}^*}^\top \boldsymbol{F}_t)\boldsymbol{F}_t] = \boldsymbol{0}. \tag{4}$$

The analysis of this paper centers on (4).

In this paper we always make the following assumption on the data generating process (DGP).

Assumption 1 (Data generating process)**.** Assume that $\boldsymbol{F}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{F}_t] \in \mathbb{R}^P$ is the factor risk premium vector, and $\Sigma = \mathbb{E}[(\boldsymbol{F}_t - \boldsymbol{\mu})(\boldsymbol{F}_t - \boldsymbol{\mu})^\top] \succ 0$ is the factor covariance matrix.

The i.i.d. assumption is standard in the literature. As noted by Kozak et al. (2020), time variation in conditional moments can be captured through dynamics in S_t. The Gaussian assumption is also common, and we leave extensions to non-Gaussian settings to future work. Further assumptions on $\boldsymbol{\mu}$ and Σ will be introduced as needed in later sections.

2.2 SDF Estimation and Ridge Estimator

If the population moments were known, the SDF coefficients $\boldsymbol{\beta}^*$ that satisfy the unconditional pricing condition (4) would be given by solving the population *Maximum Sharpe Ratio Regression* (MSRR) problem (Britten-Jones, 1999; Kelly and Xiu, 2023, Chapter 5 and the references therein):

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^P} \mathbb{E}\left[\left(1 - \boldsymbol{\beta}^\top \boldsymbol{F}\right)^2\right], \tag{5}$$

with the closed-form solution:

$$\boldsymbol{\beta}^* := \left(\mathbb{E}[\boldsymbol{F}\boldsymbol{F}^\top]\right)^{-1}\mathbb{E}[\boldsymbol{F}] = \left(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top\right)^{-1}\boldsymbol{\mu}. \tag{6}$$

In practice, however, the population moments are unknown and must be estimated from data. We now formally define the SDF estimation problem.

2.2.1 The SDF Estimation Problem

Consider a dataset consisting of empirical observations of $\boldsymbol{F}$ over $T \in \mathbb{N}$ time periods: $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_T$. Let $\boldsymbol{F}_t := (F_{1;t}, \ldots, F_{P;t})^\top$, where $F_{i;t}$ denotes the i-th factor's value at time t. Define the $T \times P$ empirical design matrix $\mathbb{F} := [\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_T]^\top$.

The objective is to use the training samples $\mathbb{F}$ to construct an estimate $\hat{\boldsymbol{\beta}} \in \mathbb{R}^P$ of $\boldsymbol{\beta}^*$, such that the estimated SDF $\hat{M}_{T+1} := 1 - \hat{\boldsymbol{\beta}}^\top \boldsymbol{F}_{T+1}$ prices the test-period assets $\boldsymbol{F}_{T+1}$ as accurately as possible.[2] This objective is an *out-of-sample* objective, as $\boldsymbol{F}_{T+1}$ is independent of the training samples.

To make this objective precise, we require a suitable notion of out-of-sample pricing error. Following a large literature (Hansen and Jagannathan, 1997; Hodrick and Zhang, 2001; Kan and Robotti, 2008; Chen and Ludvigson, 2009; Kelly and Xiu, 2023; Didisheim et al., 2024), we adopt the out-of-sample (population) *Hansen-Jagannathan distance* (HJ distance), denoted by $\delta(\cdot)$, as our metric. The squared HJ distance is defined as:

$$\begin{aligned}\delta^2(\hat{\boldsymbol{\beta}}) &:= \left(\mathbb{E}\left[(1 - \hat{\boldsymbol{\beta}}^\top \boldsymbol{F}_{T+1})\boldsymbol{F}_{T+1}\right]\right)^\top \mathbb{E}[\boldsymbol{F}_{T+1}\boldsymbol{F}_{T+1}^\top]^{-1}\mathbb{E}\left[(1 - \hat{\boldsymbol{\beta}}^\top \boldsymbol{F}_{T+1})\boldsymbol{F}_{T+1}\right] \\ &= \hat{\boldsymbol{\beta}}^\top(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top)\hat{\boldsymbol{\beta}} - 2\hat{\boldsymbol{\beta}}^\top\boldsymbol{\mu} + \boldsymbol{\mu}^\top(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top)^{-1}\boldsymbol{\mu}.\end{aligned} \tag{7}$$

The HJ distance $\delta(\cdot)$ has many desirable properties, as discussed in prior work. Most importantly for our purposes, it equates (excess) out-of-sample pricing error with (excess) out-of-sample regression error (as shown in Proposition 1), allowing us to frame the SDF estimation problem as a special case of the well-studied *random-design linear regression problem* with squared loss.

Proposition 1. *For any $\hat{\boldsymbol{\beta}} \in \mathbb{R}^P$, we have*

$$\delta^2(\hat{\boldsymbol{\beta}}) = \delta^2(\hat{\boldsymbol{\beta}}) - \delta^2(\boldsymbol{\beta}^*) = \mathbb{E}\left[(1 - \hat{\boldsymbol{\beta}}^\top \boldsymbol{F})^2\right] - \mathbb{E}\left[(1 - \boldsymbol{\beta}^{*\top} \boldsymbol{F})^2\right] = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2_{\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top}. \tag{8}$$

Remark 1 (Misspecified SDF structure)**.** While we used the linear SDF representation (3) to motivate the estimation of $\boldsymbol{\beta}^*$, note that given the definition of $\boldsymbol{\beta}^*$ in (6), the unconditional pricing condition (4) and Proposition 1 hold regardless of whether the true SDF satisfies (3). None of our theoretical results depend on the correctness of (3). This means that even if the true SDF is nonlinear in $\boldsymbol{F}$, one can still interpret $\boldsymbol{\beta}^*$ as the best linear projection of the true SDF onto the span of $\boldsymbol{F}$. Our bounds on $\delta^2(\hat{\boldsymbol{\beta}})$ remain valid and measure how far $\hat{\boldsymbol{\beta}}$ is from this projection in terms of pricing performance on test assets.

[2] In this paper, the factors in $\boldsymbol{F}$ serve both as the test assets whose returns we aim to explain and as the candidate factors that may enter the SDF. Our results can be extended to the case where the test assets are different from candidate factors.

2.2.2 Ridge Estimator

A natural estimation strategy for obtaining $\hat{\boldsymbol{\beta}}$ is to solve the *empirical* MSRR problem, which is the sample analogue of (5): $\min_{\boldsymbol{\beta}} \hat{\mathbb{E}}_T \left[\left(1 - \boldsymbol{\beta}^\top \boldsymbol{F}\right)^2 \right]$. This approach is well understood and performs effectively in low-dimensional settings where $P \ll T$. However, in high-dimensional settings where $P \gg T$, it is prone to overfitting and may lead to poor out-of-sample performance. This challenge parallels the difficulty of estimating mean-variance efficient (MVE) portfolio weights when the number of assets is large, motivating the use of regularization or alternative estimation methods.

In this paper, we focus on the ridge estimator, a widely used regularization method for high-dimensional estimation in both machine learning (Kobak et al., 2020; Wu and Xu, 2020; Hastie et al., 2022; Tsigler and Bartlett, 2023) and finance (Kozak et al., 2020; Kelly et al., 2024; Didisheim et al., 2024), with economic interpretations provided by Kozak et al. (2020). The *all-inclusive ridge estimator* solves the regularized optimization problem:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{\text{all}}(z) &:= \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^P} \hat{\mathbb{E}}_T \left[(1 - \boldsymbol{\beta}^\top \boldsymbol{F})^2 \right] + z \|\boldsymbol{\beta}\|_2^2 \\ &= \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^P} \|\mathbf{1}_{T\times 1} - \mathbb{F}\boldsymbol{\beta}\|_2^2 + Tz\|\boldsymbol{\beta}\|_2^2. \end{aligned} \tag{9}$$

Lemma 1. *For any $z > 0$, the solution to (9) admits the following equivalent closed-form expressions:*

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{\text{all}}(z) &= (\mathbb{F}^\top \mathbb{F} + Tz \cdot I_{P\times P})^{-1} \mathbb{F}^\top \mathbf{1}_{T\times 1} \\ &= \mathbb{F}^\top (\mathbb{F}\mathbb{F}^\top + Tz \cdot I_{T\times T})^{-1} \mathbf{1}_{T\times 1}. \end{aligned} \tag{10}$$

For any $z \leq 0$, we define the all-inclusive ridge estimator as $\hat{\boldsymbol{\beta}}^{\text{all}}(z) := \mathbb{F}^\top (\mathbb{F}\mathbb{F}^\top + Tz \cdot I_{T\times T})^\dagger \mathbf{1}_{T\times 1}$, where $\dagger$ denotes the Moore-Penrose pseudoinverse. Then $\hat{\boldsymbol{\beta}}^{\text{all}}(z)$ is well-defined for all $z \in \mathbb{R}$ (even when (9) is ill-defined).

The special case $\hat{\boldsymbol{\beta}}^{\text{all}}(0)$ is commonly referred to as the *minimum-norm least squares estimator* in the literature. The motivation for considering negative ridge penalties $z < 0$ is discussed in the following remark.

Remark 2 (Negative ridge)**.** Expression (10) suggests that the negative ridge estimator could remain stable when $z > -\lambda_T(\mathbb{F}\mathbb{F}^\top)/T$; see Section B.1 for details. While the possibility of negative ridge penalties was noted in earlier work (Hua and Gunst, 1983; Björkström and Sundberg, 1999), it remained largely overlooked in the statistics literature. This is likely because the optimal ridge penalty is provably non-negative (Wu and Xu, 2020) in classical underparameterized settings. However, negative ridge has attracted considerable attention in the recent machine learning literature. Contrary to the conventional wisdom that large models require strong regularization to prevent overfitting, Kobak et al. (2020) showed a striking empirical finding that, in overparameterized regimes, the optimal ridge penalty can in fact be negative. This surprising phenomenon has since been rigorously studied in Wu and Xu (2020) and Tsigler and Bartlett (2023), who prove that negative ridge can *outperform* non-negative ridge in certain high-dimensional settings.

3 Theoretical Results: Pricing Error Bounds

We develop a new class of *comparator-based* bounds that provide non-asymptotic guarantees on the out-of-sample pricing error of the ridge estimator $\hat{\boldsymbol{\beta}}^{\text{all}}(z)$. These bounds hold uniformly over all finite sample sizes T, model dimensions P, and arbitrary distributional parameters $\boldsymbol{\mu}$ and Σ. Rather than analyzing performance in isolation, our results compare $\hat{\boldsymbol{\beta}}^{\text{all}}(z)$ to a rich family of economically motivated benchmark estimators based on *principal component regression* (PCR), formulated in the rotated *principal component (PC) factor space*. This comparator framework is strongly inspired by the work of Kozak et al. (2020), who demonstrated the economic interpretability and empirical relevance of PCR-based estimators in high-dimensional characteristic-based factor models. In what follows, we first introduce the PC factor representation in Section 3.1, then define our benchmark estimators in Section 3.2, and finally present our main theoretical results in Section 3.3.

3.1 Moving to the PC Factor Space

Our goal is to derive sharp, non-asymptotic upper bounds on the out-of-sample HJ distance $\delta(\hat{\boldsymbol{\beta}}^{\text{all}}(z))$, as a function of the sample size T, the factor dimension P, and the underlying DGP parameters $\boldsymbol{\mu}$ and Σ. These results yield insight into how model complexity, sample size, and factor structure interact in high-dimensional SDF estimation. A key simplification follows from the observation that, without loss of generality, we can restrict attention to cases where the factor covariance matrix is diagonal. This is achieved via a change of basis to the *principal component (PC) factors*, as introduced by Kozak et al. (2020).

Let $\Sigma = U\Lambda U^\top$ denote the eigendecomposition of Σ, where $\Lambda := \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_P)$ is a diagonal matrix of eigenvalues ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_P > 0$, and $U \in \mathbb{R}^{P\times P}$ is an orthogonal matrix of corresponding eigenvectors.

Definition 1 (Principal component factors (Kozak et al., 2020))**.** *Define the* principal component factors *as* $\widetilde{\boldsymbol{F}} := U^\top \boldsymbol{F}$*. Then* $\widetilde{\boldsymbol{F}} \sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \Lambda)$*, where* $\widetilde{\boldsymbol{\mu}} := U^\top \boldsymbol{\mu}$ *is the risk premium vector of the PC factors, and* Λ *is their covariance matrix.*

This transformation rotates the factor space so that the resulting PC factors are uncorrelated, with variances λ_i and means $\widetilde{\mu}_i$, ordered by decreasing variance. The optimal SDF coefficients in this space, $\widetilde{\boldsymbol{\beta}}^* := U^\top \boldsymbol{\beta}^*$, admit the closed-form expression:

$$\widetilde{\boldsymbol{\beta}}^* = \frac{1}{1 + \sum_{p=1}^{P} \widetilde{\mu}_p^2/\lambda_p} \cdot \left(\frac{\widetilde{\mu}_1}{\lambda_1}, \ldots, \frac{\widetilde{\mu}_P}{\lambda_P}\right),$$

where $\widetilde{\beta}_i^*$ denotes the optimal weight on the i-th PC factor. This representation disentangles the risk-return trade-off in a transparent manner. See Section C for a derivation.

Although PC factors simplify the structure of the estimation problem, they rely on population quantities (e.g., the eigenbasis U) and are therefore not directly observable in practice. In high-dimensional settings where $P \gg T$, estimating low-variance PC factors can be highly unstable, and such errors can propagate to inflate pricing errors. This motivates the truncation approach in Kozak et al. (2020), which eliminates unstable PC factors via principal component analysis (PCA).

Despite the inaccessibility of PC factors in practice, a key observation is that both the ridge-based SDF and its associated HJ distance $\delta^2(\hat{\boldsymbol{\beta}}^{\text{ridge}}(z))$ are *rotation-invariant* with respect to orthogonal transformations of the factor space. Specifically, define

$$\hat{\boldsymbol{\beta}}^{\text{PC(all)}}(z) := \widetilde{\mathbb{F}}^\top(\widetilde{\mathbb{F}}\widetilde{\mathbb{F}}^\top + Tz \cdot I_{T\times T})^\dagger \mathbf{1}_{T\times 1},$$

which is the ridge estimator one would compute if the PC factor realizations $\widetilde{\boldsymbol{F}}_1, \ldots, \widetilde{\boldsymbol{F}}_T$ were observable. Also let $\delta_{\text{PC}}(\cdot)$ denote the corresponding HJ distance evaluated on $\widetilde{\boldsymbol{F}}_{T+1}$. We have the following proposition.

Proposition 2. *For any $z \in \mathbb{R}$, we have $\hat{\boldsymbol{\beta}}^{\text{PC(all)}}(z) = U^\top \hat{\boldsymbol{\beta}}^{\text{all}}(z)$, and hence*

$$1 - \hat{\boldsymbol{\beta}}^{\text{all}}(z)^\top \boldsymbol{F}_{T+1} = 1 - \hat{\boldsymbol{\beta}}^{\text{PC(all)}}(z)^\top \widetilde{\boldsymbol{F}}_{T+1},$$

$$\delta(\hat{\boldsymbol{\beta}}^{\text{ridge}}(z)) = \|\hat{\boldsymbol{\beta}}^{\text{all}}(z) - \boldsymbol{\beta}^*\|_{\Sigma+\boldsymbol{\mu}\boldsymbol{\mu}^\top} = \|\hat{\boldsymbol{\beta}}^{\text{PC(all)}}(z) - \widetilde{\boldsymbol{\beta}}^*\|_{\Lambda+\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top} = \delta_{\text{PC}}(\hat{\boldsymbol{\beta}}^{\text{PC(all)}}(z)).$$

Proposition 2 shows that analyzing the out-of-sample performance of $\hat{\boldsymbol{\beta}}^{\text{all}}(z)$ trained on the original factors $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_T$ and tested on $\boldsymbol{F}_{T+1}$ is equivalent to analyzing the performance of $\hat{\boldsymbol{\beta}}^{\text{PC(all)}}(z)$ trained on the PC factors $\widetilde{\boldsymbol{F}}_1, \ldots, \widetilde{\boldsymbol{F}}_T$ and tested on $\widetilde{\boldsymbol{F}}_{T+1}$. This equivalence enables us to carry out the analysis in the rotated PC factor space, noting that all results translate directly to the original space since the SDF and pricing error remain the same.

3.2 Benchmarks in the PC Factor Space

Rather than relying on the classical bias-variance decomposition (Belkin et al., 2020; Bartlett et al., 2020; Tsigler and Bartlett, 2023), our analysis is based on the *estimation-overfitting decomposition* developed by Bartlett et al. (2021), Shamir (2023), and Lecué and Shang (2024). Translated into the PC factor space for our setting, the core insight is as follows: when ridge regression is applied with $P \gg T$, there exists a fundamental threshold k^* (determined by the DGP and penalty z) that separates the PC factor space into two orthogonal subspaces: An **estimation space** spanned by the top k^* PC factors, where meaningful estimation occurs; and an **overfitting space** spanned by the bottom $P - k^*$ PC factors, where overfitting dominates.

In the estimation space, the leading k^* components of the ridge estimator, $\hat{\boldsymbol{\beta}}^{\text{PC(all)}}_{[1:k^*]}(z)$, are expected to approximate the corresponding entries of the optimal SDF coefficients:

$$\widetilde{\boldsymbol{\beta}}^*_{[1:k^*]} := \frac{1}{1 + \sum_{p=1}^P \widetilde{\mu}_p^2/\lambda_p} \cdot \left(\frac{\widetilde{\mu}_1}{\lambda_1}, \ldots, \frac{\widetilde{\mu}_{k^*}}{\lambda_{k^*}}\right).$$

These components are most critical, as they receive the largest weights in the norm $\|\cdot - \widetilde{\boldsymbol{\beta}}^*\|^2_\Lambda$, which is a major part of the out-of-sample HJ distance.

In contrast, the overfitting part $\hat{\boldsymbol{\beta}}^{\text{PC(all)}}_{[k^*+1:P]}(z)$ is not expected to approximate any meaningful target, not even

$$\widetilde{\boldsymbol{\beta}}^*_{[k^*+1:P]} := \frac{1}{1 + \sum_{p=1}^P \widetilde{\mu}_p^2/\lambda_p} \cdot \left(\frac{\widetilde{\mu}_{k^*+1}}{\lambda_{k^*+1}}, \ldots, \frac{\widetilde{\mu}_P}{\lambda_P}\right).$$

Instead, it primarily serves to (over)fit noise in the training data. Crucially, the existence of k^* is not merely an assumption on the DGP, but rather emerges from the geometry of high-dimensional ridge regression (Lecué and Shang, 2024).

This raises a natural question: why not simply retain the top k^* components and discard the rest? This is precisely the strategy of *principal component regression* (PCR), as explored in Kozak et al. (2020). However, the difficulty lies in the fact that k^* is unknown. Even mild misspecification of k can lead to substantial increases in pricing error (Didisheim et al., 2024).

Nevertheless, PCR estimators offer a useful benchmark. For any $k \in [P]$ and any $z > 0$, define the PCR ridge benchmark:

$$\hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}(z) = \begin{bmatrix} \hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}(z) \\ \\ \hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[k+1:P]}(z) \end{bmatrix} := \begin{bmatrix} \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^k} \hat{\mathbb{E}}_T \left[(1 - \boldsymbol{\beta}^\top \widetilde{\boldsymbol{F}}_{[1:k]})^2 \right] + z\|\boldsymbol{\beta}\|_2^2 \\ \\ \mathbf{0} \end{bmatrix}, \tag{11}$$

which corresponds to solving ridge regression using only the top k PC factors, with zero weights elsewhere.

Remark 3 (Comparison to the PCR estimators in Kozak et al. (2020))**.** While Kozak et al. (2020) estimate the eigenbasis U from data to obtain estimated PC factors, our benchmark estimators (11) assume oracle knowledge of U. As such, the estimators $\hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}(z)$ are *stronger*, in the sense that they are expected to outperform the implementable variants considered in Kozak et al. (2020). Remarkably, as shown in Section 3.3, the performance of the fully data-driven ridge estimator $\hat{\boldsymbol{\beta}}^{\mathrm{all}}(z)$ can be tightly bounded relative to our family of *idealized* PCR estimators even without access to U or knowledge of k^*.

3.3 Main Results: Non-Asymptotic Comparator-Based Bounds

In this subsection we present our main theoretical results. To proceed, we first introduce an economically motivated assumption (Assumption 2) and a definition commonly used in high-dimensional statistics (Definition 2).

Assumption 2 (No asymptotic arbitrage)**.** There exists a universal constant C_{NA} such that

$$\sum_{i=1}^{P} \frac{\widetilde{\mu}_i^2}{\lambda_i} \leq C_{\mathrm{NA}}.$$

The rationale of Assumption 2 is as follows. Suppose Assumption 2 is violated and $\lim_{P\to\infty} \sum_{i=1}^{P} \frac{\widetilde{\mu}_i^2}{\lambda_i} = \infty$, then for any P, consider the portfolio given by the PC factor weights

$$\frac{1}{\left(\sum_{i=1}^{P} \frac{\widetilde{\mu}_i^2}{\lambda_i}\right)} \left(\frac{\widetilde{\mu}_1}{\lambda_1}, \ldots, \frac{\widetilde{\mu}_P}{\lambda_P} \right).$$

This portfolio has variance being $\left(\sum_{i=1}^{P} \frac{\widetilde{\mu}_i^2}{\lambda_i}\right)^{-1}$ and risk premium being 1. As $P \to \infty$, we obtain a sequence of *asymptotic arbitrage* portfolios whose variance $\to 0$ and risk premium being 1. Therefore, in order to preclude asymptotic arbitrage, it is necessary that Assumption 2 holds.

Lemma 2. *Under Assumption 2, $\|\cdot\|^2_{\Lambda+\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top} \le (1+C_{\mathrm{NA}})\|\cdot\|^2_{\Lambda}$. Hence for any benchmark $\boldsymbol{\beta} \in \mathbb{R}^P$, for any $k \in [P]$,*

$$\begin{aligned}&\left|\delta_{\mathrm{PC}}(\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}(z)) - \delta_{\mathrm{PC}}(\boldsymbol{\beta})\right| \\ &\le \sqrt{1+C_{\mathrm{NA}}}\left(\left\|\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}_{[1:k]}(z) - \boldsymbol{\beta}_{[1:k]}\right\|_{\Lambda_{[1:k]}} + \left\|\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}_{[k+1:P]}(z) - \boldsymbol{\beta}_{[k+1:P]}\right\|_{\Lambda_{[k+1:P]}}\right)\end{aligned}$$

where $\Lambda_{[1:k]} := \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ and $\Lambda_{[k+1:P]} := \mathrm{diag}(\lambda_{k+1}, \ldots, \lambda_P)$.

Lemma 2 highlights the role of Assumption 2 in our analysis. It allows us to compare the HJ distances of two estimators by decomposing the comparison into two orthogonal subspaces. Intuitively, the norm $\|\cdot\|_{\Lambda_{[1:k]}}$ captures the "projection" of the HJ distance difference onto the "top k" PC factor space, while $\|\cdot\|_{\Lambda_{[k+1:P]}}$ reflects the "projection" onto the "bottom $P-k$" PC factor space. Since these spaces can be extremely high-dimensional when $P \gg T$, we introduce the following definition to quantify the effective (or intrinsic) dimension.

Definition 2 (Effective rank (Bartlett et al., 2020; Lecué and Shang, 2024))**.** *For any $p \times p$ covariance matrix $A \succ 0$, we define its* effective rank *as $r(A) := \frac{\sum_{i=1}^p \lambda_i(A)}{\lambda_1(A)}$.*

The effective rank coincides with the true rank when all eigenvalues are equal, and decreases as the spectrum becomes more uneven, i.e., when a few eigenvalues dominate. Intuitively, it captures how many directions "effectively matter" in a high-dimensional distribution. A larger effective rank indicates that the variance is more evenly distributed across many directions, whereas a smaller value reflects concentration along a few dominant components.

We are now ready to state our main theorem.

Theorem 1. *Suppose Assumption 1 and Assumption 2 hold. There exists universal constants $C_0, C_1, c > 0$, such that for any $T \in \mathbb{N}, P \in \mathbb{N}, \boldsymbol{\mu} \in \mathbb{R}^P, \Sigma \succ 0$, we have:*

For any $k \le C_0 T$ such that $r(\Lambda_{[k+1,P]}) \ge C_1 T$, for any $z > -\frac{1}{T}\sum_{j=k+1}^P \lambda_j$, with probability at least $1 - \exp(-cT)/c$:

(1) In the "top k" PC factor space,

$$\left\|\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}_{[1:k]}(z) - \hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}\left(z + \frac{1}{T}\sum_{j=k+1}^P \lambda_j\right)\right\|_{\Lambda_{[1:k]}} \lesssim \varepsilon_\Sigma(z,k) \tag{12}$$

(2) In the "bottom $P-k$" PC factor space,

$$\left\|\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}_{[k+1:P]}(z) - \mathbf{0}\right\|_{\Lambda_{[k+1:P]}} \lesssim \varepsilon_\Sigma(z,k) \tag{13}$$

(3) Overall,

$$\left|\delta_{\mathrm{PC}}(\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}(z)) - \delta_{\mathrm{PC}}\left(\hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}\left(z + \frac{1}{T}\sum_{j=k+1}^P \lambda_j\right)\right)\right| \lesssim \varepsilon_\Sigma(z,k) \tag{14}$$

Here $\lesssim$ suppress universal constants that only depend on C_{NA}, and

$$\varepsilon_\Sigma(z,k) := \frac{\lambda_{k+1} + \sqrt{\frac{1}{T}\sum_{j=k+1}^P \lambda_j^2}}{z + \frac{1}{T}\sum_{j=k+1}^P \lambda_j} \quad \overset{(if\ z > -\frac{1}{2T}\sum_{j=k+1}^P \lambda_j)}{\lesssim} \quad \frac{1}{\sqrt{r(\Lambda_{k+1,P})/T}}.$$

The same bounds apply to any $k > C_0 T$ such that $r(\Lambda_{[k+1,P]}) \geq C_1 T$, with the additional condition that $z + \frac{1}{T}\sum_{j=k+1}^P \lambda_j \gtrsim \lambda_1$.

Theorem 1 shows that the pricing error of the all-inclusive ridge estimator $\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}(z)$ will converge to the pricing error of a PCR estimator $\hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}\left(z + \frac{1}{T}\sum_{j=k+1}^P \lambda_j\right)$, with a convergence rate no slower than $1/\sqrt{r(\Lambda_{k+1,P})/T}$. The convergence *simultaneously* apply to all $k \in [P]$ such that $r(\Lambda_{k+1,P}) \gtrsim T$, with possibly different convergence rates for different k. If $k \lesssim T$, then the requirement on z is very weak and allows negative ridge; if $k \gtrsim T$, then the requirement on z becomes stronger.[3]

3.4 Interpretations of the Main Results

In this subsection, we examine the implications of Theorem 1 in detail and derive three core insights of the paper: self-induced regularization, bounded cost of overfitting, and self-adaptive property. To streamline the discussion, we introduce the following structural assumption on Σ, which is essentially an assumption on the DGP.

Assumption 3 (Low-dimensional estimation space, high-dimensional overfitting space)**.** As T, P grow (without any restriction on the growth rate of P relative to T), there exists a cutoff $k^* \leq C_0 T$ (which may grow with T, but not faster than T) such that the bottom $P - k^*$ PC factors have a covariance matrix with high effective rank:

$$r(\Lambda_{[k^*+1:P]}) \gg T.$$

Under Assumption 3, we refer to the space spanned by the top k^* PC factors as the *estimation space*. This space is low-dimensional in the sense that $k^* \lesssim T$, which allows for meaningful statistical estimation. In contrast, we refer to the space spanned by the bottom $P - k^*$ PC factors as the *overfitting space*. This space is intrinsically high-dimensional as dictated by the condition $r(\Lambda_{[k^*+1:P]}) \gg T$, making reliable estimation infeasible.

Assumption 3 is very general and is satisfied by the covariance matrices of many economically relevant DGPs. We illustrate its scope with four canonical examples:

> **Reciprocal decay.** If $\lambda_i \asymp 1/i$ for $i \in [P]$, then $r\big(\Lambda_{[k+1:P]}\big) \asymp k\log(P/k)$. Choosing $k^* \asymp T$ therefore meets Assumption 3 when $P \gg T$.
>
> **Sub-reciprocal decay.** Let $\lambda_i \asymp 1/i^\alpha$ with $\alpha \in (0,1)$. Then $r\big(\Lambda_{[k+1:P]}\big) \asymp k(P/k)^{1-\alpha}$, so a wide range of k^* satisfies Assumption 3. For example, when $\alpha = 1/2$ one may take any $k^* \leq C_0 T$ that also obeys $k^* \gg T^2/P$; if, in addition, $P \gg T^2$, every $k^* \leq C_0 T$ works. We note that $\alpha = 1/2$ matches the empirical RFF calibration in Didisheim et al. (2024).

[3] See Section 1.4 for definitions of $\asymp, \lesssim, \gtrsim, \ll, \gg$.

Equal-scale low-variance PC factors. Suppose there exists $k^* \leq C_0 T$ such that $\lambda_1, \ldots, \lambda_{k^*}$ are arbitrary while $\lambda_{k^*+1} \asymp \cdots \asymp \lambda_P \asymp \underline{\lambda}$. Then $r\big(\Lambda_{[k^*+1:P]}\big) \asymp P - k^*$, so any $k^* \leq C_0 T$ suffices when $P \gg T$, irrespective of the precise scale of $\underline{\lambda}$.

Classical small-factor model. Letting $\underline{\lambda} \to 0$ in the previous example recovers the classical, under-parameterized factor models (Ross, 1976; Fama and French, 1993, 2015). Here k^* (typically a small constant) counts the dominant factors, and making $\underline{\lambda}$ arbitrarily small renders the P-factor model arbitrarily close to a k^*-factor one.

Under Assumption 3, we can sharpen the interpretations of Theorem 1.

Corollary 1 (What large factor models learn)**.** *Suppose Assumption 1, Assumption 2, and Assumption 3 hold. There exists a universal constant $c > 0$ such that for any $T \in \mathbb{N}, P \in \mathbb{N}, \boldsymbol{\mu} \in \mathbb{R}^P, \Sigma \succ 0$, and for any $z > -\frac{1}{2T}\sum_{j=k^*+1}^{P} \lambda_j$, with probability at least $1 - \exp(-cT)/c$:*

(1) In the estimation space *spanned by the top k^* PC factors, the all-inclusive ridge estimator behaves like a PCR estimator using only the top k^* PC factors, but with an effective regularization parameter $z + \frac{1}{T}\sum_{j=k^*+1}^{P} \lambda_j$ larger than z:*

$$\left\| \hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}_{[1:k^*]}(z) - \hat{\boldsymbol{\beta}}^{\mathrm{PC}(k^*)}_{[1:k^*]}\left(z + \frac{1}{T}\sum_{j=k^*+1}^{P} \lambda_j\right) \right\|_{\Lambda_{[1:k^*]}} \lesssim \frac{1}{\sqrt{r(\Lambda_{[k^*+1:P]})/T}} \to 0.$$

*This additional penalty arises from the influence of low-variance PC factors and is termed **self-induced regularization**.*

(2) In the overfitting space *spanned by the bottom $P - k^*$ PC factors, the ridge estimator's out-of-sample behavior is close to that of the zero estimator:*

$$\left\| \hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}_{[k^*+1:P]}(z) - \mathbf{0} \right\|_{\Lambda_{[k^*+1:P]}} \lesssim \frac{1}{\sqrt{r(\Lambda_{[k^*+1:P]})/T}} \to 0.$$

*Although the estimator interpolates noise and perfectly fits the training data, its out-of-sample pricing performance is similar to simply discarding these components. The fact that **cost of overfitting** $\approx$ **cost of under-specification** is quite surprising; see more discussions at the end of this subsection.*

(3) Overall, the all-inclusive ridge estimator approximates a PCR estimator that selects only the top k^* PCs and applies a self-induced regularization:

$$\left| \delta_{\mathrm{PC}}(\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}(z)) - \delta_{\mathrm{PC}}\left(\hat{\boldsymbol{\beta}}^{\mathrm{PC}(k^*)}\left(z + \frac{1}{T}\sum_{j=k^*+1}^{P} \lambda_j\right)\right) \right| \lesssim \frac{1}{\sqrt{r(\Lambda_{[k^*+1:P]})/T}} \to 0.$$

Crucially, the all-inclusive ridge estimator achieves this without *knowledge of k^* or explicit PC factor construction, demonstrating its **self-adaptive property**. Moreover, it implicitly competes with all feasible k^*, including the best-performing one.*

Corollary 1 establishes three fundamental insights. First of all, there exists self-induced regularization in the estimation space. During training, the overfitting space absorbs noise, which indirectly leads to shrinkage of the estimated coefficients in the estimation space, as part of the signal may be absorbed as well. This mimics an increase in the effective ridge penalty. Consistent with the double descent phenomenon (Belkin et al., 2019), such implicit regularization can benefit unpenalized estimators like the min-norm least squares estimator. However, it can be detrimental when explicit regularization is already optimized, as it restricts the effective tuning range and introduces additional bias.

In addition, Corollary 1 points to the bounded cost of overfitting in the overfitting space. In the low-variance PC factor space, the ridge estimator fits noise and loses all signal, so the cost is real. However, this cost is no worse than that of under-specification (i.e., discarding the entire overfitting space). As shown in Tsigler and Bartlett (2023) and Lecué and Shang (2024), when the effective rank is high, training and test data in the overfitting space are nearly orthogonal. As a result, the interpolated coefficients have negligible projection onto test data, although they are nonzero and achieving perfect in-sample fit. This geometric decoupling ensures that overfitting does not substantially amplify the out-of-sample error and effectively behaves like the zero estimator.

Furthermore, Corollary 1 implies certain self-adaptivity. Although the ridge estimator does not explicitly perform PCA or know the best cutoff dimension k^*, it effectively prices test assets as if it retained only the top k^* PC factors while suppressing the rest. This leads to performance close to that of a well-tuned PCR estimator with the correct subspace and certain regularization, though the regularization is not necessarily optimal. The ability to adaptively focus on valuable components without explicit dimension selection constitutes a form of self-adaptivity.

4 Applications of Pricing Error Bounds

We now apply the pricing error bounds derived in Section 3.3 to address the research questions posed in Section 1. For rigor and clarity, all discussions in this section assume Assumption 1, Assumption 2, and Assumption 3.

4.1 Role of Noise Factors

We consider the setting in Section 3.4, with two additional assumptions on the rotated risk premium vector $\widetilde{\boldsymbol{\mu}}$:

(i) **Priced top-k^* factors:** The top k^* high-variance PC factors carry strictly positive risk premia.

(ii) **Unpriced noise factors:** The bottom $P - k^*$ PC factors are noise factors, i.e.,

$$\widetilde{\mu}_{k^*+1} = \cdots = \widetilde{\mu}_P = 0.$$

Now consider a model with $K = k^*$ (i.e., a well-specified model containing all priced components). To assess the impact of including noise factors, we compare the performance of

the K-factor model and the all-inclusive P-factor model, which is precisely the comparison addressed by Corollary 1. For any $z > -\frac{1}{T}\sum_{j=k^*+1}^{P}\lambda_j$, the z-penalized ridge estimator in the P-factor model is approximately equivalent to a $(z+\frac{1}{T}\sum_{j=k^*+1}^{P}\lambda_j)$-penalized ridge estimator in the K-factor model in terms of pricing error.

This leads to some notable implications for the role of noise factors. Most prominently, noise factors restrict the tuning range. The effective tuning range of z in the K-factor model is $\mathbb{R}_{\geq 0}$, whereas in the P-factor model, it becomes $\left[\frac{1}{T}\sum_{j=k^*+1}^{P}\lambda_j, \infty\right)$ which is a strictly more restrictive interval due to self-induced regularization. Since all other aspects remain approximately the same, this narrowing of the tuning range directly implies that adding noise factors strictly worsens model performance in our setting.

Also, the extent of harm caused by noise factors grows with their aggregate variance. Noise factors are especially damaging when $\frac{1}{T}\sum_{j=k^*+1}^{P}\lambda_j$ is large. They are asymptotically benign if this quantity vanishes as $P \to \infty$.

The result also implicates a surprising origin of performance loss. While it may seem intuitive that noise factors hurt performance, our results clarify the mechanism. In the current noise factor setting, the cost of under-specification is zero (noise factors carry no risk premium), hence the cost of overfitting is also approximately zero. Therefore, the harm does not stem from the overfitting space. The harm arises from self-induced regularization in the estimation space, which biases the estimated coefficients on the top k^* PC factors and reduces the attainable risk premium.

Our analysis further highlights that self-adaptivity provides value when $K < k^*$. Under-specifying the model ($K < k^*$) leads to unrecoverable loss in pricing error and Sharpe ratio, which persists even with infinite samples. In this context, self-adaptation via the all-inclusive model offers a valuable tradeoff, in particular when $K \ll k^*$ and the aggregate variance of noise factors is small.

4.2 Role of Weak Factors

Next, consider the setting in Section 3.4 but assume

$$\widetilde{\mu}_{k^*+1} \geq \cdots \geq \widetilde{\mu}_P > 0,$$

meaning that the bottom $P - k^*$ PC factors are weakly priced. While each may carry a small risk premium, their aggregate contribution could be significant.

Since Corollary 1 makes no assumption on $\widetilde{\boldsymbol{\mu}}$, all insights from the noise factor setting apply but with one difference: the cost of overfitting is now non-negligible, as true signals are lost despite interpolation. Nonetheless, the principle "in the overfitting space, cost of overfitting $\approx$ cost of under-specification" still holds.

In addition to the implications from Section 4.1, further takeaways emerge regarding the behavior of all-inclusive model. First, all-inclusive model fails to recover weak signals. In our setting, even when all factors are priced, the P-factor model performs worse than the $K = k^*$ model. This is due to Corollary 1 part (2): although interpolating the in-sample data, the estimated coefficients on the bottom $P - k^*$ PC factors are approximately equivalent to a zero estimator, in terms of the out-of-sample pricing performance. Second, benefit lies in self-adaptation, not in signal recovery. Identifying k^* ex ante is difficult. The strength of the

all-inclusive model lies in its ability to automatically prioritize valuable components, not in recovering Sharpe ratio from the bottom $P - k^*$ PC factors.

4.3 Negative Ridge

In Section 2.2, we define the ridge estimator $\hat{\boldsymbol{\beta}}^{\text{all}}(z)$ for *negative ridge penalties* $z < 0$, following earlier works by Hua and Gunst (1983); Björkström and Sundberg (1999); Kobak et al. (2020); Wu and Xu (2020); Tsigler and Bartlett (2023). Further interpretations of negative ridge, along with numerical considerations, are provided in Section B.1.

Negative ridge may seem counterintuitive at first, but it emerges naturally in our setting. The phenomenon of *self-induced regularization*, in which noise and weak factors introduce additional regularization and bias in the estimation space, motivates the use of negative penalties to counteract this implicit bias. Specifically, can a negative ridge penalty *correct* the bias introduced by overparametrization and thereby improve estimation?

This idea becomes particularly relevant when considering the optimal penalty z^* for the PCR estimator. If z^* falls below the level $\frac{1}{T}\sum_{j=k^*+1}^{P}\lambda_j$, then achieving the appropriate shrinkage for the full P-factor model requires a *negative ridge penalty*. Our Corollary 1 supports this possibility, consistent with recent theoretical results (Wu and Xu, 2020; Tsigler and Bartlett, 2023) showing that negative ridge can outperform standard non-negative ridge in certain high-dimensional regimes. More precisely, under Assumption 1, Assumption 2, and Assumption 3, if we aim for the ridge estimator $\hat{\boldsymbol{\beta}}^{\text{all}}(z)$ to compete with $\hat{\boldsymbol{\beta}}^{\text{PC}(k^*)}(z^*)$, the appropriate choice of z is around

$$z^* - \frac{1}{T}\sum_{j=k^*+1}^{P}\lambda_j, \tag{15}$$

which can be negative if the cumulative variance of low-variance PC factors is large. Moreover, as P grows, the need for negative penalties becomes even more pronounced.

However, negative ridge can only *partially* offset the effects of self-induced regularization. Its corrective power has inherent limits: once z becomes too negative, both *statistical* and *numerical instabilities* may arise.

From a statistical standpoint, although the scale of self-induced regularization in Corollary 1 is $\frac{1}{T}\sum_{j=k^*+1}^{P}\lambda_j$, the penalty z must satisfy

$$z > -\frac{1}{2T}\sum_{j=k^*+1}^{P}\lambda_j$$

in Corollary 1 to guarantee controlled out-of-sample behavior.[4] This implies that the effective regularization remains strictly positive at $\frac{1}{2T}\sum_{j=k^*+1}^{P}\lambda_j > 0$. In cases where $z^* < \frac{1}{2T}\sum_{j=k^*+1}^{P}\lambda_j$, negative ridge alone cannot fully replicate the ideal penalty level.

Why is this stricter threshold necessary, compared to the broader condition $z > -\frac{1}{T}\sum_{j=k+1}^{P}\lambda_j$ in Theorem 1? The key reason is that the error term $\varepsilon_\Sigma(k, z)$ decreases *strictly* as z increases

[4] Replacing $-\frac{1}{2T}\sum_{j=k^*+1}^{P}\lambda_j$ by $-\frac{1}{\gamma T}\sum_{j=k^*+1}^{P}\lambda_j$ for any fixed $\gamma > 1$ would yield the same convergence rate.

within $\left(-\frac{1}{T}\sum_{j=k+1}^{P}\lambda_j, \infty\right)$. Pushing z too close to $-\frac{1}{T}\sum_{j=k+1}^{P}\lambda_j$ causes $\varepsilon_\Sigma(k,z)$ to *diverge*, undermining the convergence rate. Empirical results in Section 6 corroborate this insight: negative ridge can improve SDF estimation quite significantly, but only up to a point. Beyond that point, instability behavior sets in.

From a numerical standpoint, computing $\hat{\boldsymbol{\beta}}^{\text{all}}(z)$ requires inverting the $T \times T$ matrix $\mathbb{F}\mathbb{F}^\top + Tz \cdot I_{T\times T}$. If z becomes highly negative, the *condition number* of this matrix deteriorates, rendering the numerical solution sensitive to small perturbations. This degradation can cause the numerically computed $\hat{\boldsymbol{\beta}}^{\text{all}}(z)$ to deviate from its theoretical statistical guarantee. We explain this issue further in Section B.1. As a practical guideline, we recommend selecting z sufficiently larger than $-\lambda_T(\mathbb{F}\mathbb{F}^\top)/T$, ensuring numerical stability. Encouragingly, Lemma 4 implies that this numerical threshold is of the same order as the statistical threshold $z > -\frac{1}{2T}\sum_{j=k^*+1}^{P}\lambda_j$ under mild conditions.

5 Simulations

We use three Monte Carlo simulations as analytically tractable examples to illustrate each of the three points made in Section 4. For simplicity, all simulations are implemented directly in the PC space. In all simulations we draw T i.i.d. Gaussian samples to construct a $T \times P$ matrix $\widetilde{\mathbb{F}}$ with diagonal covariance $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_P)$, consistent with Assumption 1 and Definition 1. An oracle cutoff k^* separates the estimation space and the overfitting space. We set $T = 200$, $k^* = 20$, and number of Monte Carlo replications to be 300 unless otherwise stated. Pricing performance is evaluated using the squared HJ distance δ^2 as in (7).

5.1 Simulation for Noise Factors

To match the noise factor design in Section 4.1, the leading PCs are assigned unit variance

$$\lambda_i = 1, \quad i \le k^*$$

while the remaining PCs follow a sub-reciprocal tail,

$$\lambda_{k^*+m} = 0.1\, m^{-0.2}, \quad m \in \mathbb{N}.$$

For the risk premia, we set $\widetilde{\mu}_i = \sqrt{\lambda_i/k^*}$ for $i \le k^*$ and $\widetilde{\mu}_i = 0$ for $i > k^*$. This design satisfies Assumption 2 (no asymptotic arbitrage) in a trivial way since

$$\lim_{P\to\infty}\sum_{i=1}^{P}\widetilde{\mu}_i^2/\lambda_i = \sum_{i=1}^{k^*}1/k^* = 1.$$

We calibrate a baseline penalty z_0 that approximately minimizes the baseline top-k^* error, then hold z_0 fixed while varying P. Using the calibrated z_0 is not essential. Any fixed z value in an appropriate range suffices for the qualitative comparisons here. We run 300 replications and report the mean squared HJ distance.

Figure 1 compares the all-inclusive ridge estimator at the fixed penalty z_0 to the "shifted" top-k^* proxy from Corollary 1, which uses only the first k^* PCs but increases the penalty

by the empirical shift term $\frac{1}{T}\sum_{j=k^*+1}^{P}\lambda_j$. The close agreement between the circle-marked and square-marked curves supports the self-induced regularization mechanism. When adding unpriced factors, they mainly acts as an effective increase in the penalty applied to the leading PCs. The no-shift baseline is flat, consistent with the fact that the added factors have zero risk premia.

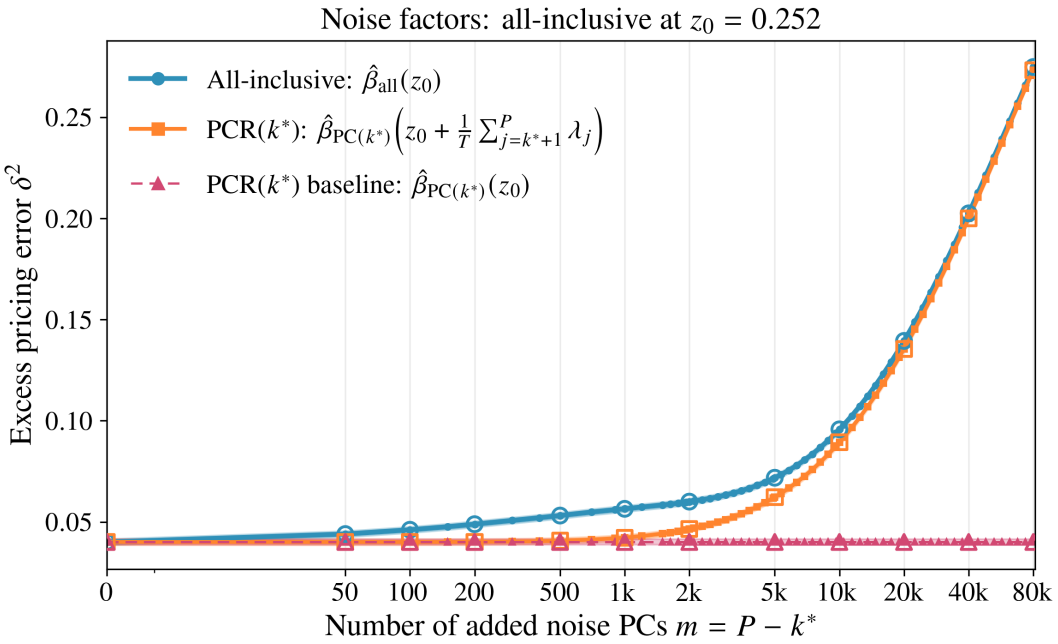

Figure 1. Simulation for the Noise Factors Case

5.2 Simulation for Weak Factors

To examine the role of weak factors discussed in Section 4.2, we repeat the same Λ design but assigns strictly positive, yet decaying, risk premia to the bottom PCs. Concretely, we parameterize risk premia through per-PC Sharpe ratios $s_i = \widetilde{\mu}_i/\sqrt{\lambda_i}$. To ensure the design satisfies Assumption 2, we take $s_i = 1/\sqrt{k^*}$ for $i \le k^*$ and

$$s_{k^*+m} = 0.8\,(1/\sqrt{k^*})\,m^{-0.55}, \quad m \in \mathbb{N}.$$

This gives

$$\lim_{P\to\infty}\sum_{i=1}^{P} s_i^2 = 1 + \frac{0.8^2}{k^*}\sum_{m=1}^{\infty} m^{-2\times 0.55}.$$

This is why we choose the decay exponent 0.55. Since $2 \times 0.55 > 1$, the series $\sum_{m=1}^{\infty} m^{-2\times 0.55}$ converges and Assumption 2 holds.

As in Figure 1, we compare the all-inclusive ridge estimator with the penalty-shift proxy in Figure 2. Consistent with Section 4.2, the all-inclusive estimator behaves largely like a top-k^* PCR with a penalty shift, so weak premia in low-variance directions are difficult to recover in such a setting even though they are priced. Notably, the top-k^* PCR baseline that omits weak signals in Figure 2 is no longer flat, which is exactly what we would expect when the priced tail factors are omitted.

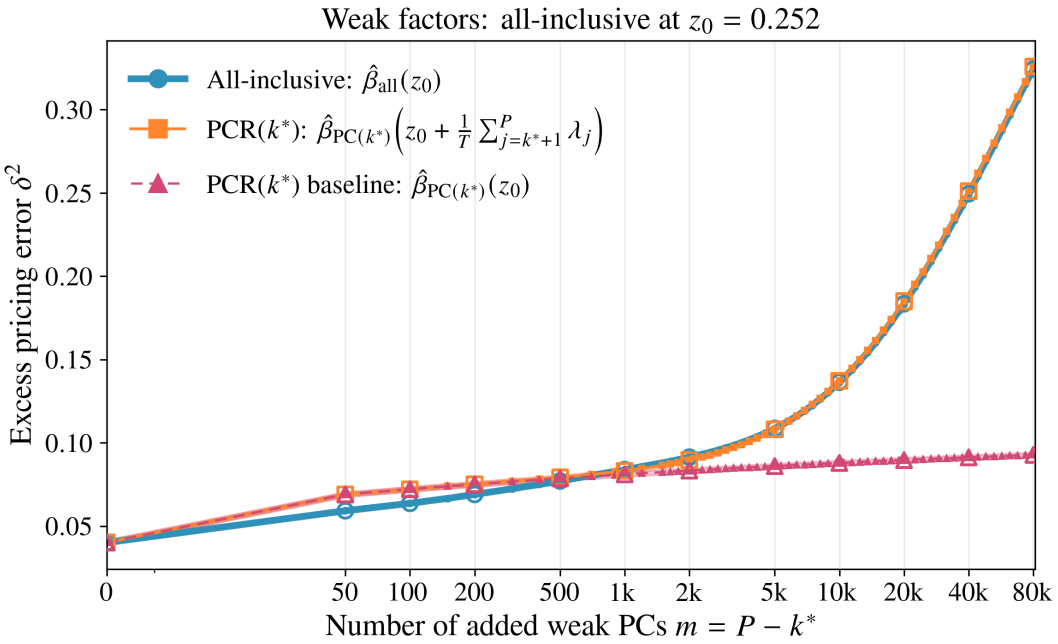

Figure 2. Simulation for the Weak Factors Case

5.3 Simulation for Negative Ridge

We use a separate design to show when the optimal ridge penalty becomes negative. We set $P \in \{10T, 25T, 50T, 100T\}$ and consider two eigenvalue regimes: a reciprocal spectrum $\lambda_j = j^{-1}$, and a heavier sub-reciprocal spectrum $\lambda_j = j^{-\alpha}$ with $\alpha = 0.5$.

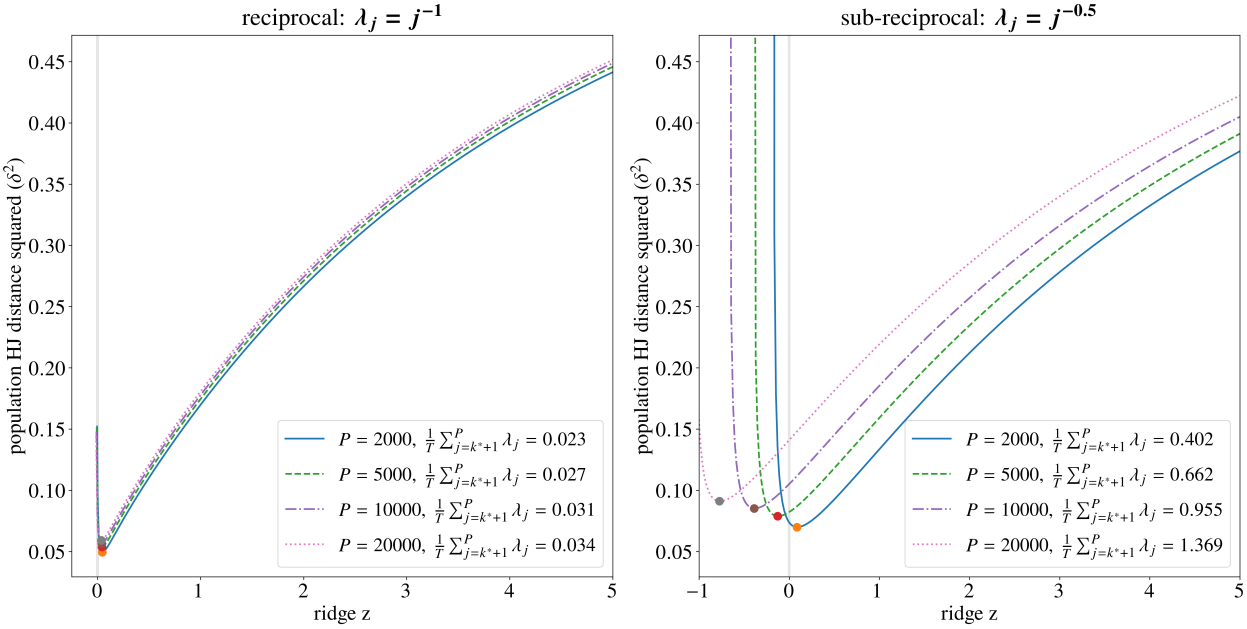

Figure 3. Simulation for Negative Ridge

For each P, we compute the mean $\delta^2(z)$ over replications on a grid that extends into

negative values down to

$$z_{\min} = -\kappa \cdot \frac{1}{T} \sum_{j=k^*+1}^{P} \lambda_j$$

with $\kappa = 0.9$ to avoid numerical instability. For risk premia, we adopt a design analogous to the weak factor simulation to ensure Assumption 2 holds. Specifically, we set $\widetilde{\mu}_i = \sqrt{\lambda_i} \cdot i^{-0.55}$ for $i \leq k^*$ and $\widetilde{\mu}_i = (\sqrt{\lambda_i}/2) \cdot i^{-0.55}$ for $i > k^*$r. We note that using a noise factor design here yields similar results.

Figure 3 plots $\delta^2(z)$ and marks the minimizer z^*. In the sub-reciprocal simulations, the self-induced shift grows with P and z^* becomes negative for sufficiently large P, suggesting the role of negative ridge as a partial correction for self-induced regularization as we discussed in Section 4.3. Besides, the ridgeless case (i.e., $z = 0$) seems to work reasonably well under reciprocal decay and in some sub-reciprocal cases.

6 Empirical Results

6.1 Factor Construction

In our empirical experiments, we closely follow the framework developed in Didisheim et al. (2024). To transform a limited set of stock characteristics into an arbitrarily large number of factors, we employ Random Fourier Features (RFF), as in Didisheim et al. (2024), which builds on earlier work by Kelly et al. (2024). Although RFF represents only one possible architecture for factor construction, we focus on RFF in this paper to ensure consistent benchmarking against existing results.

Originally introduced by Rahimi and Recht (2007), RFF provides an explicit feature mapping to approximate shift-invariant kernels (e.g., the RBF/Gaussian kernel) using a finite number of random features. Conceptually, RFF can be viewed as a shallow neural network with a single hidden layer, potentially with very large width (when P is large), where the weights feeding into the hidden layer are randomly drawn rather than learned through optimization.

In this paper, we adopt the same RFF structure as in Didisheim et al. (2024), as specified in equation (16).

$$S_{t-1} = \left\{ \begin{bmatrix} \cos(\gamma_p Z_{t-1} \xi_p) \\ \sin(\gamma_p Z_{t-1} \xi_p) \end{bmatrix} \right\}_{p=1}^{P/2} \tag{16}$$

where γ is a scaler; Z_{t-1} is the stock-characteristic at time t; and ξ_p are i.i.d. drawn from the standard normal distribution $\mathcal{N}(0, 1)$. Denote N_{t-1} the total number of stocks we have in data at time $t-1$, then S_{t-1} is an $N_{t-1} \times P$ matrix with a sin part and a cos part vertically stacked.

6.2 Empirical Estimation

Consistent with previous notations in Section 2.1, once we have S_{t-1} we are ready to get the "factors" by $\boldsymbol{F}_t = (N_{t-1}^{-1/2}) S_{t-1}^\top \boldsymbol{R}_t$, here scaled by $N_{t-1}^{-1/2}$ for adjusting with the difference

of number of stocks vary across sample period. Our empirical design matrix, as defined in Section 2.2, would be $\mathbb{F} := [\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_T]^\top$. We set the estimation sample size $T = 360$ (i.e., 360 months rolling window) unless otherwise specified, again this is identical to Didisheim et al. (2024). Plug this $P \times T$ matrix $\mathbb{F}$ into the closed-form solution (6), we get the empirical estimation $\hat{\boldsymbol{\beta}}^{\text{all}}(z)$ for ridge regression with regularization parameter z.

Following the empirical setup in Didisheim et al. (2024), we obtained the U.S. 1963-2024 stock-month data used in Jensen et al. (2023) from their WRDS repository. Detailed documentation of this dataset can be found on their website[5]. The stocks include NYSE/AMEX/NASDAQ securities that have CRSP share code in 10, 11, or 12. "Nano-cap" stocks are excluded. There are in total 153 characteristics for the stocks. Out of the 153 characteristics, only the least-missing 130 characteristics in the sample period are selected. After fixing the list of 130 characteristics, we drop any stock-month observations that have $\geq 30\%$ missing values (i.e., ≥ 39 characteristics). Next, each of the 130 characteristics are standardized to the $[-0.5, +0.5]$ interval by ranking of the value. Lastly, missing values are filled by zeros. This is the Z_t we used for RFF input.

6.3 Evaluation Metrics

For out-of-sample evaluation, we report two metrics: the *squared* HJ distance, which quantifies pricing error, and the Sharpe ratio of the factors' tangency portfolio. We use the squared HJ distance rather than the standard (non-squared) version, with the purpose to align with the empirical metrics in Didisheim et al. (2024) and to ensure consistency in comparison. Following a rolling window approach, we estimate the SDF using past data and evaluate its performance over the subsequent one-month period, ensuring a genuinely *out-of-sample* assessment.

Consistent with the setup in Section 2.2, the estimated SDF in each test period takes the form

$$\hat{M}_{\text{test}}(z) := 1 - \hat{\boldsymbol{\beta}}^{\text{all}}(z)^\top \boldsymbol{F}_{\text{test}}.$$

By construction, the HJ distance is unit-free, while Sharpe ratios are annualized by scaling the monthly Sharpe ratios by $\sqrt{12}$. Because the one-month-ahead excess returns are available in the Jensen et al. (2023) dataset through November 2024, we obtain a total of $L = 383$ rolling estimation-evaluation rounds. For each evaluation metric, we report the average across these 383 rounds. For HJ distance evaluation, we use as test assets the U.S. monthly data of 153 anomaly factor-mimicking portfolios (capped value-weighted) constructed by Jensen et al. (2023), also downloaded from the same source. The empirical formula we use to estimate the out-of-sample squared HJ distance is

$$\hat{\mathbb{E}}_L[\hat{M}_{\text{test}}(z)\boldsymbol{F}_{\text{test}}]^\top \left(\hat{\mathbb{E}}_L[\boldsymbol{F}_{\text{test}}\boldsymbol{F}_{\text{test}}^\top]\right)^\dagger \hat{\mathbb{E}}_L[\hat{M}_{\text{test}}(z)\boldsymbol{F}_{\text{test}}], \tag{17}$$

where $\hat{\mathbb{E}}_L[\cdot]$ denotes the empirical average across the $L = 383$ rolling periods.

It is important to highlight a difference between the *empirical* estimate (17) and the *theoretical* out-of-sample squared HJ distance defined in (7). The empirical estimate (17) aggregates $L = 383$ different SDFs, each estimated on a distinct rolling training window

[5] `https://jkpfactors.com`, accessed February 2026.

and evaluated on a single one-month test period. In contrast, the theoretical definition (7) conceptually evaluates a *single* SDF across many independent test samples. As a result of the rolling window construction, the $L = 383$ estimation-evaluation rounds are highly correlated. This induces a constant-scale, non-vanishing sampling bias in (17) relative to the idealized expectation in (7). However, because this bias affects all estimators simultaneously, comparisons across different methods remain fair. Thus, the empirical estimate (17) can be interpreted as an estimate of the theoretical squared HJ distance (7), up to a common sampling bias across all estimators.

With a trade-off between randomness and replicability, we ran all our empirical analyses on 20 independent random seeds spawned from a base seed 42. We report averaged results across the 20 seeds wherever randomness is involved in the estimation process, unless otherwise noted. We did not observe any substantial differences in values or trends across the random seeds we tested.

6.4 Experimental Results

Our empirical experiments aim to validate the theoretical insights developed in Section 4. It is important to note that we do not claim that our theoretical results can be directly applied to the empirical data in a literal sense. Specifically, we do not assert that the data perfectly satisfies Assumption 1, Assumption 2, or Assumption 3, nor that the theoretical bounds hold exactly. Rather, our goal is to test the insights suggested by the theory using real-world data.

Reassuringly, we find that the qualitative predictions developed in Section 4 are strongly supported by the empirical results. We view this as evidence that the underlying insights are broadly applicable beyond the stylized assumptions made in our theoretical analysis. Establishing a more precise connection between theoretical guarantees and empirical behavior is left for future work.

6.4.1 Adding Noise Factors Can Hurt Pricing Performance

We begin by establishing two baseline models that separately illustrate the behavior of "strong" and "noise" factor constructions. Figure 4 presents the results for $\gamma = 0.5$, with $T = 360$ and P ranging from 1 to 120,000. This setting corresponds to a strong model, slightly different from the original specification in Didisheim et al. (2024) due to our use of a fixed $\gamma = 0.5$ instead of $\gamma \sim \text{Unif}(\{0.5, 0.6, \ldots, 1.0\})$. Consistent with the "virtue of complexity" hypothesis (Kelly et al., 2024), we observe a nearly monotonic improvement in out-of-sample performance as P increases.

Figure 5 presents the results for $\gamma = 2$ under otherwise identical conditions. In stark contrast, out-of-sample performance is very poor not only relative to strong large-factor models such as Figure 4, but also compared to classical few-factor models.[6] An examination of the factor structure reveals that the constructed factors carry near-zero individual risk

[6]The stark contrast between $\gamma = 0.5$ and $\gamma = 2$ also highlights that, under the RFF architecture, tuning γ may be more critical than tuning P. While this observation does not affect our subsequent experiments (in fact, our experiments build upon this contrast), it points to an interesting direction for future research: exploring alternative factor construction architectures to better understand the role of complexity in factor models.

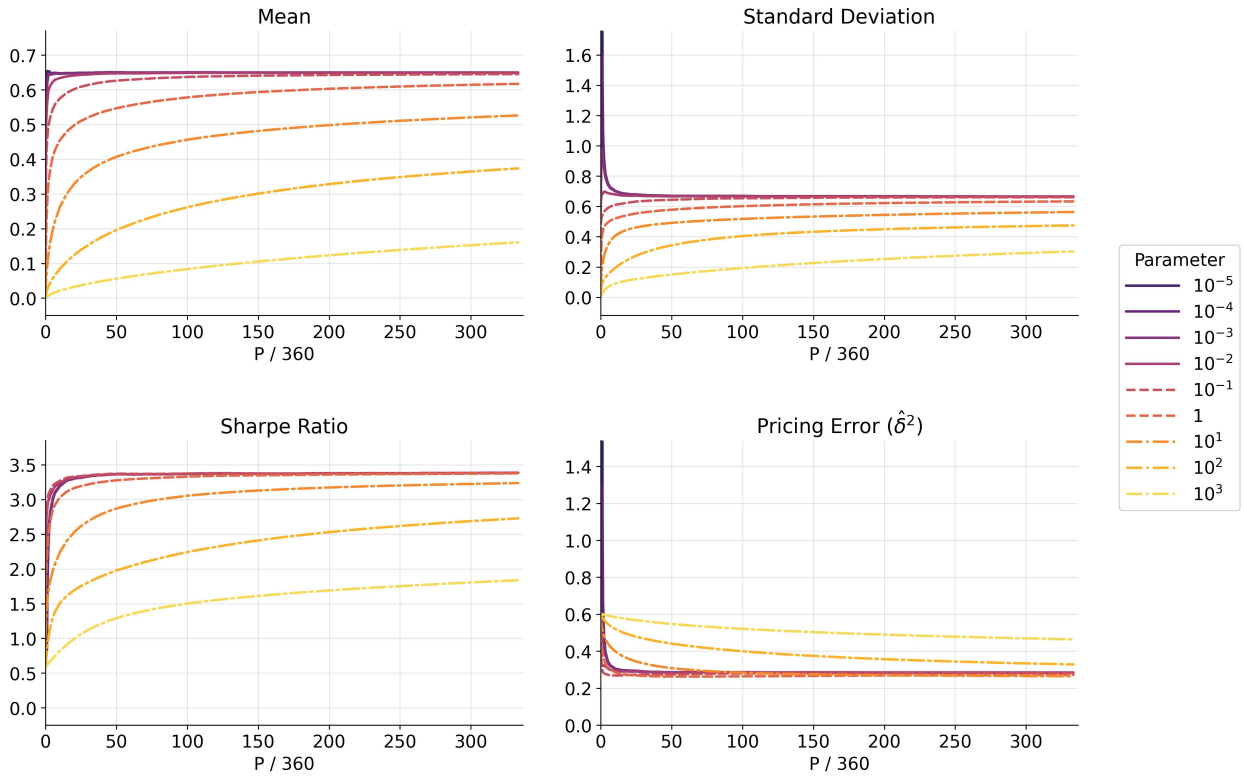

Figure 4. Baseline model with $\gamma = 0.5$ and P varying up to 120,000.
Note: Curves correspond to different ridge penalties z.

premia, and even the tangency portfolio formed from all 120,000 factors fails to achieve a significantly positive Sharpe ratio. Thus, the $\gamma = 2$ factors behave similar to near-pure "noise factors," as defined in Section 4.1.

Building on these two baseline cases, Figure 6 implements a controlled experiment designed to directly test the impact of adding noise factors to an otherwise well-performing model. Specifically, we first sequentially add 30,000 strong ($\gamma = 0.5$) factors, then progressively introduce 90,000 noise ($\gamma = 2$) factors. This allows us to isolate and measure the incremental effect of noise factor inclusion on out-of-sample performance.

The patterns in Figure 6 conform with several of our theoretical predictions. First, noise factors induce a clear out-of-sample performance deterioration. As shown in Figure 6, out-of-sample performance improves steadily up to approximately $P = 30{,}000$, after which it begins to decline. This turning point coincides with the addition of noise factors, validating a core prediction of Section 4.1 that adding noise factors can hurt out-of-sample performance in a way that cannot be corrected even with optimal regularization.

Moreover, Figure 6 aligns with our theoretical expectation that the magnitude of degradation grows with the aggregate variance of noise factors. As P increases, the cumulative variance of the newly added noise factors rises, leading to an upward drift in the self-induced regularization term $\frac{1}{T}\sum_{j=k^*+1}^{P}\lambda_j$. This systematic bias shrinks the SDF loadings on economically meaningful factors, directly reducing the attainable Sharpe ratio. The empirical patterns in Figure 6 also shows that the performance declines gradually, without abrupt collapses, consistent with a smooth but persistent increase in regularization burden.

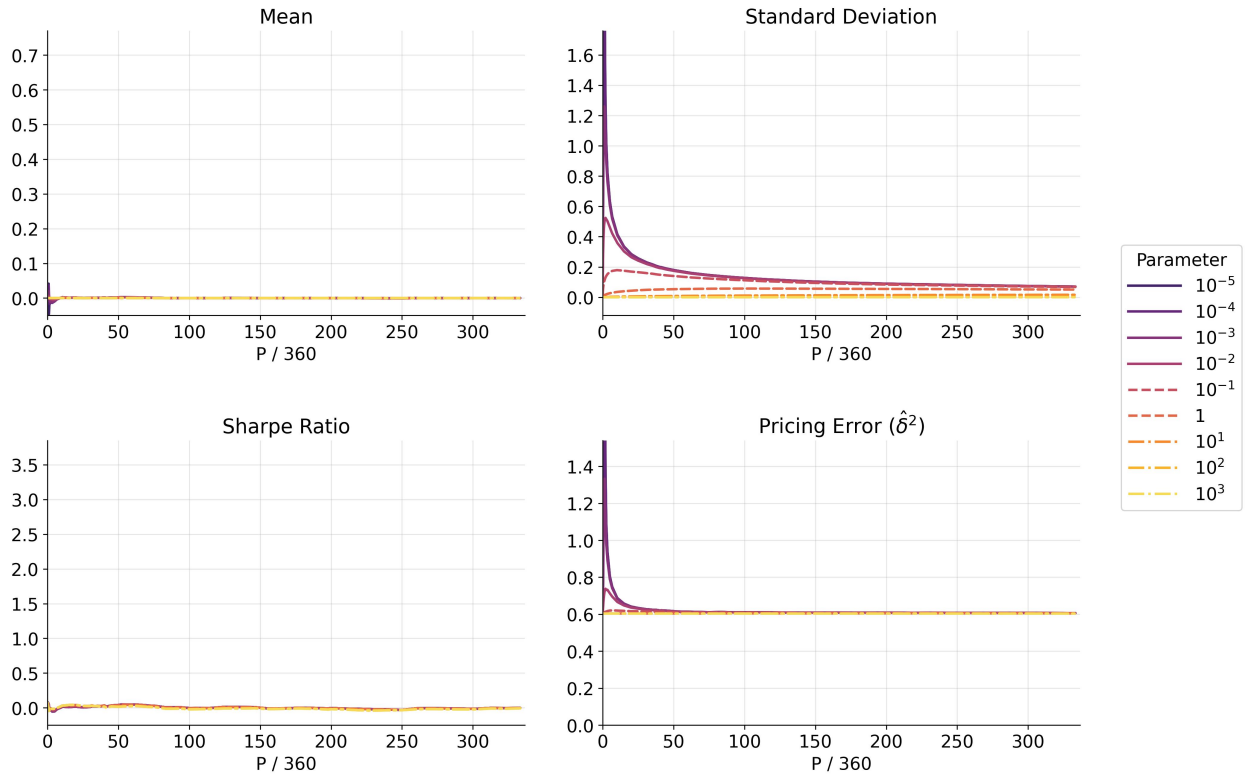

Figure 5. Baseline model with $\gamma = 2$ and P varying up to 120,000.
Note: Curves correspond to different ridge penalties z.

A natural concern when adding noise factors is extreme variance inflation, following the classical bias-variance trade-off. Figure 6 suggests that adding noise factors does not increase the volatility of estimated SDFs. Instead, the variance of the tangency portfolio declines as P increases. The dominant effect is excessive shrinkage of priced exposures, reducing the attainable risk premium. This indicates that the mechanism of performance loss is bias-induced shrinkage, rather than variance inflation, consistent with the theoretical explanations in Section 4.1 and highlighting the distinctive nature of overfitting costs in high-dimensional factor models.

The sensitivity to noise factors, as illustrated in Figure 6, depends on the ridge penalty z. Larger z values "absorb" the impact of added noise factors, while small z values become nearly indistinguishable after noise is introduced. As in Corollary 1, when z is large relative to the self-induced regularization term $\frac{1}{T}\sum_{j=k^*+1}^{P}\lambda_j$, noise has negligible marginal effect, whereas when z is small, self-induced regularization dominates and erodes the distinction across nearby penalties.

Another notable pattern is the robustness of ridge estimator, driven by self-adaptivity. Despite the introduction of 90,000 noise factors, the degradation in performance is moderate. Even when $P = 120{,}000$, the model substantially outperforms conventional small factor models where $T/P < 1$. Two mechanisms can explain this robustness. First, the variances of individual noise factors are very small, resulting in a relatively slow accumulation of self-induced regularization as P increases. Second, as predicted by Corollary 1, the all-inclusive ridge estimator automatically prioritizes high-variance, priced components while suppressing

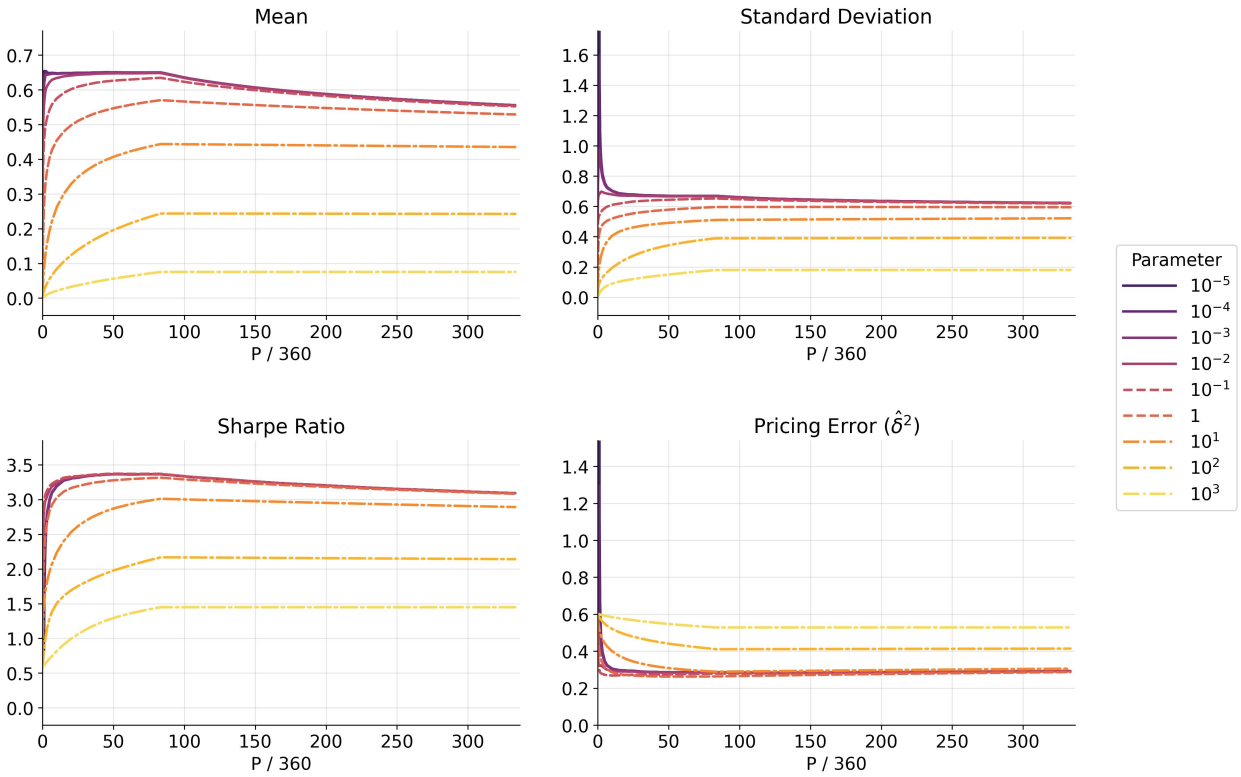

Figure 6. Controlled experiment. First 30,000 factors generated with $\gamma = 0.5$, followed by 90,000 factors generated with $\gamma = 2$.
Note: Curves correspond to different ridge penalties z.

the impact of noise factors, effectively mimicking oracle factor selection without explicit screening (modulo some cost).

One potential concern is that by $P = 30{,}000$ the model is already "saturated," with the kernel limit effectively achieved, so that adding more factors would not improve performance regardless of their quality. However, our robustness tests suggest that this is not the case. We conducted an experiment using a much smaller setting with $P = 7{,}200$. Figure A1 and Figure A2 show the model's performance up to 7,200, while Figure A3 presents the controlled experiment in which the first 1,800 factors were generated with $\gamma = 0.5$, and the remaining 5,400 factors with $\gamma = 2$. The figures are qualitatively similar to those in main results, confirming that the observed patterns are not artifacts of saturation.

Taken together, these results provide strong empirical validation for the theoretical insights developed in this paper. They illustrate that the "virtue of complexity" can fail in practice, not because of classical overfitting, but because of systematic bias introduced by self-induced regularization. Nonetheless, the all-inclusive ridge estimator exhibits considerable resilience to noise, offering a compelling trade-off between flexibility and robustness in high-dimensional SDF estimation.

6.4.2 Adding Weak Factors Can Also Degrade Performance

We now examine the consequences of adding weak factors, complementing the noise factor analysis. Throughout this section, we refer to "weak factors" as those that are weaker relative

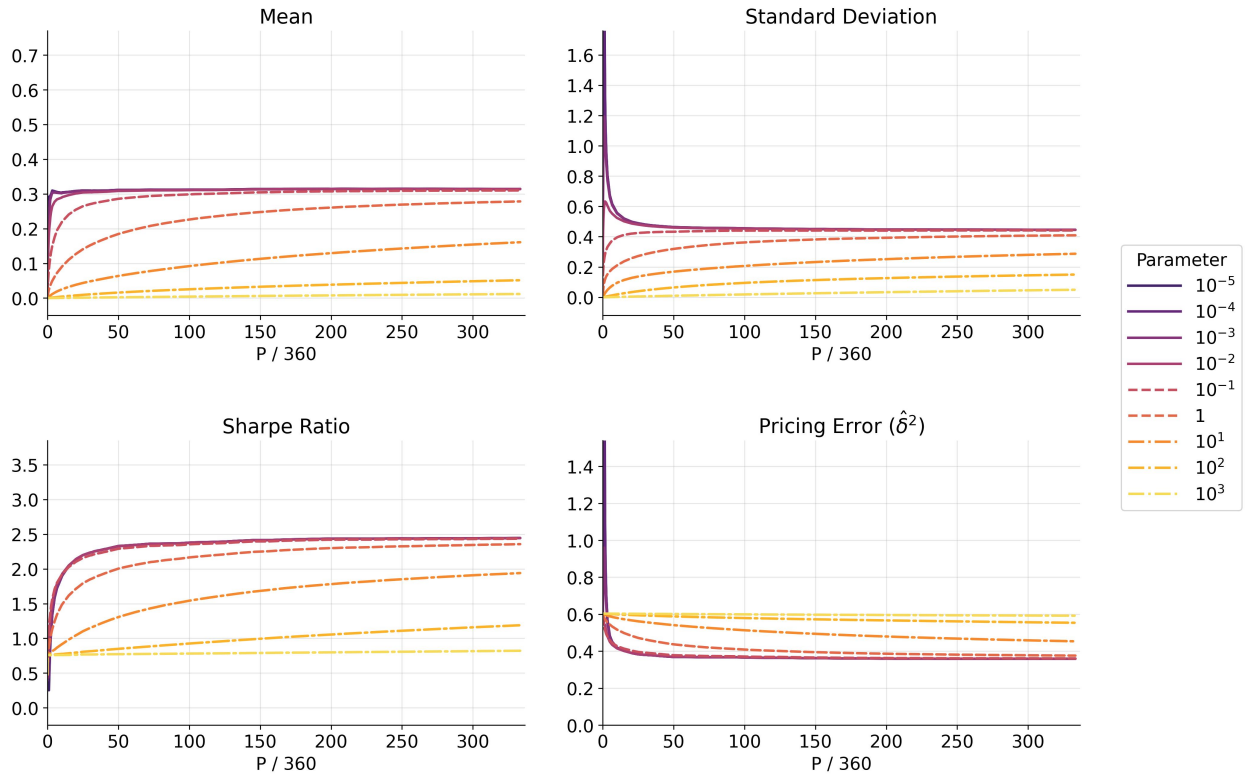

Figure 7. Baseline model with $\gamma = 1$ and P varying up to 120,000.
Note: Curves correspond to different ridge penalties z.

to the strong factors generated with $\gamma = 0.5$, although they still carry economically meaningful risk premia.

Figure 7 presents results for a new baseline model with $\gamma = 1$, with $T = 360$ and P varying up to 120,000. Out-of-sample performance remains strong as P increases, indicating that $\gamma = 1$ factors are reasonably priced and valuable when used alone. However, compared to $\gamma = 0.5$ factors, their signals are significantly weaker: the risk premium of the tangency portfolio constructed from $\gamma = 1$ factors is less than half that of $\gamma = 0.5$, and the Sharpe ratio is also substantially lower.

Similar to the noise factor experiment in Section 6.4.1, we conducted a controlled experiment where we first added 30,000 strong factors generated with $\gamma = 0.5$, followed by 90,000 weak factors generated with $\gamma = 1$. The results are presented in Figure 8.

Adding weak factors degrades performance, similarly to adding noise factors. Performance peaks after the inclusion of the 30,000 strong factors and deteriorates steadily as $\gamma = 1$ factors are added. The magnitude of performance decay is comparable to that observed when adding $\gamma = 2$ noise factors, despite $\gamma = 1$ factors performing much better than $\gamma = 2$ when used alone. This phenomenon is intriguing. A potential explanation is that although $\gamma = 1$ factors carry positive risk premia, they have substantially larger variances than $\gamma = 2$ factors (as seen by comparing Figure 7 and Figure 5). Consequently, the cumulative self-induced regularization term $\frac{1}{T}\sum_{j=k^*+1}^{P} \lambda_j$ grows more rapidly, leading to a trade-off between added risk premia and increased regularization burden.

In addition to the effects of weak factor inclusion, the results further reinforce the role of

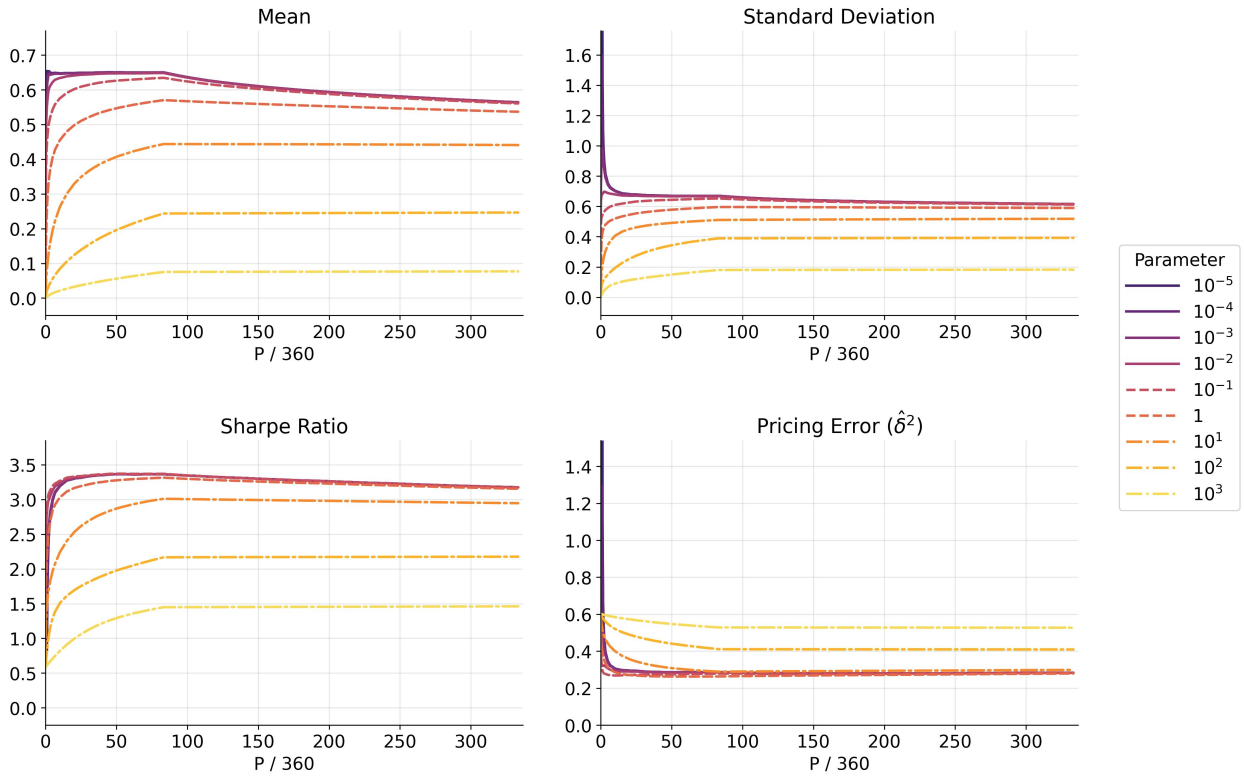

Figure 8. Controlled experiment: first 30,000 factors generated with $\gamma = 0.5$, followed by 90,000 factors generated with $\gamma = 1$.

Note: Curves correspond to different ridge penalties z.

self-adaptivity. The all-inclusive ridge estimator seems to primarily benefit from prioritizing strong $\gamma = 0.5$ factors rather than aggregating many weaker ones. This behavior is consistent with Corollary 1, which emphasizes that the advantages of large factor models stem mainly from self-adaptivity and prioritization of strong signals, rather than from recovering and combining weaker ones.

As a robustness check, we perform the same smaller-scale experiment with P up to 7,200 and find similar patterns (see Figure A4 and Figure A5). Overall, these results highlight that even weak factors, defined relative to a stronger benchmark, can degrade pricing performance when introduced in large numbers. Moreover, in terms of their marginal contribution when added, they do not exhibit significant advantages compared to noise factors, even though they perform much better when used alone. This observation is intriguing, and we plan to explore additional variants of weak factor constructions in future experiments.

6.4.3 Negative Ridge Can Be Optimal in Practice

We now investigate the potential benefits of using negative ridge penalties in large factor models, as motivated by the theoretical results in Section 4.3. To be as close as possible to setups previously studied in the finance literature (where negative ridge has not been considered), we adopt the specification of Didisheim et al. (2024), generating factors using $\gamma \sim \text{Unif}(\{0.5, 0.6, \ldots, 1.0\})$ and setting $P = 360{,}000$. The only modification is that we

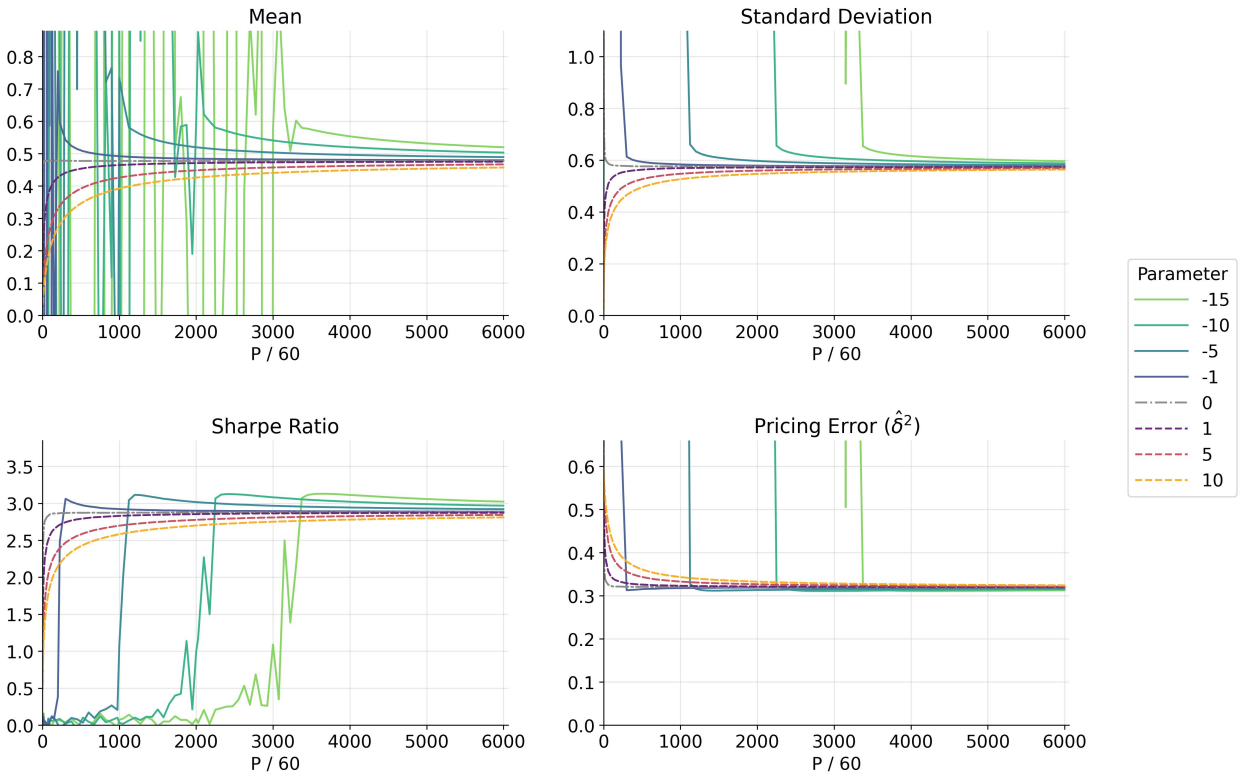

Figure 9. Negative ridge performance with a 60-month window and $P = 360{,}000$.

Note: Each curve corresponds to a fixed ridge penalty z.

reduce the training window size to $T = 60$, making the self-induced regularization more pronounced and facilitating the observation of negative ridge effects, consistent with the threshold characterization in Corollary 1.

Figure 9 shows several of our benchmark results. First, negative ridge can significantly outperform standard ridge. Across a wide range of P, the best-performing ridge penalties are negative. Both the Sharpe ratio and HJ distance improve substantially by introducing a moderately negative z, demonstrating that negative ridge is not merely a theoretical hypothesis but can meaningfully enhance out-of-sample pricing performance.

Second, the optimal level of negativity increases with P. Consistent with the bias-correction formula in (15), larger P magnifies the self-induced regularization and thus requires more negative ridge penalties to offset it. For example, in Figure 9, when $P \approx 18{,}000$, $z = -1$ delivers noticeable improvements; when $P \approx 72{,}000$, $z = -5$ becomes preferable; when $P \approx 144{,}000$, $z = -10$ achieves the best results; and when $P \approx 210,000$, $z = -15$ performs the best. This systematic shift confirms that in the current setup, the optimal penalty level becomes increasingly negative as P grows.

Third, the benefits of negative ridge have intrinsic limits. For any fixed P, there exists a threshold level of negativity beyond which performance collapses. For instance, at $P \approx 72{,}000$, $z = -5$ improves performance relative to $z = -1$ and $z = 0$, but pushing further to $z = -10$ sharply degrades both the Sharpe ratio and HJ distance. This pattern aligns precisely with the theoretical predictions in Section 4.3 and Section B.1: if z falls below $-\frac{1}{T}\sum_{j=k^*+1}^{P} \lambda_j$, the effective regularization becomes negative and destabilizes the model.

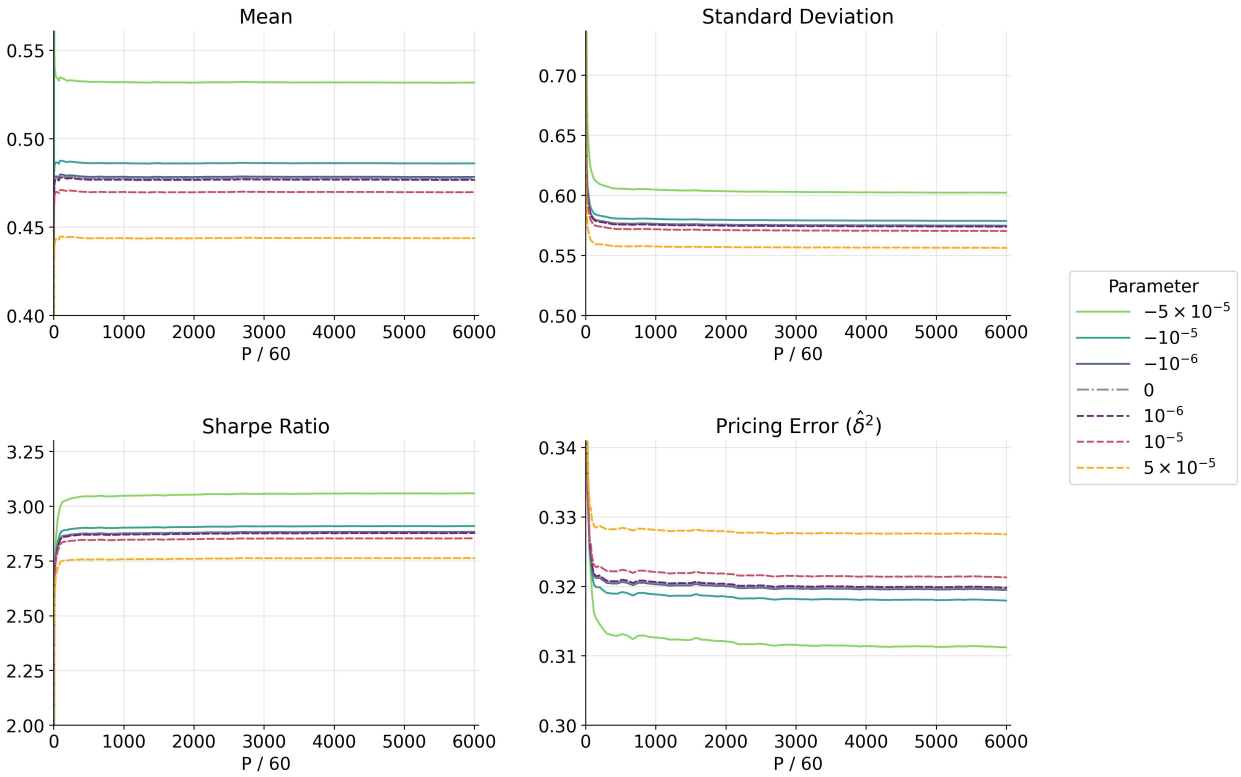

Figure 10. Negative ridge performance with a 60-month window and $P = 360{,}000$.
Note: Each curve corresponds to a scaled penalty $z = z_0 P$, with z_0 indicated in the legend.

Fourth, optimal penalty choices are inherently P-dependent. A penalty z that is effective at large P may destabilize the model at smaller P. This sensitivity reflects the dependence of the self-induced regularization level on the number of factors included. Thus, while negative ridge can enhance performance, its effectiveness is inherently model-dependent, and careful calibration accounting for both the model size and the underlying DGP is essential.

Motivated by these findings, we explore whether a simple scaling rule for z can consistently capture the gains from negative ridge across all P. Figure 10 shows an alternative parametrization where z is chosen proportionally to P, with each curve corresponding to a fixed z_0 value where $z = z_0 P$. As Figure 10 shows, selecting z proportional to P with an appropriate negative z_0 leads to uniform improvements in out-of-sample performance. In particular, negative ridge with a fixed z_0 dominates non-negative ridge across nearly all P.

To sum up, our findings in Figure 9 and Figure 10 empirically confirm the theoretical logic of Section 4.3. In large, overparameterized factor models, optimal regularization must balance both explicit and implicit sources of shrinkage, and allowing for negative penalties can offer an effective correction for self-induced bias.

To the best of our knowledge, this is the first empirical documentation of the practical advantages of negative ridge in the finance literature. Given the foundational role of ridge estimators in asset pricing and economic modeling, we view this finding as a major empirical contribution of this paper. Moreover, the close alignment between our theoretical predictions and the empirical evidence further validates the relevance and robustness of our theoretical framework.

7 Discussions

This paper analyzes the out-of-sample performance of large, overparameterized linear factor models for SDF estimation. Factor libraries in finance keep expanding. Meanwhile, machine learning theory has identified high-dimensional regimes where even unregularized, interpolating estimators can generalize well. In this context, we investigate the behavior of the all-inclusive ridge estimator, which incorporates all candidate factors without any ex-ante screening or dimension reduction. Through new non-asymptotic pricing error bounds, we uncover several important insights into the mechanisms governing the performance of such models.

Our theoretical results reveal three key phenomena: self-induced regularization, where the inclusion of numerous low-variance PC factors implicitly biases estimation of high-variance PC factors; the bounded cost of overfitting, where exact interpolation in the low-variance space does not materially amplify out-of-sample pricing error compared to under-specification; and self-adaptivity, where ridge regression effectively prioritizes the most economically meaningful components without explicit dimension reduction. These findings extend the theory of benign overfitting to the inherently misspecified setting of SDF estimation, where existing benign overfitting conditions fail.

Evidence from both simulations and U.S. stock market data supports our theoretical results. In the simulations, we isolate the predicted mechanisms under controlled spectral designs in the PC space. In the empirical analysis, we use RFF constructions to generate large numbers of factors following prior literature. We show that adding noise factors systematically degrades out-of-sample performance through bias rather than variance inflation, and that adding weak but priced factors can similarly hurt model performance. These experiments also provide evidence that mildly negative ridge penalties have the potential to enhance pricing accuracy, consistent with the theoretical prediction that negative penalties can partially offset self-induced regularization.

Many interesting directions remain open for future exploration. First, while the current experiments in Section 6.4.1 validate the predicted role of self-induced regularization (in setups closely following prior work with minimal adjustments), one can also go beyond the "minimal adjustments" and extend the analysis by introducing more severe noise structures. Specifically, if one constructs settings where added noise factors have substantially larger individual variances, then the aggregate noise variance can grow faster with P. We believe this can induce stronger self-regularization and lead to a sharper deterioration in out-of-sample performance. These experiments would further validate the theoretical insight that the harm from noise factors scales with their cumulative variance (in the PC factor space) and highlight the practical limits of model complexity in overparameterized factor models.

Beyond this specific extension, several broader directions remain open and highly valuable. Refining the theoretical error bounds under more general (possibly non-Gaussian) factor structures would strengthen the link between theory and empirical observations. Exploring alternative architectures for factor construction could further illuminate the role of complexity in affecting the out-of-sample performance. Moreover, developing principled methods for adaptive penalty selection could improve empirical performance, especially when self-induced regularization is significant.

References

Stanislav Anatolyev and Anna Mikusheva. Factor Models With Many Assets: Strong Factors, Weak Factors, and the Two-Pass Procedure. *Journal of Econometrics*, 229(1):103–126, 2022.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign Overfitting in Linear Regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep Learning: A Statistical Viewpoint. *Acta Numerica*, 30:87–201, 2021.

Daniel Barzilai and Ohad Shamir. Generalization in Kernel Regression Under Realistic Assumptions, 2024.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling Modern Machine-Learning Practice and the Classical Bias–Variance Trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two Models of Double Descent for Weak Features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Anders Björkström and Rolf Sundberg. A Generalized View on Continuum Regression. *Scandinavian Journal of Statistics*, 26(1):17–30, 1999.

Mark Britten-Jones. The Sampling Error in Estimates of Mean-Variance Efficient Portfolio Weights. *The Journal of Finance*, 54(2):655–671, 1999.

Svetlana Bryzgalova, Markus Pelger, and Jason Zhu. Forest through the trees: Building cross-sections of stock returns. *Available at SSRN 3493458*, 2019.

Luyang Chen, Markus Pelger, and Jason Zhu. Deep learning in asset pricing. *Management Science*, 70(2):714–750, 2024.

Xiaohong Chen and Sydney C. Ludvigson. Land of Addicts? An Empirical Investigation of Habit-Based Asset Pricing Models. *Journal of Applied Econometrics*, 24(7):1057–1093, 2009.

Antoine Didisheim, Shikun (Barry) Ke, Bryan T. Kelly, and Semyon Malamud. APT or "AIPT"? The Surprising Dominance of Large Factor Models. Working Paper 33012, National Bureau of Economic Research, 2024.

Eugene F. Fama and Kenneth R. French. Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

Eugene F. Fama and Kenneth R. French. A Five-Factor Asset Pricing Model. *Journal of Financial Economics*, 116(1):1–22, 2015.

Georgios Gavrilopoulos, Guillaume Lecué, and Zong Shang. A Geometrical Analysis of Kernel Ridge Regression and Its Applications, 2024.

Stefano Giglio, Dacheng Xiu, and Dake Zhang. Test Assets and Weak Factors. *The Journal of Finance*, 80(1):259–319, 2025.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

Lars Peter Hansen and Ravi Jagannathan. Assessing Specification Errors in Stochastic Discount Factor Models. *The Journal of Finance*, 52(2):557–590, 1997.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in High-Dimensional Ridgeless Least Squares Interpolation. *Annals of Statistics*, 50(2):949, 2022.

Robert J. Hodrick and Xiaoyan Zhang. Evaluating the Specification Errors of Asset Pricing Models. *Journal of Financial Economics*, 62(2):327–376, 2001.

Tsushung A. Hua and Richard F. Gunst. Generalized Ridge Regression: A Note on Negative Ridge Parameters. *Communications in Statistics - Theory and Methods*, 12(1):37–45, 1983.

Theis Ingerslev Jensen, Bryan Kelly, and Lasse Heje Pedersen. Is There a Replication Crisis in Finance? *The Journal of Finance*, 78(5):2465–2518, 2023.

Raymond Kan and Cesare Robotti. Model Comparison Using the Hansen-Jagannathan Distance. *The Review of Financial Studies*, 22(9):3449–3490, 2008.

Bryan Kelly and Dacheng Xiu. Financial Machine Learning. *Foundations and Trends® in Finance*, 13(3-4):205–363, 2023.

Bryan Kelly, Semyon Malamud, and Kangying Zhou. The Virtue of Complexity in Return Prediction. *The Journal of Finance*, 79(1):459–503, 2024.

Bryan T. Kelly and Semyon Malamud. Understanding the virtue of complexity. *Swiss Finance Institute Research Paper No. 25-96*, 2025.

Bryan T Kelly, Seth Pruitt, and Yinan Su. Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524, 2019.

Bryan T. Kelly, Semyon Malamud, Mohammad Pourmohammadi, and Fabio Trojani. Universal Portfolio Shrinkage. Working Paper 32004, National Bureau of Economic Research, 2023.

Bryan T Kelly, Boris Kuznetsov, Semyon Malamud, and Teng Andrea Xu. Artificial intelligence asset pricing models. Technical report, National Bureau of Economic Research, 2025.

Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The Optimal Ridge Penalty for Real-world High-dimensional Data Can Be Zero or Negative Due to the Implicit Ridge Regularization. *Journal of Machine Learning Research*, 21(169):1–16, 2020.

Serhiy Kozak. Kernel Trick for the Cross-Section. *Available at SSRN 3307895*, 2020.

Serhiy Kozak, Stefan Nagel, and Shrihari Santosh. Shrinking the Cross-Section. *Journal of Financial Economics*, 135(2):271–292, 2020.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pages 1302–1338, 2000.

Guillaume Lecué and Zong Shang. A Geometrical Viewpoint on the Benign Overfitting Property of the Minimum ℓ_2-Norm Interpolant Estimator and Its Universality. *Probability Theory and Related Fields*, 2024.

Martin Lettau and Markus Pelger. Estimating Latent Asset-Pricing Factors. *Journal of Econometrics*, 218(1):1–31, 2020.

Bin Li, Alberto G. Rossi, Xuemin (Sterling) Yan, and Lingling Zheng. Machine learning from a "universe" of signals: The role of feature engineering. *Journal of Financial Economics*, 172:104138, 2025.

Stefan Nagel. Seemingly virtuous complexity in return prediction. *Available at SSRN 5335012*, 2025.

Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Stephen A. Ross. The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory*, 13(3):341–360, 1976.

Ohad Shamir. The Implicit Bias of Benign Overfitting. *Journal of Machine Learning Research*, 24(113):1–40, 2023.

Zhouyu Shen and Dacheng Xiu. Can Machines Learn Weak Signals? Working Paper 33421, National Bureau of Economic Research, 2025.

Alexander Tsigler and Peter L. Bartlett. Benign Overfitting in Ridge Regression. *Journal of Machine Learning Research*, 24(1):123, 2023.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Denny Wu and Ji Xu. On the Optimal Weighted ℓ_2 Regularization in Overparameterized Linear Regression. In *Advances in Neural Information Processing Systems*, volume 33, pages 10112–10123, 2020.

Tong Zhang. *Mathematical Analysis of Machine Learning Algorithms*. Cambridge University Press, 2023.

Appendix

A Proof of Proposition 1

Proof of Proposition 1. Since

$$\delta^2(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top)\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\boldsymbol{\mu} + \boldsymbol{\mu}^\top(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top)^{-1}\boldsymbol{\mu} \tag{18}$$

and

$$\mathbb{E}\big[(1 - \boldsymbol{\beta}^\top \boldsymbol{F})^2\big] = \boldsymbol{\beta}^\top \mathbb{E}[\boldsymbol{F}\boldsymbol{F}^\top]\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbb{E}[\boldsymbol{F}] + 1 = \boldsymbol{\beta}^\top(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top)\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\boldsymbol{\mu} + 1,$$

we have $\delta^2(\hat{\boldsymbol{\beta}}) - \delta^2(\boldsymbol{\beta}^*) = \mathbb{E}\Big[(1 - \hat{\boldsymbol{\beta}}^\top \boldsymbol{F})^2\Big] - \mathbb{E}\big[(1 - \boldsymbol{\beta}^{*\top} \boldsymbol{F})^2\big]$.

Since (18) defines a quadratic function of $\boldsymbol{\beta}$, we can apply Taylor expansion around $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ to obtain

$$\begin{aligned}
\delta^2(\hat{\boldsymbol{\beta}}) - \delta^2(\boldsymbol{\beta}^*) &= 2\left[(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top)\boldsymbol{\beta}^* - \boldsymbol{\mu}^\top\right]^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \\
&= 0 + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \\
&= \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2_{\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top},
\end{aligned}$$

where the second equality uses the fact that $\boldsymbol{\beta}^* = (\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top)^{-1}\boldsymbol{\mu}$, which makes the linear term vanish.

Finally, since $\boldsymbol{\beta}^*$ satisfies (4), by (7),

$$\delta^2(\boldsymbol{\beta}^*) = \left(\mathbb{E}\left[(1 - \boldsymbol{\beta}^{*\top}\boldsymbol{F}_{T+1})\boldsymbol{F}_{T+1}\right]\right)^\top \mathbb{E}[\boldsymbol{F}_{T+1}\boldsymbol{F}_{T+1}^\top]^{-1}\mathbb{E}\left[(1 - \boldsymbol{\beta}^{*\top}\boldsymbol{F}_{T+1})\boldsymbol{F}_{T+1}\right] = 0.$$

Hence $\delta^2(\hat{\boldsymbol{\beta}}) - \delta^2(\boldsymbol{\beta}^*) = \delta^2(\hat{\boldsymbol{\beta}})$. □

B Proof of Lemma 1

Proof of Lemma 1. When $z > 0$, $\mathbb{F}^\top\mathbb{F} + Tz \cdot I_{P\times P}$ is invertible and (9) admits a unique solution. Applying the first-order condition gives

$$\hat{\boldsymbol{\beta}}^{\text{all}}(z) = (\mathbb{F}^\top\mathbb{F} + Tz \cdot I_{P\times P})^{-1}\mathbb{F}^\top \mathbf{1}_{T\times 1}.$$

For $\mathbb{F}^\top\mathbb{F} + Tz \cdot I_{P\times P}$, apply the Sherman-Morrison-Woodbury formula:

$$(\mathbb{F}^\top\mathbb{F} + Tz \cdot I_{P\times P})^{-1} = (Tz)^{-1}I_{P\times P} - (Tz)^{-2}\mathbb{F}^\top((Tz)^{-1}\mathbb{F}\mathbb{F}^\top + I_{T\times T})^{-1}\mathbb{F}$$

Therefore we can rewrite the ridge expression (10) with the algebraic identity:

$$
\begin{aligned}
&(\mathbb{F}^\top\mathbb{F} + Tz \cdot I_{P\times P})^{-1}\mathbb{F}^\top \mathbf{1}_{T\times 1} \\
&= (Tz)^{-1}\mathbb{F}^\top \mathbf{1}_{T\times 1} - (Tz)^{-2}\mathbb{F}^\top \Big((Tz)^{-1}\mathbb{F}\mathbb{F}^\top + I_{T\times T}\Big)^{-1}\mathbb{F}\mathbb{F}^\top \mathbf{1}_{T\times 1} \\
&= (Tz)^{-1}\mathbb{F}^\top \left(I_{T\times T} - \Big((Tz)^{-1}\mathbb{F}\mathbb{F}^\top + I_{T\times T}\Big)^{-1}(Tz)^{-1}\mathbb{F}\mathbb{F}^\top\right)\mathbf{1}_{T\times 1} \\
&= (Tz)^{-1}\mathbb{F}^\top \left(I_{T\times T} - \Big((Tz)^{-1}\mathbb{F}\mathbb{F}^\top + I_{T\times T}\Big)^{-1}\Big((Tz)^{-1}\mathbb{F}\mathbb{F}^\top + I_{T\times T} - I_{T\times T}\Big)\right)\mathbf{1}_{T\times 1} \\
&= (Tz)^{-1}\mathbb{F}^\top \left(I_{T\times T} - I_{T\times T} + \Big((Tz)^{-1}\mathbb{F}\mathbb{F}^\top + I_{T\times T}\Big)^{-1} I_{T\times T}\right)\mathbf{1}_{T\times 1} \\
&= (Tz)^{-1}\mathbb{F}^\top \Big((Tz)^{-1}\mathbb{F}\mathbb{F}^\top + I_{T\times T}\Big)^{-1} I_{T\times T}\Big)\mathbf{1}_{T\times 1} \\
&= \mathbb{F}^\top \Big(\mathbb{F}\mathbb{F}^\top + Tz \cdot I_{T\times T}\Big)^{-1}\mathbf{1}_{T\times 1}. \qquad (19)
\end{aligned}
$$

We note that (19) is a standard result; the proof here is adapted from Appendix B of Tsigler and Bartlett (2023).

When $z = 0$, the objective in (9) admits multiple minimizers, and $\hat{\boldsymbol{\beta}}^{\text{all}}(0) = \mathbb{F}^\top(\mathbb{F}\mathbb{F}^\top)^\dagger \mathbf{1}_{T\times 1}$ corresponds to the minimum-norm solution among them. In this case, the identity

$$(\mathbb{F}^\top\mathbb{F})^\dagger \mathbb{F}^\top \mathbf{1}_{T\times 1} = \mathbb{F}^\top(\mathbb{F}\mathbb{F}^\top)^\dagger \mathbf{1}_{T\times 1}$$

holds by a standard property of the Moore-Penrose pseudoinverse, which can be viewed as an extension of (19) to the case $z = 0$.

When $z < 0$, although the objective in (9) may become non-convex and ill-posed, the expression $\hat{\boldsymbol{\beta}}^{\text{all}}(z) = \mathbb{F}^\top(\mathbb{F}\mathbb{F}^\top + Tz \cdot I_{T\times T})^\dagger \mathbf{1}_{T\times 1}$ remains well-defined. Moreover, for all $z \in \mathbb{R}$, we have

$$\left(\mathbb{F}^\top\mathbb{F} + Tz \cdot I_{P\times P}\right)^\dagger \mathbb{F}^\top \mathbf{1}_{T\times 1} = \mathbb{F}^\top \left(\mathbb{F}\mathbb{F}^\top + Tz \cdot I_{T\times T}\right)^\dagger \mathbf{1}_{T\times 1}$$

almost surely. This follows because both $\mathbb{F}^\top\mathbb{F} + Tz \cdot I_{P\times P}$ and $\mathbb{F}\mathbb{F}^\top + Tz \cdot I_{T\times T}$ are invertible almost surely (as $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_T$ are drawn i.i.d. from a continuous distribution), and (19) holds whenever these two matrices are invertible. Since $(\mathbb{F}^\top\mathbb{F} + Tz \cdot I_{P\times P})^\dagger \mathbb{F}^\top \mathbf{1}_{T\times 1}$ represents the minimum-norm solution to the linear system

$$\left(\mathbb{F}^\top\mathbb{F} + Tz \cdot I_{P\times P}\right)\boldsymbol{\beta} = \mathbb{F}^\top \mathbf{1}_{T\times 1},$$

which coincides with the first-order condition for the objective in (9), we conclude that $\hat{\boldsymbol{\beta}}^{\text{all}}(z)$ is the minimum-norm stationary point of the objective in (9), almost surely. □

B.1 Additional Explanations for Negative Ridge

In this subsection, we provide additional insights into the interpretation of $\hat{\boldsymbol{\beta}}^{\text{all}}(z)$ and explain why it can remain numerically stable (i.e., insensitive to small errors in matrix inversion) even when $z < 0$, provided that z is not "too" negative.

First, when $z < 0$, the objective in (9),

$$\|\mathbf{1}_{T\times 1} - \mathbb{F}\boldsymbol{\beta}\|_2^2 + Tz\|\boldsymbol{\beta}\|_2^2,$$

may be ill-posed in the sense that its minimum may be attained at infinity. Nevertheless, as shown in the proof of Lemma 1, $\hat{\boldsymbol{\beta}}^{\text{all}}(z)$ corresponds to the minimum-norm stationary point of the objective, which need not be a global minimizer due to the nonconvexity introduced when $z < 0$. When z is negative but close to zero, this minimum-norm stationary point tends to be close to the minimum-norm least squares estimator $\hat{\boldsymbol{\beta}}^{\text{all}}(0)$, whose statistical properties have been well studied and shown to yield strong empirical and theoretical performance in certain overparameterized settings (Belkin et al., 2019; Bartlett et al., 2020; Kelly et al., 2024; Didisheim et al., 2024). This perspective offers additional theoretical motivation for exploring negative ridge penalties.

Second, when $z < 0$, the numerical computation of $\hat{\boldsymbol{\beta}}^{\text{all}}(z)$ only requires inverting the $T \times T$ matrix $\mathbb{F}\mathbb{F}^\top + Tz \cdot I_{T\times T}$, rather than the $P \times P$ matrix $\mathbb{F}^\top\mathbb{F} + Tz \cdot I_{P\times P}$. Since $\text{rank}(\mathbb{F}\mathbb{F}^\top) = \text{rank}(\mathbb{F}^\top\mathbb{F}) = \min\{T, P\}$ almost surely (because $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_T$ are drawn i.i.d. from a continuous distribution), in the overparameterized regime $P \gg T$, $\mathbb{F}\mathbb{F}^\top$ is almost surely of full rank T, and hence positive definite. The eigenvalues of $\mathbb{F}\mathbb{F}^\top + Tz \cdot I_{T\times T}$ are

$$\lambda_1(\mathbb{F}\mathbb{F}^\top) + Tz, \quad \ldots, \quad \lambda_T(\mathbb{F}\mathbb{F}^\top) + Tz,$$

and its condition number is given by $(\lambda_1(\mathbb{F}\mathbb{F}^\top) + Tz)/(\lambda_T(\mathbb{F}\mathbb{F}^\top) + Tz)$. Therefore, as long as z remains "sufficiently" larger than $-\lambda_T(\mathbb{F}\mathbb{F}^\top)/T$, the condition number can remain moderate, and the numerical computation of $\hat{\boldsymbol{\beta}}^{\text{all}}(z)$ can remain stable.

C Closed Form of the Optimal SDF Coefficients in the PC Factor Space

The purpose of this section it to show $\widetilde{\beta}_i^*$ is proportional to $\widetilde{\mu}_i/\lambda_i$ in the PC factor space. We have

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}^* &= U^\top \boldsymbol{\beta}^* \\
&= U^\top \left(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top\right)^{-1} \boldsymbol{\mu} \\
&= U^\top \left(U\Lambda U^\top + U\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top U^\top\right)^{-1} \boldsymbol{\mu} \\
&= U^\top \left(U\left(\Lambda + \widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top\right)U^\top\right)^{-1} \boldsymbol{\mu} \\
&= U^\top U \left(\Lambda + \widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top\right)^{-1} U^\top U\widetilde{\boldsymbol{\mu}} \\
&= \left(\Lambda + \widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top\right)^{-1} \widetilde{\boldsymbol{\mu}}.
\end{aligned}$$

By applying the Sherman-Morrison-Woodbury formula,

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}^* &= \left(\Lambda + \widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top\right)^{-1}\widetilde{\boldsymbol{\mu}} \\
&= \left(\Lambda^{-1} - \frac{\Lambda^{-1}\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top\Lambda^{-1}}{1+\widetilde{\boldsymbol{\mu}}^\top\Lambda^{-1}\widetilde{\boldsymbol{\mu}}}\right)\widetilde{\boldsymbol{\mu}} \\
&= \Lambda^{-1}\widetilde{\boldsymbol{\mu}} - \frac{\Lambda^{-1}\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top\Lambda^{-1}\widetilde{\boldsymbol{\mu}}}{1+\widetilde{\boldsymbol{\mu}}^\top\Lambda^{-1}\widetilde{\boldsymbol{\mu}}} \\
&= \Lambda^{-1}\widetilde{\boldsymbol{\mu}}\left(1 - \frac{\widetilde{\boldsymbol{\mu}}^\top\Lambda^{-1}\widetilde{\boldsymbol{\mu}}}{1+\widetilde{\boldsymbol{\mu}}^\top\Lambda^{-1}\widetilde{\boldsymbol{\mu}}}\right) \\
&= \frac{\Lambda^{-1}\widetilde{\boldsymbol{\mu}}}{1+\widetilde{\boldsymbol{\mu}}^\top\Lambda^{-1}\widetilde{\boldsymbol{\mu}}}.
\end{aligned}$$

So we have,

$$\widetilde{\beta}_i^* = \left(\frac{1}{1+\sum_{p=1}^P \frac{\widetilde{\mu}_p^2}{\lambda_p}}\right)\cdot\frac{\widetilde{\mu}_i}{\lambda_i}.$$

D Proof of Proposition 2

Proof of Proposition 2. Since $\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}(z) = \widetilde{\mathbb{F}}^\top(\widetilde{\mathbb{F}}\widetilde{\mathbb{F}}^\top + Tz\cdot I_{T\times T})^\dagger \mathbf{1}_{T\times 1}$ and

$$\widetilde{\mathbb{F}} = \left[\widetilde{\boldsymbol{F}}_1,\ldots,\widetilde{\boldsymbol{F}}_T\right]^\top = \left[U^\top\boldsymbol{F}_1,\ldots,U^\top\boldsymbol{F}_T\right]^\top = \left[\boldsymbol{F}_1,\ldots,\boldsymbol{F}_T\right]^\top U = \mathbb{F}U,$$

we have $\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}(z) = U^\top\mathbb{F}^\top(\mathbb{F}UU^\top\mathbb{F}^\top + Tz\cdot I_{T\times T})^\dagger\mathbf{1}_{T\times 1} = U^\top\hat{\boldsymbol{\beta}}^{\mathrm{all}}(z)$. Hence

$$1 - \hat{\boldsymbol{\beta}}^{\mathrm{all}}(z)^\top\boldsymbol{F}_{T+1} = 1 - \hat{\boldsymbol{\beta}}^{\mathrm{all}}(z)^\top UU^\top\boldsymbol{F}_{T+1} = 1 - \hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}(z)^\top\widetilde{\boldsymbol{F}}_{T+1},$$

$$\delta^2(\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}(z)) = \|\hat{\boldsymbol{\beta}}^{\mathrm{all}}(z) - \boldsymbol{\beta}^*\|^2_{\Sigma+\boldsymbol{\mu}\boldsymbol{\mu}^\top} = \|\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}(z) - \widetilde{\boldsymbol{\beta}}^*\|^2_{\Lambda+\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top}.$$

Finally, since

$$\begin{aligned}
\delta^2_{\mathrm{PC}}(\boldsymbol{\beta}) &:= \left(\mathbb{E}\left[(1-\boldsymbol{\beta}^\top\widetilde{\boldsymbol{F}}_{T+1})\widetilde{\boldsymbol{F}}_{T+1}\right]\right)^\top\mathbb{E}[\widetilde{\boldsymbol{F}}_{T+1}\widetilde{\boldsymbol{F}}_{T+1}^\top]^{-1}\mathbb{E}\left[(1-\boldsymbol{\beta}^\top\widetilde{\boldsymbol{F}}_{T+1})\widetilde{\boldsymbol{F}}_{T+1}\right] \\
&= \boldsymbol{\beta}^\top(\Lambda+\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top)\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\widetilde{\boldsymbol{\mu}} + \widetilde{\boldsymbol{\mu}}^\top(\Lambda+\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top)^{-1}\widetilde{\boldsymbol{\mu}}
\end{aligned}$$

defines a quadratic function of $\boldsymbol{\beta}$, we can apply Taylor expansion around $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}^*$ to obtain

$$\begin{aligned}
\delta^2_{\mathrm{PC}}(\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}) &= \delta^2_{\mathrm{PC}}(\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}) - \delta^2_{\mathrm{PC}}(\widetilde{\boldsymbol{\beta}}^*) \\
&= 2\left[(\Lambda+\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top)\widetilde{\boldsymbol{\beta}}^* - \boldsymbol{\mu}\right]^\top(\widetilde{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}) + (\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}} - \widetilde{\boldsymbol{\beta}}^*)^\top(\Lambda+\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top)(\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}} - \widetilde{\boldsymbol{\beta}}^*) \\
&= 0 + (\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}} - \widetilde{\boldsymbol{\beta}}^*)^\top(\Lambda+\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top)(\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}} - \widetilde{\boldsymbol{\beta}}^*) \\
&= \|\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}} - \widetilde{\boldsymbol{\beta}}^*\|^2_{\Lambda+\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top},
\end{aligned}$$

where the second equality uses the fact that $\widetilde{\boldsymbol{\beta}}^* = (\Lambda+\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top)^{-1}\widetilde{\boldsymbol{\mu}}$, which makes the linear term vanish.

□

E Proof of Lemma 2

Proof of Lemma 2. For any $\boldsymbol{\theta} \in \mathbb{R}^P$, by the Cauchy-Schwarz inequality, we have

$$\|\boldsymbol{\theta}\|^2_{\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top} = \|\widetilde{\boldsymbol{\mu}}^\top \boldsymbol{\theta}\|_2^2 \le \left(\sum_{i=1}^P \frac{\widetilde{\mu}_i^2}{\lambda_i}\right) \|\boldsymbol{\theta}\|_\Lambda^2 \le C_{\mathrm{NA}} \|\boldsymbol{\theta}\|_\Lambda^2.$$

Hence $\|\boldsymbol{\theta}\|^2_{\Lambda+\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top} = \|\boldsymbol{\theta}\|^2_\Lambda + \|\boldsymbol{\theta}\|^2_{\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top} \le (1 + C_{\mathrm{NA}})\|\boldsymbol{\theta}\|^2_\Lambda$.

For any benchmark $\boldsymbol{\beta} \in \mathbb{R}^P$, for any $k \in [P]$,

$$\begin{aligned}
&\left|\delta_{\mathrm{PC}}(\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}(z)) - \delta_{\mathrm{PC}}(\boldsymbol{\beta})\right| \\
&= \left|\|\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}(z) - \widetilde{\boldsymbol{\beta}}^*\|_{\Lambda+\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top} - \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}^*\|_{\Lambda+\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top}\right| \\
&\le \|\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}(z) - \boldsymbol{\beta}\|_{\Lambda+\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top} \\
&\le \sqrt{1 + C_{\mathrm{NA}}} \left\|\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}(z) - \boldsymbol{\beta}\right\|_\Lambda \\
&\le \sqrt{1 + C_{\mathrm{NA}}} \left(\left\|\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}_{[1:k]}(z) - \boldsymbol{\beta}_{[1:k]}\right\|_{\Lambda_{[1:k]}} + \left\|\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}_{[k+1:P]}(z) - \boldsymbol{\beta}_{[k+1:P]}\right\|_{\Lambda_{[k+1:P]}}\right).
\end{aligned}$$

The second inequality uses the fact $\|\cdot\|^2_{\Lambda+\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top} \le (1 + C_{\mathrm{NA}})\|\cdot\|^2_\Lambda$ which we just proved, and the last inequality uses the fact that Λ is diagonal. □

F Proof of the Theorem 1

F.1 Definitions

We begin by introducing some definitions and notation.

Matrix notation Let I_n denote the $n \times n$ identity matrix, and let $\mathbf{1}_n$ denote the $n \times 1$ vector of ones. For any matrix $A \in \mathbb{R}^{m \times n}$, let $\|A\|$ denote the Euclidean-induced operator norm of A (i.e., its largest singular value), and $\|A\|_F$ denote its Frobenius norm.

De-meaned PC factors Let $\boldsymbol{G}_t = \widetilde{\boldsymbol{F}}_t - \widetilde{\boldsymbol{\mu}}$. Then $\boldsymbol{G}_t \sim \mathcal{N}(\mathbf{0}, \Lambda)$. For any $i \in [P]$, let $G_{i,t} := \widetilde{F}_{i,t} - \widetilde{\mu}_i$. Then $\boldsymbol{G}_t = (G_{1;t}, \ldots, G_{P;t})^\top$.

Estimation-overfitting decomposition As discussed in Section 3.2, for some $k \in [P]$ we decompose the PC factor space to two subspaces: the "top k" PC factor space and the "bottom $P - k$" PC factor space. This corresponds to split the coordinates $[1 : P]$ into two subgroups: $[1 : k]$ and $[k + 1 : P]$. Thus we introduce the following notation.

- Let $\widetilde{\boldsymbol{\mu}}_{[1:k]} := (\widetilde{\mu}_1, \ldots, \widetilde{\mu}_k)$ and $\widetilde{\boldsymbol{\mu}}_{[k+1:P]} := (\widetilde{\mu}_{k+1}, \ldots, \widetilde{\mu}_P)$. Recall that $\Lambda_{[1:k]} := \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ and $\Lambda_{[k+1:P]} := \mathrm{diag}(\lambda_{k+1}, \ldots, \lambda_P)$.
- Let $\widetilde{\boldsymbol{F}}_{[1:k];t} := (\widetilde{F}_{1;t}, \ldots, \widetilde{F}_{k;t})^\top$ and $\widetilde{\boldsymbol{F}}_{[k+1:P];t} := (\widetilde{F}_{k+1;t}, \ldots, \widetilde{F}_{P;t})^\top$. Similarly, let $\boldsymbol{G}_{[1:k];t} := (G_{1;t}, \ldots, G_{k;t})^\top$ and $\boldsymbol{G}_{[k+1:P];t} := (G_{k+1;t}, \ldots, G_{P;t})^\top$.

- Let

$$\widetilde{\mathbb{F}}_{[1:k]} := [\widetilde{\boldsymbol{F}}_{[1:k];1}, \ldots, \widetilde{\boldsymbol{F}}_{[1:k];T}]^\top = \begin{bmatrix} \widetilde{F}_{1;1} & \widetilde{F}_{2;1} & \cdots & \widetilde{F}_{k;1} \\ \widetilde{F}_{1;2} & \widetilde{F}_{2;2} & \cdots & \widetilde{F}_{k;2} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{F}_{1;T} & \widetilde{F}_{2;T} & \cdots & \widetilde{F}_{k;T} \end{bmatrix},$$

$$\widetilde{\mathbb{F}}_{[k+1:P]} := [\widetilde{\boldsymbol{F}}_{[k+1:P];1}, \ldots, \widetilde{\boldsymbol{F}}_{[k+1:P];T}]^\top = \begin{bmatrix} \widetilde{F}_{k+1;1} & \widetilde{F}_{k+2;1} & \cdots & \widetilde{F}_{P;1} \\ \widetilde{F}_{k+1;2} & \widetilde{F}_{k+2;2} & \cdots & \widetilde{F}_{P;2} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{F}_{k+1;T} & \widetilde{F}_{k+2;T} & \cdots & \widetilde{F}_{P;T} \end{bmatrix}.$$

  Similarly, let

$$\mathbb{G}_{[1:k]} := [\boldsymbol{G}_{[1:k];1}, \ldots, \boldsymbol{G}_{[1:k];T}]^\top = \begin{bmatrix} G_{1;1} & G_{2;1} & \cdots & G_{k;1} \\ G_{1;2} & G_{2;2} & \cdots & G_{k;2} \\ \vdots & \vdots & \ddots & \vdots \\ G_{1;T} & G_{2;T} & \cdots & G_{k;T} \end{bmatrix},$$

$$\mathbb{G}_{[k+1:P]} := [\boldsymbol{G}_{[k+1:P];1}, \ldots, \boldsymbol{G}_{[k+1:P];T}]^\top = \begin{bmatrix} G_{k+1;1} & G_{k+2;1} & \cdots & G_{P;1} \\ G_{k+1;2} & G_{k+2;2} & \cdots & G_{P;2} \\ \vdots & \vdots & \ddots & \vdots \\ G_{k+1;T} & G_{k+2;T} & \cdots & G_{P;T} \end{bmatrix}.$$

- Let $A_{[k+1:P];z} := \widetilde{\mathbb{F}}_{[k+1:P]}\widetilde{\mathbb{F}}_{[k+1:P]}^\top + Tz \cdot I_T$. Let $B_{[k+1,P]} := \mathbb{G}_{[k+1:P]}\mathbb{G}_{[k+1:P]}^\top$. Then we have

$$\mathbb{E}[B_{[k+1,P]}] = \left(\sum_{j=k+1}^{P} \lambda_j\right) \cdot I_T \ ,$$

$$A_{[k+1:P];z} = B_{[k+1:P]} + \|\widetilde{\boldsymbol{\mu}}_{[k+1:P]}\|_2^2 \cdot \mathbf{1}_T\mathbf{1}_T^\top + 2\mathbb{G}_{[k+1:P]}\widetilde{\boldsymbol{\mu}}_{[k+1:P]}\mathbf{1}_T^\top + Tz \cdot I_T \ . \tag{20}$$

  Define the effective ridge penalty $\bar{z} := z + \frac{1}{T}\sum_{j=k+1}^{P} \lambda_j$ and

$$\begin{aligned} A^\circ_{[k+1:P];z} :&= A_{[k+1:P];z} - T\bar{z} \cdot I_T \\ &= \big(B_{[k+1:P]} - \mathbb{E}[B_{[k+1:P]}]\big) + \|\widetilde{\boldsymbol{\mu}}_{[k+1:P]}\|_2^2 \cdot \mathbf{1}_T\mathbf{1}_T^\top + 2\mathbb{G}_{[k+1:P]}\widetilde{\boldsymbol{\mu}}_{[k+1:P]}\mathbf{1}_T^\top. \end{aligned} \tag{21}$$

F.2 Algebraic Identities

Since we focus on the overparameterized regime, we assume $P > T$ and restrict attention to $k < P - T$; this is essentially without loss of generality, as Theorem 1 becomes either standard or trivial when $P \le T$ or $k \ge P - T$.

Because $\widetilde{\boldsymbol{F}}_1, \ldots, \widetilde{\boldsymbol{F}}_T$ are drawn i.i.d. from a continuous distribution, the matrix $\widetilde{\mathbb{F}}_{[k+1:P]}\widetilde{\mathbb{F}}_{[k+1:P]}^\top$ is almost surely of full rank T. Consequently, the matrices $A_{[k+1:P];z}$ and $\widetilde{\mathbb{F}}\widetilde{\mathbb{F}}^\top + Tz \cdot I_T$ are

almost surely invertible.[7] Since events of zero probability do not affect the value of the out-of-sample HJ distance, we condition throughout the proof on the event that $A_{[k+1:P];z}$ and $\widetilde{\mathbb{F}}\widetilde{\mathbb{F}}^\top + Tz \cdot I_T$ are invertible.

Our analysis relies on the following algebraic identities, extensively used in prior work (Tsigler and Bartlett, 2023; Shamir, 2023; Lecué and Shang, 2024).

Lemma 3. *For any $k \in [P]$ and any $z \in \mathbb{R}$, the following hold:*

$$\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}_{[1:k]}(z) = \widetilde{\mathbb{F}}^\top_{[1:k]} \left(\widetilde{\mathbb{F}}_{[1:k]} \widetilde{\mathbb{F}}^\top_{[1:k]} + A_{[k+1:P];z} \right)^{-1} \mathbf{1}_T, \tag{22}$$

$$\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}_{[k+1:P]}(z) = \widetilde{\mathbb{F}}^\top_{[k+1:P]} \left(\widetilde{\mathbb{F}}_{[1:k]} \widetilde{\mathbb{F}}^\top_{[1:k]} + A_{[k+1:P];z} \right)^{-1} \mathbf{1}_T, \tag{23}$$

and

$$\left(I_k + \widetilde{\mathbb{F}}^\top_{[1:k]} A^{-1}_{[k+1:P];z} \widetilde{\mathbb{F}}_{[1:k]} \right) \hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}_{[1:k]}(z) = \widetilde{\mathbb{F}}^\top_{[1:k]} A^{-1}_{[k+1:P];z} \mathbf{1}_T. \tag{24}$$

The proof of Lemma 3 can be found in Appendix F of Tsigler and Bartlett (2023) and in the proof of Proposition 3 in Lecué and Shang (2024).

(24) becomes especially revealing when set beside the PCR benchmark. By definition,

$$\left(I_k + \widetilde{\mathbb{F}}^\top_{[1:k]} (T\bar{z} \cdot I_T)^{-1} \widetilde{\mathbb{F}}_{[1:k]} \right) \hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}(\bar{z}) = \widetilde{\mathbb{F}}^\top_{[1:k]} (T\bar{z} \cdot I_T)^{-1} \mathbf{1}_T. \tag{25}$$

A term-by-term comparison of (24) with (25) shows that, once $A_{[k+1:P];z}$ is "sufficiently close" to $T\bar{z} \cdot I_T$, we can show that $\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}_{[1:k]}(z)$ must also lie close to their PCR comparator $\hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}(\bar{z})$. Demonstrating precisely this proximity is the focus of Section F.4.

F.3 Concentration Inequalities

Our analysis relies on several concentration inequalities to control the norms and eigenvalues of random matrices, which we introduce in this subsection. This is the only part of the proof that involves randomness; all subsequent arguments are purely algebraic and deterministic, conditional on the high-probability events established here.

Lemma 4 (Dvoretzky-Milman theorem)**.** *There exist universal constants $C_1, C_2 > 0$ such that whenever $r(\Lambda_{[k+1:P]}) \geq C_1 T$, with probability at least $1 - \exp\big(1 - C_2 r(\Lambda_{[k+1:P]})\big)$,*

$$\frac{1}{4} \sum_{j=k+1}^{P} \lambda_j \;\leq\; \lambda_T\big(\mathbb{G}_{[k+1:P]} \mathbb{G}^\top_{[k+1:P]}\big) \;\leq\; \lambda_1\big(\mathbb{G}_{[k+1:P]} \mathbb{G}^\top_{[k+1:P]}\big) \;\leq\; \frac{3}{2} \sum_{j=k+1}^{P} \lambda_j.$$

Lemma 4 is Proposition 1 in Lecué and Shang (2024); see that paper for a proof. For later use, we fix $C_1 > 16 C_{\mathrm{NA}}$ (which is always feasible) and only use the following consequence of Lemma 4: conditional on the event in Lemma 4,

$$\lambda_T\big(A_{[k+1:P];z}\big) \;\geq\; \frac{1}{16} T\bar{z}. \tag{26}$$

[7] Invertibility is a very weak requirement for matrices formed from i.i.d. rows or columns drawn from continuous distributions; it holds almost surely and does not require positive definiteness at this stage (high-probability lower bounds on the minimum eigenvalues of $A_{[k+1:P];z}$ and $\widetilde{\mathbb{F}}\widetilde{\mathbb{F}}^\top + Tz \cdot I_T$ will be established later).

Indeed, deterministically,

$$\begin{aligned}
\lambda_T\big(A_{[k+1:P];z}\big) &= \lambda_T\big(\widetilde{\mathbb{F}}_{[k+1:P]}\widetilde{\mathbb{F}}_{[k+1:P]}^\top\big) + Tz \\
&\geq \left(\frac{1}{2}\lambda_T\big(\mathbb{G}_{[k+1:P]}\mathbb{G}_{[k+1:P]}^\top\big) - \big\|\widetilde{\mathbb{F}}_{[k+1:P]} - \mathbb{G}_{[k+1:P]}\big\|^2\right) + Tz \\
&= \frac{1}{2}\lambda_T\big(\mathbb{G}_{[k+1:P]}\mathbb{G}_{[k+1:P]}^\top\big) - T\big\|\widetilde{\boldsymbol{\mu}}_{[k+1:P]}\big\|_2^2 + Tz \\
&\geq \frac{1}{2}\lambda_T\big(\mathbb{G}_{[k+1:P]}\mathbb{G}_{[k+1:P]}^\top\big) - TC_{\mathrm{NA}}\lambda_{k+1} + Tz \\
&= \frac{1}{2}\lambda_T\big(\mathbb{G}_{[k+1:P]}\mathbb{G}_{[k+1:P]}^\top\big) - \frac{C_{\mathrm{NA}}T}{r(\Lambda_{[k+1:P]})}\sum_{j=k+1}^{P}\lambda_j + Tz.
\end{aligned}$$

Combining this with the lower bound in Lemma 4 and $r(\Lambda_{[k+1:P]}) \geq C_1 T > 16 C_{\mathrm{NA}} T$ yields (26).

Lemma 5 (Alternative upper bound of the Dvoretzky-Milman theorem)**.** *For all $T \in \mathbb{N}$, with probability at least $1 - \exp(-T)$:*

$$\lambda_1(\mathbb{G}_{[k+1:P]}\Lambda_{[k+1:P]}\mathbb{G}_{[k+1:P]}^\top) \leq 72T\lambda_{k+1}^2 + 72\sum_{j=k+1}^{P}\lambda_j^2.$$

Lemma 5 is equation (53) in Lecué and Shang (2024), which is a corollary of Proposition 2 in Lecué and Shang (2024). The proof can be found therein.

Lemma 6 (Random matrices with independent Gaussian rows.)**.** *Let $D := \mathbb{G}_{[1:k]}\Lambda_{[1:k]}^{-1/2} = \Big(\widetilde{\mathbb{F}}_{[1:k]} - \mathbf{1}_T\widetilde{\boldsymbol{\mu}}_{[1:k]}^\top\Big)\Lambda_{[1:k]}^{-1/2}$, then D has T i.i.d. rows following $\mathcal{N}(\mathbf{0}, I_k)$. There exists universal constants $C_3, C_4 > 0$, such that for all $T, k \in \mathbb{N}$, with probability at least $1 - \exp(-C_3 T)$:*

$$\frac{\sqrt{T}}{2} - C_4\sqrt{k} \leq s_{\min}(D) \leq s_{\max}(D) \leq \frac{3\sqrt{T}}{2} + C_4\sqrt{k},$$

where $s_{\min}(\cdot), s_{\max}(\cdot)$ denote minimum and maximum singular values. Consequentially, there exists a universal constant $C_0 > 0$ determined by C_3, C_4 and C_{NA}, such that for all $T > 1$,

$$\frac{T - k/(2C_0)}{8} \leq \lambda_k\Big(\Lambda_{[1:k]}^{-1/2}\widetilde{\mathbb{F}}_{[1:k]}^\top\widetilde{\mathbb{F}}_{[1:k]}\Lambda_{[1:k]}^{-1/2}\Big) \leq \lambda_1\Big(\Lambda_{[1:k]}^{-1/2}\widetilde{\mathbb{F}}_{[1:k]}^\top\widetilde{\mathbb{F}}_{[1:k]}\Lambda_{[1:k]}^{-1/2}\Big) \leq (9+4C_{\mathrm{NA}})T + 2C_4^2 k$$

for all $k \leq T$, with probability at least $1 - \exp(-C_0 T)/C_0$.

The proof of Lemma 6, provided below, relies on Theorem 5.39 from Vershynin (2010), with necessary algebraic modifications introduced to manage the non-centered nature of the PC factors.

Proof of Lemma 6. The proof proceeds in two steps. First, we bound the singular values for the standard Gaussian matrix D using standard concentration inequalities. Second, we use a geometric projection argument to control the eigenvalues of the Gram matrix involving the non-centered principal component factors $\widetilde{\mathbb{F}}_{[1:k]}$.

Step 1: Singular values of D. Let $D := (\widetilde{\mathbb{F}}_{[1:k]} - \mathbf{1}_T \widetilde{\boldsymbol{\mu}}_{[1:k]}^\top)\Lambda_{[1:k]}^{-1/2}$. By Assumption 1, the rows of $\widetilde{\mathbb{F}}_{[1:k]}$ are independent Gaussian vectors. Consequently, D is a $T \times k$ random matrix with independent standard normal entries, $D_{ij} \sim \mathcal{N}(0,1)$.

We invoke Theorem 5.39 in Vershynin (2010), which provides non-asymptotic bounds for the singular values of matrices with sub-Gaussian entries. It says that there exists universal constants $C_3 > 1$ and C_4 (corresponding to $c/8$ and C in Theorem 5.39 in Vershynin (2010)), such that for any $t \geq 0$, with probability at least $1 - 2\exp(-(8C_3)t^2)$:

$$\sqrt{T} - C_4\sqrt{k} - t \leq s_{\min}(D) \leq s_{\max}(D) \leq \sqrt{T} + C_4\sqrt{k} + t. \tag{27}$$

We choose the deviation parameter $t = \frac{\sqrt{T}}{2}$. Substituting t into (27), we obtain: with probability at least $1 - 2\exp(-2C_3T) \geq 1 - \exp(-C_3T)$,

$$\frac{\sqrt{T}}{2} - C_4\sqrt{k} \leq s_{\min}(D) \leq s_{\max}(D) \leq \frac{3\sqrt{T}}{2} + C_4\sqrt{k}.$$

Thus, the singular value bounds in Lemma 6 are proved.

Step 2: Eigenvalues of the Gram matrix. Let $Q := \Lambda_{[1:k]}^{-1/2}\widetilde{\mathbb{F}}_{[1:k]}^\top\widetilde{\mathbb{F}}_{[1:k]}\Lambda_{[1:k]}^{-1/2}$. We verify the bounds on $\lambda_1(Q)$ and $\lambda_k(Q)$. Decompose the normalized factor matrix as:

$$\widetilde{\mathbb{F}}_{[1:k]}\Lambda_{[1:k]}^{-1/2} = D + \mathbf{1}_T\mathbf{v}^\top, \quad \text{where } \mathbf{v} = \Lambda_{[1:k]}^{-1/2}\widetilde{\boldsymbol{\mu}}_{[1:k]}.$$

By Assumption 2 (No Asymptotic Arbitrage), $\|\mathbf{v}\|_2^2 = \sum_{i=1}^k \frac{\widetilde{\mu}_i^2}{\lambda_i} \leq C_{\text{NA}}$.

Upper Bound: By the triangle inequality for the operator norm:

$$\sqrt{\lambda_1(Q)} = \|D + \mathbf{1}_T\mathbf{v}^\top\| \leq \|D\| + \|\mathbf{1}_T\|_2\|\mathbf{v}\|_2.$$

Using $\|\mathbf{1}_T\|_2 = \sqrt{T}$, $\|\mathbf{v}\|_2 \leq \sqrt{C_{\text{NA}}}$, and the bound on $\|D\| = s_{\max}(D)$ from Step 1:

$$\sqrt{\lambda_1(Q)} \leq \left(\frac{3\sqrt{T}}{2} + C_4\sqrt{k}\right) + \sqrt{C_{\text{NA}}T} = \left(\frac{3}{2} + \sqrt{C_{\text{NA}}}\right)\sqrt{T} + C_4\sqrt{k}.$$

Squaring both sides and utilizing the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ gives:

$$\lambda_1(Q) \leq 2(3/2 + \sqrt{C_{\text{NA}}})^2T + 2C_4^2k \leq 2(9/2 + 2C_{\text{NA}})T + 2C_4^2k = (9 + 4C_{\text{NA}})T + 2C_4^2k.$$

Lower Bound: We use the variational characterization: $\lambda_k(Q) = \inf_{\|x\|=1} \|(D+\mathbf{1}_T\mathbf{v}^\top)x\|^2$. Let $P_{\mathbf{1}}^\perp = I_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^\top$ be the orthogonal projection onto the subspace orthogonal to $\mathbf{1}_T$. Since projections are contractions, we have:

$$\|(D + \mathbf{1}_T\mathbf{v}^\top)x\|^2 \geq \|P_{\mathbf{1}}^\perp(D + \mathbf{1}_T\mathbf{v}^\top)x\|^2.$$

Since $P_{\mathbf{1}}^\perp\mathbf{1}_T = 0$, the non-zero mean component is eliminated:

$$\|P_{\mathbf{1}}^\perp(D + \mathbf{1}_T\mathbf{v}^\top)x\|^2 = \|P_{\mathbf{1}}^\perp Dx\|^2 \geq s_{\min}^2(P_{\mathbf{1}}^\perp D).$$

We now establish the distribution of the singular values of $\tilde{D} := P_{\mathbf{1}}^{\perp} D$. Since $P_{\mathbf{1}}^{\perp}$ is a rank-$(T-1)$ projection matrix, there exists an orthogonal matrix $U \in \mathbb{R}^{T\times T}$ such that $P_{\mathbf{1}}^{\perp} = U\mathrm{diag}(1,\dots,1,0)U^{\top}$. Let $Z := U^{\top}D$. Since D has i.i.d. $\mathcal{N}(0,1)$ entries and the Gaussian distribution is rotationally invariant, Z also has i.i.d. $\mathcal{N}(0,1)$ entries. We observe:

$$\tilde{D}^{\top}\tilde{D} = D^{\top}P_{\mathbf{1}}^{\perp}P_{\mathbf{1}}^{\perp}D = D^{\top}U\mathrm{diag}(1,\dots,1,0)U^{\top}D = Z^{\top}\mathrm{diag}(1,\dots,1,0)Z.$$

Let $Z_{[1:T-1]}$ denote the first $T-1$ rows of Z. The equation above simplifies to $Z_{[1:T-1]}^{\top}Z_{[1:T-1]}$. Since $Z_{[1:T-1]}$ is a $(T-1)\times k$ matrix with i.i.d. $\mathcal{N}(0,1)$ entries, the non-zero singular values of $\tilde{D}$ are exactly the singular values of a $(T-1)\times k$ standard Gaussian matrix.

Applying Theorem 5.39 from Vershynin (2010) to the $(T-1)\times k$ matrix $\tilde{D}$: for any $t \geq 0$, with probability at least $1 - 2\exp(-8C_3t^2)$,

$$s_{\min}(\tilde{D}) \geq \sqrt{T-1} - C_4\sqrt{k} - t.$$

We choose $t = \frac{\sqrt{T-1}}{4}$. Since $T \geq 2$, we have $T-1 \geq T/2$, so the failure probability is bounded by $2\exp\left(-8C_3\frac{T/2}{16}\right) = 2\exp\left(-\frac{C_3}{4}T\right)$. Substituting t into the singular value bound yields:

$$s_{\min}(\tilde{D}) \geq \frac{3}{4}\sqrt{T-1} - C_4\sqrt{k} \geq \frac{1}{2}\sqrt{T} - C_4\sqrt{k}.$$

Squaring this result yields a lower bound for the eigenvalue:

$$\lambda_k(Q) \geq s_{\min}^2(\tilde{D}) \geq \frac{1}{4}\left(\sqrt{T} - 2C_4\sqrt{k}\right)_+^2 \geq \frac{1}{8}\left(T - \frac{k}{2C_0}\right)$$

for $C_0 \leq \frac{1}{16C_4^2}$.

Combining Step 1 and Step 2, the lemma holds with probability at least $1-2\exp(-2C_3T)-2\exp(-(C_3/4)T) \geq 1-\exp(-C_0T)/C_0$ for $C_0 \leq \min\{1/4, C_3/8, 1/(16C_4^2)\}$. □

Lemma 7 (Deviation of Gram matrix)**.** *There exists a universal constant $C_5 > 0$ depending only on C_{NA}, such that with probability at least $1 - C_5\exp(-T/C_5)$:*

$$\left\|A^{\circ}_{[k+1,P];z}\right\| \leq C_5 T\lambda_{k+1} + C_5\sqrt{T}\sqrt{\sum_{j=k+1}^{P}\lambda_j^2}.$$

Lemma 7 adapts Lemma 23 from Tsigler and Bartlett (2023) (which controls $B_{[k+1:P]} - \mathbb{E}[B_{[k+1:P]}]$ in (21) with high probability) to our setting, combined with using Assumption 2 to control $\|\widetilde{\boldsymbol{\mu}}_{[k+1:P]}\|_2^2 \leq C_{\mathrm{NA}}\lambda_{k+1}$ and $\|\widetilde{\boldsymbol{\mu}}_{[k+1:P]}\|_{\Lambda_{[k+1:P]}} \leq \sqrt{C_{\mathrm{NA}}}\lambda_{k+1}$. The proof of Lemma 23 can be found in Appendix D of Tsigler and Bartlett (2023).

Proof of Lemma 7. Recall the decomposition of the deviation matrix $A^{\circ}_{[k+1:P];z}$ defined in (21):

$$A^{\circ}_{[k+1:P];z} = \underbrace{\left(B_{[k+1:P]} - \mathbb{E}[B_{[k+1:P]}]\right)}_{E_1} + \underbrace{\|\widetilde{\boldsymbol{\mu}}_{[k+1:P]}\|_2^2 \cdot \mathbf{1}_T\mathbf{1}_T^{\top}}_{E_2} + \underbrace{2\mathbb{G}_{[k+1:P]}\widetilde{\boldsymbol{\mu}}_{[k+1:P]}\mathbf{1}_T^{\top}}_{E_3},$$

where $B_{[k+1:P]} = \mathbb{G}_{[k+1:P]}\mathbb{G}_{[k+1:P]}^{\top}$ and $\mathbb{E}[B_{[k+1:P]}] = (\sum_{j=k+1}^{P}\lambda_j)I_T$. We bound the operator norm of each term separately. Let $\Sigma := \Lambda_{[k+1:P]}$ and $\mathbf{v} := \widetilde{\boldsymbol{\mu}}_{[k+1:P]}$. Note that $\|\Sigma\| = \lambda_{k+1}$ and $\|\Sigma\|_F = \sqrt{\sum_{j=k+1}^{P}\lambda_j^2}$.

Part 1: Bounding E_1. The matrix E_1 represents the deviation of the Gram matrix of the centered Gaussian factors. The rows of $\mathbb{G}_{[k+1:P]}$ are i.i.d. draws from $\mathcal{N}(\mathbf{0}, \Sigma)$. We invoke Lemma 23 from Tsigler and Bartlett (2023), which provides a concentration inequality for such matrices. Specifically, for any $t > 0$, with probability at least $1 - 2\exp(-t)$:

$$\|E_1\| \lesssim \|\Sigma\|(T + t) + \|\Sigma\|_F\sqrt{T + t}.$$

Setting $t = T$, there exists a universal constant $c_1 > 0$ such that with probability at least $1 - 2\exp(-T)$:

$$\|E_1\| \le c_1 \left(T\lambda_{k+1} + \sqrt{T}\sqrt{\sum_{j=k+1}^{P} \lambda_j^2} \right). \tag{28}$$

Part 2: Bounding E_2. This term is deterministic. The operator norm of the rank-1 matrix $\mathbf{1}_T\mathbf{1}_T^\top$ is T. Thus, $\|E_2\| = T\|\mathbf{v}\|_2^2$. Using Assumption 2, we have $\sum_{j=1}^{P} \frac{\widetilde{\mu}_j^2}{\lambda_j} \le C_{\mathrm{NA}}$. Since λ_j are non-increasing, for any $j \ge k+1$, we have $\lambda_j \le \lambda_{k+1}$. Therefore:

$$\|\mathbf{v}\|_2^2 = \sum_{j=k+1}^{P} \widetilde{\mu}_j^2 = \sum_{j=k+1}^{P} \frac{\widetilde{\mu}_j^2}{\lambda_j}\lambda_j \le \lambda_{k+1} \sum_{j=k+1}^{P} \frac{\widetilde{\mu}_j^2}{\lambda_j} \le C_{\mathrm{NA}}\lambda_{k+1}.$$

Consequently,

$$\|E_2\| \le C_{\mathrm{NA}}T\lambda_{k+1}. \tag{29}$$

Part 3: Bounding E_3. Let $\mathbf{z} := \mathbb{G}_{[k+1:P]}\mathbf{v}$. The term $E_3 = 2\mathbf{z}\mathbf{1}_T^\top$ is a rank-1 matrix with operator norm $\|E_3\| = 2\|\mathbf{z}\|_2\|\mathbf{1}_T\|_2 = 2\sqrt{T}\|\mathbf{z}\|_2$. Since $\mathbb{G}_{[k+1:P]}$ has i.i.d. $\mathcal{N}(\mathbf{0}, \Sigma)$ rows, the vector $\mathbf{z}$ follows a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \sigma_z^2 I_T)$, where the scalar variance is $\sigma_z^2 := \mathbf{v}^\top\Sigma\mathbf{v} = \sum_{j=k+1}^{P} \widetilde{\mu}_j^2\lambda_j$. We bound σ_z^2 using Assumption 2 and the fact that $\lambda_j \le \lambda_{k+1}$:

$$\sigma_z^2 = \sum_{j=k+1}^{P} \frac{\widetilde{\mu}_j^2}{\lambda_j}\lambda_j^2 \le \lambda_{k+1}^2 \sum_{j=k+1}^{P} \frac{\widetilde{\mu}_j^2}{\lambda_j} \le C_{\mathrm{NA}}\lambda_{k+1}^2.$$

Thus, $\|\mathbf{z}\|_2^2/(C_{\mathrm{NA}}\lambda_{k+1}^2)$ is stochastically dominated by a χ_T^2 variable. Using standard concentration bounds for χ^2 variables (e.g., Lemma 1 in Laurent and Massart (2000)), for any $t > 0$, $\mathbb{P}(\|\mathbf{z}\|_2^2 > \sigma_z^2(T + 2\sqrt{Tt} + 2t)) \le \exp(-t)$. Setting $t = T$:

$$\mathbb{P}\left(\|\mathbf{z}\|_2 > \sqrt{5T}\sigma_z \right) \le \exp(-T).$$

Conditional on this event:

$$\|E_3\| = 2\sqrt{T}\|\mathbf{z}\|_2 \le 2\sqrt{T} \cdot \sqrt{5T}\sqrt{C_{\mathrm{NA}}}\lambda_{k+1} = 2\sqrt{5C_{\mathrm{NA}}}T\lambda_{k+1}.$$

Defining $c_3 := 2\sqrt{5C_{\mathrm{NA}}}$, we have with probability at least $1 - \exp(-T)$:

$$\|E_3\| \le c_3T\lambda_{k+1}. \tag{30}$$

Conclusion. Combining the bounds (28), (29), and (30), there exists a constant C_5 depending only on C_{NA} (and universal constants) such that with probability at least $1 - 3\exp(-T)$:

$$\begin{aligned}\|A^{\circ}_{[k+1:P];z}\| &\le \|E_1\| + \|E_2\| + \|E_3\| \\ &\le (c_1 + C_{\mathrm{NA}} + c_3)T\lambda_{k+1} + c_1\sqrt{T}\sqrt{\sum_{j=k+1}^{P}\lambda_j^2}.\end{aligned}$$

Pick $C_5 \ge \max\{3, c_1 + C_{\mathrm{NA}} + c_3\}$ yields the result. □

Since each of the events described in Lemmas 4 to 7 fails with exponentially small probability (in $-T$, independent of P), there exichsts a universal constant $c > 0$ such that the events in Lemmas 4 to 7 simultaneously hold with probability at least $1 - \exp(-cT)/c$. We assume throughout the rest of the proof that all these events hold.

F.4 Analysis of the Top-k PC Factor Space

The analysis in this section builds on techniques developed in Appendix A of Shamir (2023) and Appendix H.1 of Tsigler and Bartlett (2023). Compared with Tsigler and Bartlett (2023), the key difference is that we use $\hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}(\bar{z})$ as the comparator, motivated by Kozak et al. (2020) and Shamir (2023) (see Section 3.2), and bound $\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}_{[1:k]} - \hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}(\bar{z})$ rather than $\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}_{[1:k]} - \widetilde{\boldsymbol{\beta}}^{*}_{[1:k]}$. This allows us to avoid the well-specified linear regression assumption required by Tsigler and Bartlett (2023). Compared with Shamir (2023), we extend their OLS analysis to the ridge estimator and improve the convergence rate by focusing on pricing error rather than ℓ_2 distance.

Recall that $\bar{z} = z + \frac{1}{T}\sum_{j=k+1}^{P}\lambda_j$. By (24), we have

$$\begin{aligned}&\left(I_k + \widetilde{\mathbb{F}}^{\top}_{[1:k]}A^{-1}_{[k+1:P];z}\widetilde{\mathbb{F}}_{[1:k]}\right)\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}_{[1:k]}(z) \\ &= \widetilde{\mathbb{F}}^{\top}_{[1:k]}A^{-1}_{[k+1:P];z}\mathbf{1}_T \\ &= \widetilde{\mathbb{F}}^{\top}_{[1:k]}A^{-1}_{[k+1:P];z}\widetilde{\mathbb{F}}_{[1:k]}\hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}(\bar{z}) + \widetilde{\mathbb{F}}^{\top}_{[1:k]}A^{-1}_{[k+1:P];z}\left(\mathbf{1}_T - \widetilde{\mathbb{F}}_{[1:k]}\hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}(\bar{z})\right),\end{aligned}$$

hence

$$\begin{aligned}&\left(I_k + \widetilde{\mathbb{F}}^{\top}_{[1:k]}A^{-1}_{[k+1:P];z}\widetilde{\mathbb{F}}_{[1:k]}\right)\left(\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}_{[1:k]}(z) - \hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}(\bar{z})\right) \\ &= \widetilde{\mathbb{F}}^{\top}_{[1:k]}A^{-1}_{[k+1:P];z}\left(\mathbf{1}_T - \widetilde{\mathbb{F}}_{[1:k]}\hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}(\bar{z})\right) - \hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}(\bar{z}) \\ &= \widetilde{\mathbb{F}}^{\top}_{[1:k]}\left(A^{-1}_{[k+1:P];z} - (T\bar{z}\cdot I_T)^{-1}\right)\left(\mathbf{1}_T - \widetilde{\mathbb{F}}_{[1:k]}\hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}(\bar{z})\right) \\ &= \frac{1}{T\bar{z}}\widetilde{\mathbb{F}}^{\top}_{[1:k]}\left(T\bar{z}\cdot A^{-1}_{[k+1:P];z} - I_T\right)\left(\mathbf{1}_T - \widetilde{\mathbb{F}}_{[1:k]}\hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}(\bar{z})\right) \\ &= -\frac{1}{T\bar{z}}\widetilde{\mathbb{F}}^{\top}_{[1:k]}A^{\circ}_{[k+1:P];z}A^{-1}_{[k+1:P];z}\left(\mathbf{1}_T - \widetilde{\mathbb{F}}_{[1:k]}\hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}(\bar{z})\right), \qquad (31)\end{aligned}$$

where the second equality uses the fact that

$$\begin{aligned}
\widetilde{\mathbb{F}}_{[1:k]}^\top\Big(\mathbf{1}_T - \widetilde{\mathbb{F}}_{[1:k]}\hat{\boldsymbol{\beta}}_{[1:k]}^{\mathrm{PC}(k)}(\bar{z})\Big) &= \widetilde{\mathbb{F}}_{[1:k]}^\top\mathbf{1}_T - \widetilde{\mathbb{F}}_{[1:k]}^\top\widetilde{\mathbb{F}}_{[1:k]}\Big(T\bar{z}\cdot I_k + \widetilde{\mathbb{F}}_{[1:k]}^\top\widetilde{\mathbb{F}}_{[1:k]}\Big)^{-1}\widetilde{\mathbb{F}}_{[1:k]}^\top\mathbf{1}_T \\
&= \Big(I_k - \widetilde{\mathbb{F}}_{[1:k]}^\top\widetilde{\mathbb{F}}_{[1:k]}\Big(T\bar{z}\cdot I_k + \widetilde{\mathbb{F}}_{[1:k]}^\top\widetilde{\mathbb{F}}_{[1:k]}\Big)^{-1}\Big)\widetilde{\mathbb{F}}_{[1:k]}^\top\mathbf{1}_T \\
&= T\bar{z}\cdot\Big(T\bar{z}\cdot I_k + \widetilde{\mathbb{F}}_{[1:k]}^\top\widetilde{\mathbb{F}}_{[1:k]}\Big)^{-1}\widetilde{\mathbb{F}}_{[1:k]}^\top\mathbf{1}_T \\
&= T\bar{z}\cdot\hat{\boldsymbol{\beta}}_{[1:k]}^{\mathrm{PC}(k)}(\bar{z}).
\end{aligned}$$

Now we apply the techniques in Appendix H.1 in Tsigler and Bartlett (2023), with a modified error vector determined by the comparator. Define $\boldsymbol{\xi} := \hat{\boldsymbol{\beta}}_{[1:k]}^{\mathrm{PC(all)}}(z) - \hat{\boldsymbol{\beta}}_{[1:k]}^{\mathrm{PC}(k)}(\bar{z})$. Multiplying both sides of (31) by $\xi^\top$ from the left, we obtain

$$\boldsymbol{\xi}^\top\boldsymbol{\xi} + \boldsymbol{\xi}^\top\widetilde{\mathbb{F}}_{[1:k]}^\top A_{[k+1:P];z}^{-1}\widetilde{\mathbb{F}}_{[1:k]}\boldsymbol{\xi} \le \frac{1}{T\bar{z}}\Big|\boldsymbol{\xi}^\top\widetilde{\mathbb{F}}_{[1:k]}^\top A_{[k+1:P];z}^{\circ}A_{[k+1:P];z}^{-1}\Big(\mathbf{1}_T - \widetilde{\mathbb{F}}_{[1:k]}\hat{\boldsymbol{\beta}}_{[1:k]}^{\mathrm{PC}(k)}(\bar{z})\Big)\Big|.$$

Next, divide and multiply by $\Lambda_{[1:k]}^{1/2}$ in several places to get quadratic forms:

$$\begin{aligned}
&\boldsymbol{\xi}^\top\Lambda_{[1:k]}^{1/2}\Lambda_{[1:k]}^{-1}\Lambda_{[1:k]}^{1/2}\boldsymbol{\xi} + \boldsymbol{\xi}^\top\Lambda_{[1:k]}^{1/2}\Lambda_{[1:k]}^{-1/2}\widetilde{\mathbb{F}}_{[1:k]}^\top A_{[k+1:P];z}^{-1}\widetilde{\mathbb{F}}_{[1:k]}\Lambda_{[1:k]}^{-1/2}\Lambda_{[1:k]}^{1/2}\boldsymbol{\xi} \\
&\le \frac{1}{T\bar{z}}\Big|\boldsymbol{\xi}^\top\Lambda_{[1:k]}^{1/2}\Lambda_{[1:k]}^{-1/2}\widetilde{\mathbb{F}}_{[1:k]}^\top A_{[k+1:P];z}^{\circ}A_{[k+1:P];z}^{-1}\Big(\mathbf{1}_T - \widetilde{\mathbb{F}}_{[1:k]}\hat{\boldsymbol{\beta}}_{[1:k]}^{\mathrm{PC}(k)}(\bar{z})\Big)\Big| \qquad (32)
\end{aligned}$$

We lower bound the two quadratic forms on the LHS by extracting the appropriate minimum eigenvalues via the Rayleigh-Ritz variational principle (equivalently, Rayleigh quotient bounds). In particular, for any symmetric matrix B and any $\boldsymbol{v}$,

$$\boldsymbol{v}^\top B\boldsymbol{v} \;\ge\; \lambda_{\min}(B)\,\|\boldsymbol{v}\|_2^2, \qquad \text{and} \qquad \lambda_{\min}(X^\top BX) \;\ge\; \lambda_{\min}(B)\,\lambda_{\min}(X^\top X).$$

Let $\boldsymbol{u} = \Lambda_{[1:k]}^{1/2}\boldsymbol{\xi}$, so that $\|\boldsymbol{u}\|_2^2 = \|\boldsymbol{\xi}\|_{\Lambda_{[1:k]}}^2$. Then the first term on the LHS of (32) becomes

$$\boldsymbol{\xi}^\top\Lambda_{[1:k]}^{1/2}\Lambda_{[1:k]}^{-1}\Lambda_{[1:k]}^{1/2}\boldsymbol{\xi} = \boldsymbol{u}^\top\Lambda_{[1:k]}^{-1}\boldsymbol{u} \;\ge\; \lambda_{\min}(\Lambda_{[1:k]}^{-1})\|\boldsymbol{u}\|_2^2 = \lambda_1^{-1}\,\|\boldsymbol{\xi}\|_{\Lambda_{[1:k]}}^2.$$

For the second term on the LHS of (32):

$$\begin{aligned}
&\boldsymbol{\xi}^\top\Lambda_{[1:k]}^{1/2}\Lambda_{[1:k]}^{-1/2}\widetilde{\mathbb{F}}_{[1:k]}^\top A_{[k+1:P];z}^{-1}\widetilde{\mathbb{F}}_{[1:k]}\Lambda_{[1:k]}^{-1/2}\Lambda_{[1:k]}^{1/2}\boldsymbol{\xi} \\
&\quad= \boldsymbol{u}^\top\Big(\Lambda_{[1:k]}^{-1/2}\widetilde{\mathbb{F}}_{[1:k]}^\top A_{[k+1:P];z}^{-1}\widetilde{\mathbb{F}}_{[1:k]}\Lambda_{[1:k]}^{-1/2}\Big)\boldsymbol{u} \\
&\quad\ge \lambda_{\min}\Big(\Lambda_{[1:k]}^{-1/2}\widetilde{\mathbb{F}}_{[1:k]}^\top A_{[k+1:P];z}^{-1}\widetilde{\mathbb{F}}_{[1:k]}\Lambda_{[1:k]}^{-1/2}\Big)\,\|\boldsymbol{u}\|_2^2 \\
&\quad= \lambda_{\min}\Big(\big(\widetilde{\mathbb{F}}_{[1:k]}\Lambda_{[1:k]}^{-1/2}\big)^\top A_{[k+1:P];z}^{-1}\big(\widetilde{\mathbb{F}}_{[1:k]}\Lambda_{[1:k]}^{-1/2}\big)\Big)\,\|\boldsymbol{u}\|_2^2 \\
&\quad\ge \lambda_{\min}(A_{[k+1:P];z}^{-1})\,\lambda_{\min}\Big(\Lambda_{[1:k]}^{-1/2}\widetilde{\mathbb{F}}_{[1:k]}^\top\widetilde{\mathbb{F}}_{[1:k]}\Lambda_{[1:k]}^{-1/2}\Big)\,\|\boldsymbol{u}\|_2^2 \\
&\quad= \lambda_T(A_{[k+1:P];z}^{-1})\,\lambda_k\Big(\Lambda_{[1:k]}^{-1/2}\widetilde{\mathbb{F}}_{[1:k]}^\top\widetilde{\mathbb{F}}_{[1:k]}\Lambda_{[1:k]}^{-1/2}\Big)\,\|\boldsymbol{\xi}\|_{\Lambda_{[1:k]}}^2.
\end{aligned}$$

For the RHS of (32), we apply Cauchy-Schwarz to obtain an upper bound. Let

$$\boldsymbol{\zeta} = \mathbf{1}_T - \widetilde{\mathbb{F}}_{[1:k]}\hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}(\bar{z}).$$

Then

$$\begin{aligned}
\frac{1}{T\bar{z}}\left|\boldsymbol{\xi}^\top \Lambda^{1/2}_{[1:k]}\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}A^\circ_{[k+1:P];z}A^{-1}_{[k+1:P];z}\boldsymbol{\zeta}\right| &= \frac{1}{T\bar{z}}\left|\left(\Lambda^{1/2}_{[1:k]}\boldsymbol{\xi}\right)^\top\left(\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}A^\circ_{[k+1:P];z}A^{-1}_{[k+1:P];z}\boldsymbol{\zeta}\right)\right| \\
&\le \frac{1}{T\bar{z}}\left\|\Lambda^{1/2}_{[1:k]}\boldsymbol{\xi}\right\|_2\left\|\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}A^\circ_{[k+1:P];z}A^{-1}_{[k+1:P];z}\boldsymbol{\zeta}\right\|_2 \\
&= \frac{1}{T\bar{z}}\|\boldsymbol{\xi}\|_{\Lambda_{[1:k]}}\left\|\widetilde{\mathbb{F}}^\top_{[1:k]}A^\circ_{[k+1:P];z}A^{-1}_{[k+1:P];z}\boldsymbol{\zeta}\right\|_{\Lambda^{-1}_{[1:k]}},
\end{aligned}$$

Combining the lower bound for the LHS and the upper bound for the RHS of (32), yielding

$$\begin{aligned}
&\|\boldsymbol{\xi}\|^2_{\Lambda_{[1:k]}}\left(\lambda_1^{-1} + \lambda_T(A^{-1}_{[k+1:P];z})\,\lambda_k\Big(\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}\widetilde{\mathbb{F}}_{[1:k]}\Lambda^{-1/2}_{[1:k]}\Big)\right) \\
&\le \frac{1}{T\bar{z}}\|\boldsymbol{\xi}\|_{\Lambda_{[1:k]}}\left\|\widetilde{\mathbb{F}}^\top_{[1:k]}A^\circ_{[k+1:P];z}A^{-1}_{[k+1:P];z}\big(\mathbf{1}_T - \widetilde{\mathbb{F}}_{[1:k]}\hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}(\bar{z})\big)\right\|_{\Lambda^{-1}_{[1:k]}}. \qquad (33)
\end{aligned}$$

Hence

$$\begin{aligned}
\|\boldsymbol{\xi}\|_{\Lambda_{[1:k]}} &\le \frac{1}{T\bar{z}}\,\frac{\left\|\widetilde{\mathbb{F}}^\top_{[1:k]}A^\circ_{[k+1:P];z}A^{-1}_{[k+1:P];z}\boldsymbol{\zeta}\right\|_{\Lambda^{-1}_{[1:k]}}}{\lambda_1^{-1} + \lambda_T(A^{-1}_{[k+1:P];z})\,\lambda_k\Big(\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}\widetilde{\mathbb{F}}_{[1:k]}\Lambda^{-1/2}_{[1:k]}\Big)} \\
&= \frac{1}{T\bar{z}}\,\frac{\lambda_1(A_{[k+1:P];z})\,\lambda_1\left\|\widetilde{\mathbb{F}}^\top_{[1:k]}A^\circ_{[k+1:P];z}A^{-1}_{[k+1:P];z}\boldsymbol{\zeta}\right\|_{\Lambda^{-1}_{[1:k]}}}{\lambda_1(A_{[k+1:P];z}) + \lambda_1\,\lambda_k\Big(\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}\widetilde{\mathbb{F}}_{[1:k]}\Lambda^{-1/2}_{[1:k]}\Big)} \\
&= \frac{1}{T\bar{z}}\,\frac{\lambda_1(A_{[k+1:P];z})\,\lambda_1\left\|\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}A^\circ_{[k+1:P];z}A^{-1}_{[k+1:P];z}\boldsymbol{\zeta}\right\|_2}{\lambda_1(A_{[k+1:P];z}) + \lambda_1\,\lambda_k\Big(\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}\widetilde{\mathbb{F}}_{[1:k]}\Lambda^{-1/2}_{[1:k]}\Big)} \\
&\le \frac{1}{T\bar{z}}\,\frac{\lambda_1(A_{[k+1:P];z})\,\lambda_1\left\|\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}\right\|\left\|A^\circ_{[k+1:P];z}\right\|\left\|A^{-1}_{[k+1:P];z}\right\|\|\boldsymbol{\zeta}\|_2}{\lambda_1(A_{[k+1:P];z}) + \lambda_1\,\lambda_k\Big(\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}\widetilde{\mathbb{F}}_{[1:k]}\Lambda^{-1/2}_{[1:k]}\Big)} \\
&= \frac{1}{T\bar{z}}\,\frac{\lambda_1(A_{[k+1:P];z})}{\lambda_T(A_{[k+1:P];z})}\,\frac{\lambda_1\left\|\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}\right\|\left\|A^\circ_{[k+1:P];z}\right\|\|\boldsymbol{\zeta}\|_2}{\lambda_1(A_{[k+1:P];z}) + \lambda_1\,\lambda_k\Big(\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}\widetilde{\mathbb{F}}_{[1:k]}\Lambda^{-1/2}_{[1:k]}\Big)} \\
&= \frac{1}{T\bar{z}}\,\frac{\lambda_1(A_{[k+1:P];z})}{\lambda_T(A_{[k+1:P];z})}\,\frac{\lambda_1\sqrt{\lambda_1\Big(\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}\widetilde{\mathbb{F}}_{[1:k]}\Lambda^{-1/2}_{[1:k]}\Big)}\left\|A^\circ_{[k+1:P];z}\right\|\|\boldsymbol{\zeta}\|_2}{\lambda_1(A_{[k+1:P];z}) + \lambda_1\,\lambda_k\Big(\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}\widetilde{\mathbb{F}}_{[1:k]}\Lambda^{-1/2}_{[1:k]}\Big)} \\
&\le \frac{1}{T\bar{z}}\,\frac{\lambda_1(A_{[k+1:P];z})}{\lambda_T(A_{[k+1:P];z})}\,\frac{\lambda_1\sqrt{\lambda_1\Big(\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}\widetilde{\mathbb{F}}_{[1:k]}\Lambda^{-1/2}_{[1:k]}\Big)}\left\|A^\circ_{[k+1:P];z}\right\|\|\mathbf{1}_T\|_2}{\lambda_1(A_{[k+1:P];z}) + \lambda_1\,\lambda_k\Big(\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}\widetilde{\mathbb{F}}_{[1:k]}\Lambda^{-1/2}_{[1:k]}\Big)}. \qquad (34)
\end{aligned}$$

where $\boldsymbol{\zeta} = \mathbf{1}_T - \widetilde{\mathbb{F}}_{[1:k]}\hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}(\bar{z})$. Here the first inequality is obtained by dividing the (33) by $\|\boldsymbol{\xi}\|_{\Lambda_{[1:k]}}$ on both sides. The second line uses $\lambda_T(A^{-1}_{[k+1:P];z}) = 1/\lambda_1(A_{[k+1:P];z})$. The third line rewrites the norm $\|\cdot\|_{\Lambda^{-1}_{[1:k]}} = \|\Lambda^{-1/2}_{[1:k]}\cdot\|_2$. The fourth line applies submultiplicativity of the operator norm $\|\cdot\|$ (See Section F.1 for the operator norm notation). The fifth line uses that A is symmetric positive definite thus $\|A^{-1}_{[k+1:P];z}\| = \lambda_1(A^{-1}_{[k+1:P];z}) = 1/\lambda_T(A_{[k+1:P];z})$. The sixth line uses $\|\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}\|^2 = \lambda_1(\Lambda^{-1/2}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]}\widetilde{\mathbb{F}}_{[1:k]}\Lambda^{-1/2}_{[1:k]})$. The last inequality follows from (11) which implies

$$\left\|\mathbf{1}_T - \widetilde{\mathbb{F}}_{[1:k]}\hat{\boldsymbol{\beta}}^{\mathrm{PC}(k)}_{[1:k]}(\bar{z})\right\|_2 \le \|\mathbf{1}_T\|_2.$$

Note that $\|\mathbf{1}_T\|_2 \le \sqrt{T}$, which is a good property of the SDF estimation problem.

Conditional on the events in Lemmas 4 to 7, we can further upper-bound the RHS of (34) and obtain

$$\begin{aligned}\|\boldsymbol{\xi}\|_{\Lambda_{[1:k]}} &\lesssim \frac{1}{T\bar{z}}\,\frac{T\bar{z} + T\lambda_{k+1} + \sqrt{T}\sqrt{\sum_{j=k+1}^P \lambda_j^2}}{T\bar{z}}\,\frac{T\lambda_1\left(T\lambda_{k+1} + \sqrt{T}\sqrt{\sum_{j=k+1}^P \lambda_j^2}\right)}{T\bar{z} + (T - 0.5k/C_0)_+\lambda_1} \\ &\lesssim \frac{\lambda_1}{\lambda_1(1 - k/(2C_0T))_+ + \bar{z}}\,\frac{T\lambda_{k+1} + \sqrt{T}\sqrt{\sum_{j=k+1}^P \lambda_j^2}}{T\bar{z}} \lesssim \varepsilon_\Sigma(k,z),\end{aligned}$$

where the last $\lesssim$ holds for both $k \le C_0T$ (only requiring $\bar{z} > 0$) and $k > C_0T$ (requiring $\bar{z} \gtrsim \lambda_1$).

F.5 Analysis of the Bottom-(P-k) PC Factor Space

The analysis in this section combines ideas from Appendix A in Shamir (2023) and Appendix A.3 in Lecué and Shang (2024). Compared with Shamir (2023), we extend their OLS analysis to the ridge estimator and improve the convergence rate by focusing on pricing error rather than ℓ_2 distance. Compared to Lecué and Shang (2024), we do not need their well-specified linear regression assumption; as a result, we obtain a slower convergence rate, which nonetheless suffices for proving Theorem 1.

By (23), we have

$$\begin{aligned}\left\|\hat{\boldsymbol{\beta}}^{\mathrm{PC(all)}}_{[k+1:P]}(z) - \mathbf{0}\right\|_{\Lambda_{[k+1:P]}} &= \left\|\widetilde{\mathbb{F}}^\top_{[k+1:P]}\left(\widetilde{\mathbb{F}}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]} + A_{[k+1:P];z}\right)^{-1}\mathbf{1}_T\right\|_{\Lambda_{[k+1:P]}} \\ &\le \frac{1}{\lambda_T\left(\widetilde{\mathbb{F}}_{[1:k]}\widetilde{\mathbb{F}}^\top_{[1:k]} + A_{[k+1:P];z}\right)}\left\|\widetilde{\mathbb{F}}_{[k+1:P]}\right\|_{\Lambda_{[k+1:P]}}\|\mathbf{1}_T\|_2 \\ &\le \frac{1}{\lambda_T\left(A_{[k+1:P];z}\right)}\left\|\widetilde{\mathbb{F}}_{[k+1:P]}\right\|_{\Lambda_{[k+1:P]}}\|\mathbf{1}_T\|_2.\end{aligned}$$

Conditional on the events in Lemmas 4 to 7, we can further upper-bound the RHS of (34), up to a universal constant, by

$$\frac{1}{T\bar{z}}\cdot\left(\sqrt{T}\lambda_{k+1} + \sum_{j=k+1}^P \lambda_j^2\right)\cdot\sqrt{T} \lesssim \varepsilon_\Sigma(k,z).$$

F.6 Final Step

We have established (12) and (13) in the preceding subsections. The final step is simple: combining these with Lemma 2 yields (14). We note that the condition $r(\Lambda_{[k+1:P]}) \geq C_1 T$ in Theorem 1 arises directly from Lemma 4. Moreover, when $z > -\frac{1}{2T}\sum_{j=k+1}^{P}\lambda_j$, we have $\varepsilon_\Sigma(k,z) \lesssim 1/\sqrt{r(\Lambda_{[k+1,P]})/T}$ because

$$\frac{\sum_{j=k+1}^{P}\lambda_j}{\sqrt{T\sum_{j=k+1}^{P}\lambda_j^2}} \geq \sqrt{\frac{\sum_{j=k+1}^{P}\lambda_j}{T\lambda_{k+1}}}.$$

G Robustness Empirical Tests

G.1 Smaller-Scale Experiments with P up to 7,200

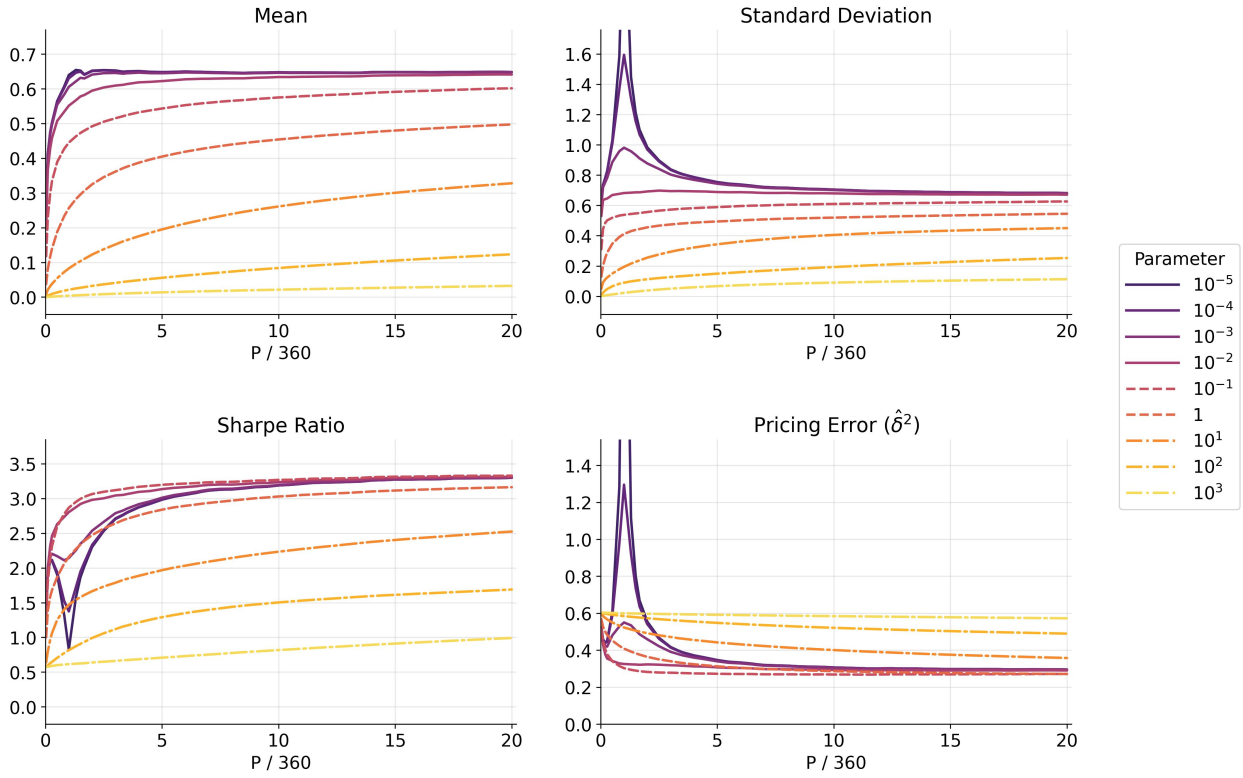

Figure A1. Baseline model with $\gamma = 0.5$ and P varying up to 7,200. Curves correspond to different ridge penalties z.

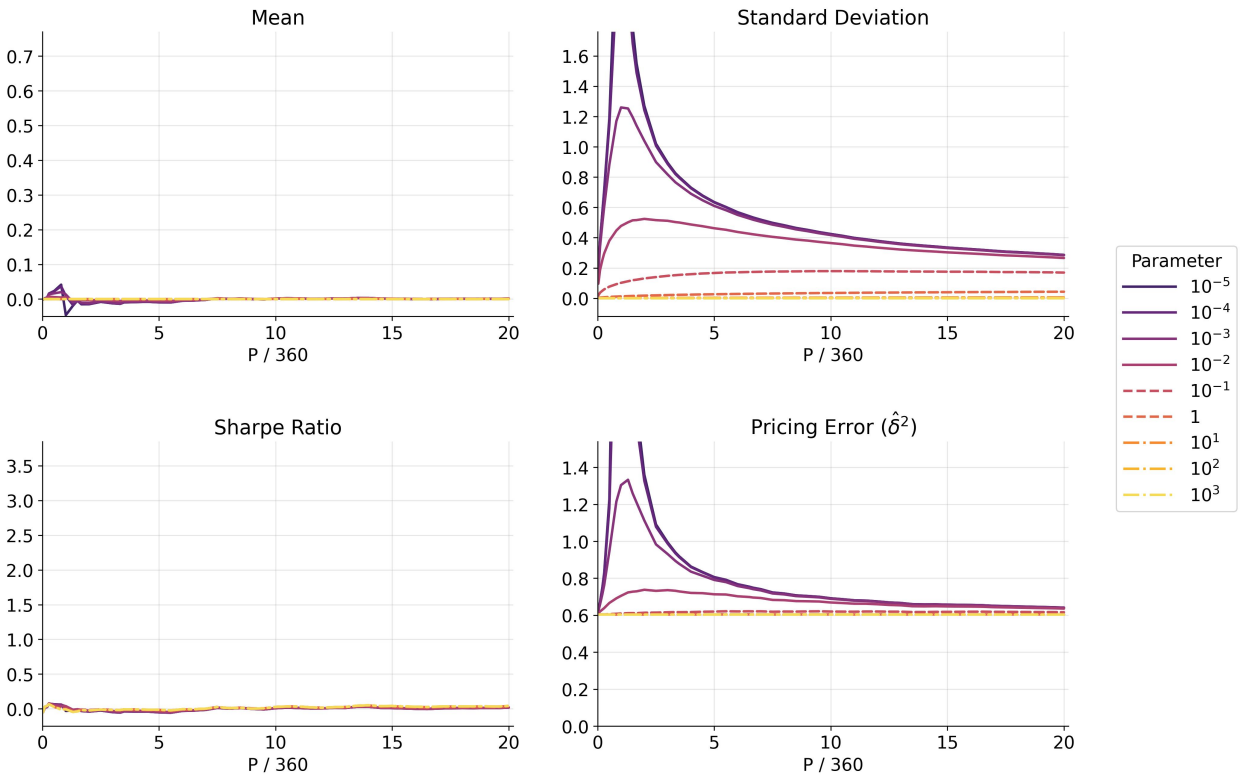

Figure A2. Baseline model with $\gamma = 2$ and P varying up to 7,200. Curves correspond to different ridge penalties z.

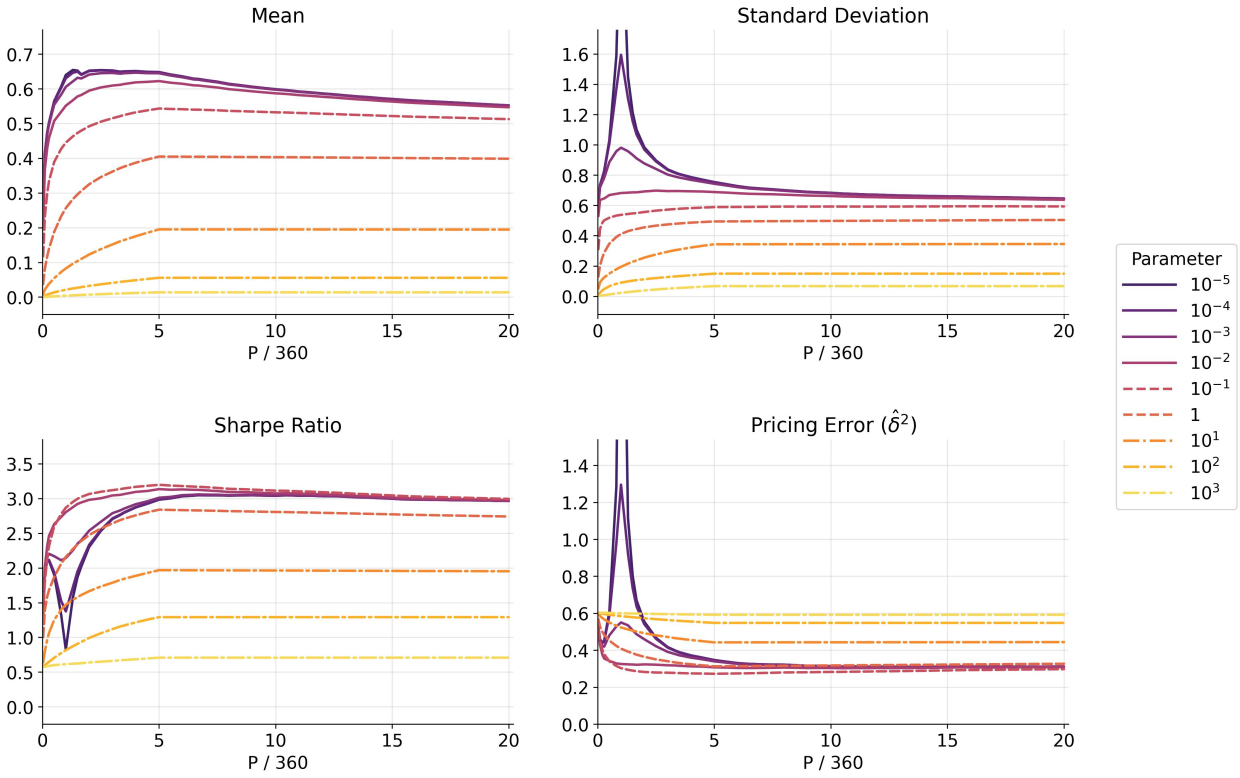

Figure A3. Controlled experiment: first 1,800 factors generated with $\gamma = 0.5$, followed by 5,400 factors generated with $\gamma = 2$. Curves correspond to different ridge penalties z.

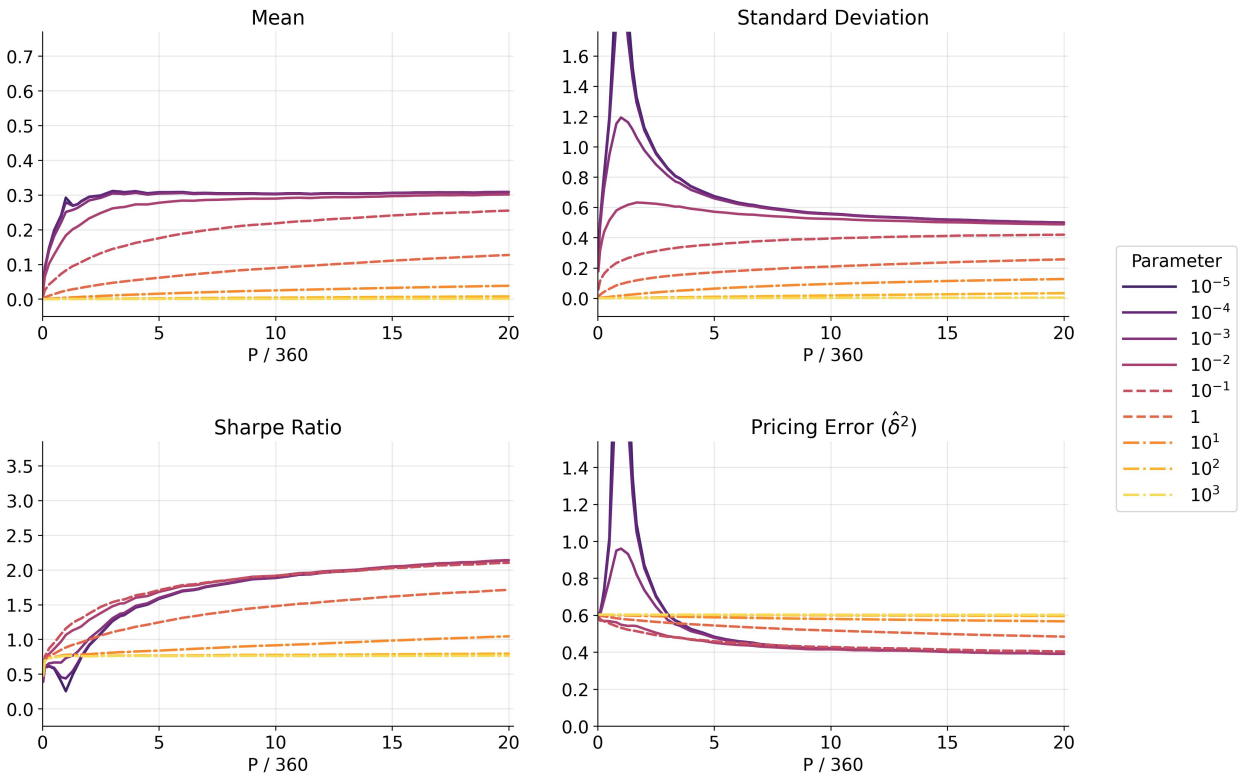

Figure A4. Baseline model with $\gamma = 1$ and P varying up to 7,200. Curves correspond to different ridge penalties z.

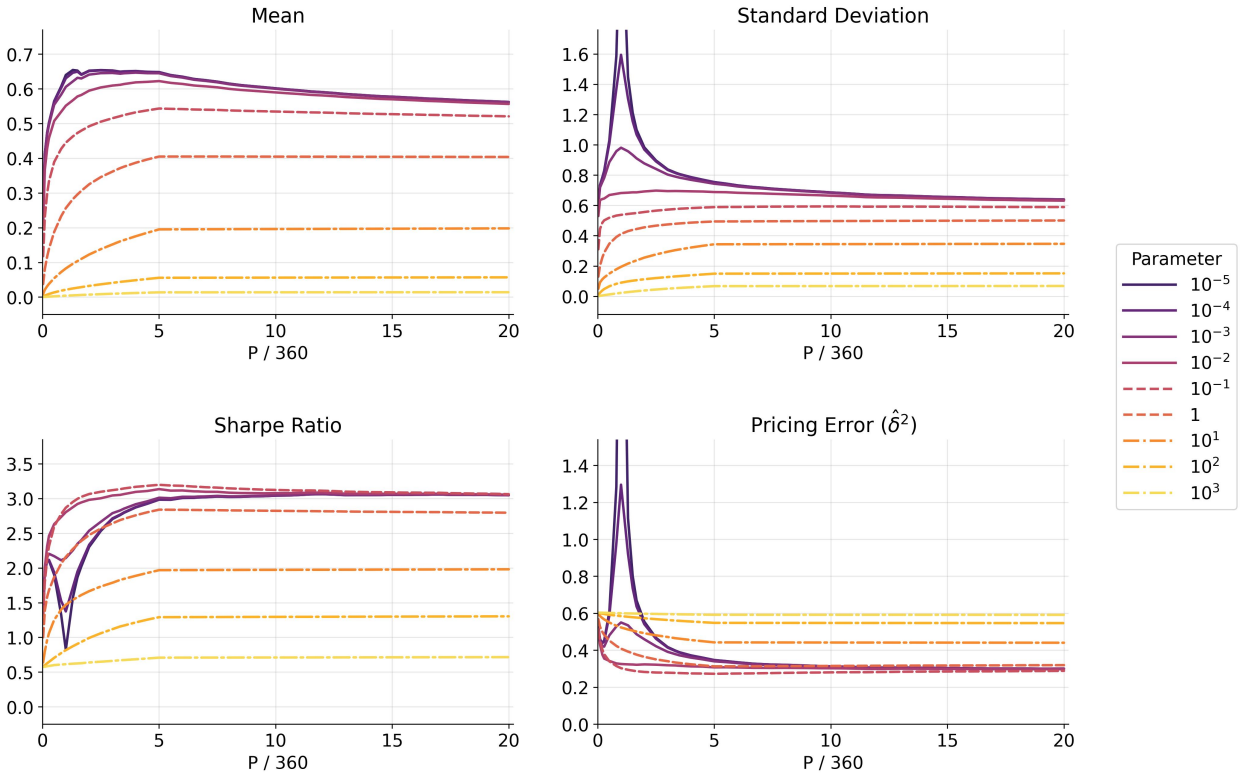

Figure A5. Controlled experiment: first 1,800 factors generated with $\gamma = 0.5$, followed by 5,400 factors generated with $\gamma = 1$. Curves correspond to different ridge penalties z.