NBER WORKING PAPER SERIES

APT OR "AIPT"? THE SURPRISING DOMINANCE OF LARGE FACTOR MODELS

Antoine Didisheim
Shikun (Barry) Ke
Bryan T. Kelly
Semyon Malamud

Working Paper 33012
http://www.nber.org/papers/w33012

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2024

A previously circulated version of this article was titled "Complexity in Factor Pricing Models." Antoine Didisheim is at the University of Melbourne. Shikun Ke is at Yale School of Management. Bryan Kelly is at Yale School of Management, AQR Capital Management, and NBER; www.bryankellyacademic.org. Semyon Malamud is at Swiss Finance Institute, EPFL, and CEPR, and is a consultant to AQR. We are grateful for helpful comments from Federico Baldi-Lanfranchi, John Campbell, Mike Chernov, Martin Lettau, Mohammad Pourmohammadi, Mirela Sandulescu, Fabio Trojani, Neng Wang, and participants at many seminars and conferences. Semyon Malamud gratefully acknowledges the financial support of the Swiss Finance Institute and the Swiss National Science Foundation, Grant 100018 192692. AQR Capital Management is a global investment management firm that may or may not apply similar investment techniques or methods of analysis as described herein. The views expressed here are those of the authors and not necessarily those of AQR. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w33012

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

APT or "AIPT"? The Surprising Dominance of Large Factor Models
Antoine Didisheim, Shikun (Barry) Ke, Bryan T. Kelly, and Semyon Malamud
NBER Working Paper No. 33012
September 2024
JEL No. C1, C40, C45, C55, G1, G11, G12, G14, G17

ABSTRACT

We introduce artificial intelligence pricing theory (AIPT). In contrast with the APT's foundational assumption of a low dimensional factor structure in returns, the AIPT conjectures that returns are driven by a large number of factors. We first verify this conjecture empirically and show that nonlinear models with an exorbitant number of factors (many more than the number of training observations or base assets) are far more successful in describing the out-of-sample behavior of asset returns than simpler standard models. We then theoretically characterize the behavior of large factor pricing models, from which we show that the AIPT's "many factors" conjecture faithfully explains our empirical findings, while the APT's "few factors" conjecture is contradicted by the data.

Antoine Didisheim
Extranef 251
University of Lausanne
Lausanne, Swit 1015
Switzerland
antoine.didisheim@unimelb.edu.au

Shikun (Barry) Ke
Yale School of Management
165 Whitney Ave.
New Haven, CT 06511
barry.ke@yale.edu

Bryan T. Kelly
Yale School of Management

and NBER
bryan.kelly@yale.edu

Semyon Malamud
Swiss Finance Institute @ EPFL
Quartier UNIL-Dorigny, Extranef 213
CH - 1015 Lausanne
Switzerland
semyon.malamud@epfl.ch

1 Introduction

Ross's (1976) APT conjectures that a small number of common factors govern the joint variation of returns. This premise, combined with no-arbitrage arguments, delivers the prediction that assets' expected returns are determined by exposures to a few common factors. Most empirical investigations of the risk-return tradeoff in the past fifty years have occurred within the confines of the APT's assumption—small linear factor models.

In this paper, we consider a different conjecture: that asset pricing models with an exorbitant number of factors are better suited to describe the behavior of asset returns. Our conjecture is rooted in the burgeoning theory of artificial intelligence, which suggests that "complex" statistical models, those with many more parameters (P) than available training observations (T), tend to achieve better out-of-sample success than smaller models.[1] Given its origins, we refer to this conjecture as artificial intelligence pricing theory, or AIPT, to contrast it with the parsimony assumption at the heart of the APT.

One reason for the APT's position as a cornerstone of empirical asset pricing is the convenience of working with small linear factor models. The most obvious convenience stems from the tractability of estimating small models. More subtle is the convenience of *comparing* small models. The APT sets up conditions—that the number of parameters is small compared to the number of training observations—necessary for comparing models based on *in-sample* test statistics. Large factor models, on the other hand, are admittedly inconvenient. With so many parameters, they can be costly to train, and due to their high

[1] See, among others, Hastie et al. (2019), Bartlett et al. (2020), and Kelly et al. (2024b). Following these papers, we define model complexity as the ratio of factor dimension to the number of training observations $c = P/T$. Our theoretical derivations rely crucially on the ratio c, hence our focus on this definition of complexity. The statistics literature proposes and analyzes other measures of model complexity as well, each capturing somewhat different qualities. As noted by Bartlett et al. (2020): "The classical perspective in statistical learning theory is that there should be a tradeoff between the fit to the training data and the complexity of the prediction rule. Whether complexity is measured in terms of the number of parameters, the number of nonzero parameters in a high-dimensional setting, the number of neighbors averaged in a nearest neighbor estimator, the scale of an estimate in a reproducing kernel Hilbert space, or the bandwidth of a kernel smoother, this tradeoff has been ubiquitous in statistical learning theory."

statistical complexity, standard in-sample asset pricing tests are not applicable. Instead, they require out-of-sample model performance comparisons.

We show that, despite their inconveniences, large factor models are far better at pricing assets than the standard low-dimensional factor models in the literature. In the first half of this paper, we present extensive empirical evidence that challenges the APT parsimony conjecture and favors the complexity conjecture of AIPT. The second half of the paper develops a statistical theory that characterizes the behavior of complex factor models and rationalizes their surprising dominance in pricing assets.

1.1 Empirical Behavior of Large Factor Models

Our central empirical finding is that the out-of-sample performance of factor pricing models is increasing in the number of factors. We focus on two metrics of model performance— the Sharpe ratio of the tangency portfolio among factors and the magnitude of pricing errors among test assets.

Our analysis gradually increases the number of factors while holding the underlying information set fixed. The information set includes 130 well-known characteristics for the cross-section of US stocks. To vary the number of pricing factors, we use a neural network to generate new stock characteristics—up to hundreds of thousands—that are nonlinear transformations of the original 130 variables. For each nonlinear characteristic, we build the corresponding factor (i.e., using a standard characteristic-managed portfolio approach). Training the factor model amounts to building the stochastic discount factor (SDF) as the tangency portfolio of factors (using ridge shrinkage where applicable). We then track the two model performance metrics out-of-sample. Our main empirical results are presented in the form of "VoC curves" that plot performance as a function of the number of factors, as introduced by Kelly et al. (2024b) (KMZ henceforth). These curves document a "virtue of complexity" in factor pricing models; out-of-sample Sharpe ratios increase, and pricing

errors decrease as we expand the number of factors. Our results show that there is a virtue of complexity even when—in fact, especially when—P far exceeds T.

Large factor models outperform well-known, low-dimensional factor models from the literature by a wide margin. The largest model we consider (with 360,000 factors, or a complexity ratio of 1,000 given our 360-month rolling training window) reduces pricing errors by 54.8% relative to the six-factor Fama and French (2015) model (including momentum), by 50.4% relative to the five-factor Hou et al. (2021) model, and by 54.8% relative to the six-factor Barillas and Shanken (2018) model. Meanwhile, our largest factor model produces an out-of-sample tangency Sharpe ratio of 3.7, compared to 0.8, 1.2, and 0.8 for the three aforementioned benchmarks, respectively.

The asset pricing benefits of complexity and nonlinearity accrue even when the conditioning information set is relatively small. For example, we construct a "nonlinear Fama-French model" using a large set of nonlinear factors derived from *only* five characteristics: size, value, investment, profitability, and momentum, rather than the full set of 130 characteristics in our main analysis. The added complexity more than doubles the out-of-sample tangency Sharpe ratio relative to the baseline Fama-French model. The same result obtains for complex nonlinear versions of other benchmark models in the literature.

The insights of Kozak et al. (2020) suggest that a successful asset pricing model may not require many factors because the returns of most anomaly factors are adequately explained by a small number of their principal components. This begs the question: Can our complex pricing model be compressed to achieve similar performance with potentially many fewer parameters? Evidently, the answer is no. We consider replacing the large number of factors in our complex model with a smaller number of their principal components and find that dimension reduction significantly impairs out-of-sample performance relative to the full complex model.

Lastly, we conduct a variety of robustness analyses that consider different input char-

acteristics, various subsets of the stock universe, different training windows, and so forth. Every robustness test reinforces our main findings. In summary, we establish the empirical fact that large factor models are better at pricing assets than existing parsimonious models.

1.2 Theoretical Findings

In order to understand our empirical findings, we develop a statistical theory of large factor pricing models. Consider the following interpretation of a large factor model from the asset pricing theory perspective of an SDF. A conditional SDF can be represented as a tradable portfolio of risky assets:

$$M_{t+1} = 1 - w(Z_t)' R_{t+1}. \tag{1}$$

R_{t+1} contains excess returns of risky assets, and $w(Z_t)$ are the SDF's weights on those assets. The weights are determined by variables Z_t that span the time t information set. Without further guidance about the functional form of w, a machine learning-based empirical approach would approximate the SDF with a nonparametric model such as a shallow neural network: $w(Z_t) \approx \sum_{p=1}^{P} \lambda_p S_p(Z_t)$, where each $S_p(Z_t)$ is some nonlinear basis function of Z_t. Thus, a neural network SDF delivers a linear factor pricing model,

$$M_{t+1} \approx 1 - \sum_{p} \lambda_p F_{p,t+1}, \tag{2}$$

where each "factor" $F_{p,t+1}$ is a characteristic-managed portfolio that uses the transformed "characteristics" $S_p(Z_t)$ as portfolio weights, and $\lambda = (\lambda_1, ..., \lambda_P)$ are parameters to be estimated.

This formulation captures the tension in machine learning asset pricing models. On the one hand, the more terms in the nonparametric representation of w (and thus the larger the number of factors), the better the model can approximate the true SDF. But as we

improve the approximation, we simultaneously increase the number of parameters that must be estimated, so when P becomes large, the estimated model may be unstable and suffer out-of-sample. Our theory makes it possible to understand (and quantitatively characterize) this cost/benefit tradeoff in large factor models.

On the cost side of the tradeoff is a phenomenon we refer to as "limits to learning." In low complexity settings—with many more observations than parameters to estimate—the law of large numbers kicks in, and appropriate estimators can recover the true model.[2] Traditional econometric theory deals with estimator behavior in this data-rich environment. By contrast, in high-complexity settings, the number of factors is large relative to the number of observations. Thus, the law of large numbers breaks down. Even correctly specified estimators fail to converge on the true model because there is not enough data to go around. The first key aspect of our theory is an explicit characterization of the limits to the learning effect. We prove that data limitations impose an unavoidable implicit shrinkage of the model, which depresses the model's expressive power. Complexity induces bias through its implicit shrinkage. But on the benefit side of the tradeoff, complexity also *reduces* specification bias. More factors lead to a more accurate approximation of the true SDF. Our main theoretical result shows that which effect dominates depends on the eigenvalue distribution of the factors. When there is a concentrated eigenvalue distribution and thus a few dominant factors, asset pricing model performance tops out after adding only a few factors, and there is no virtue of complexity—a counterfactual prediction in light of our empirical findings.

However, when the underlying factor structure is not too concentrated—that is, when the data-generating process aligns with the AIPT conjecture—a virtue of complexity emerges that closely mimics the patterns we find in the data. Our theory rationalizes our surprising empirical facts, demonstrating that even if arbitrage is absent and an SDF exists, it is possible to continually find new empirical "risk" factors that are unpriced by others and that

[2]I.e., if the model is correctly specified. If the model is mis-specified, estimators recover the nearest "pseudo-true" parameters (e.g. White, 1996).

adding these factors to the pricing model continually improves its out-of-sample performance. The theory also helps rationalize the prominence of "anomaly" portfolios in empirical asset pricing. The abundance of anomalies (the so-called "factor zoo") is not a puzzle to be solved or evidence of a corrupt research process.[3] Instead, it is the theoretically expected outcome in a complex asset pricing environment because data limitations hamper our ability to learn the true nature of markets. In fact, our theory argues that the extant factor zoo is *too small* and that an SDF model can be beneficially expanded to incorporate a teeming Noah's ark of factors by transforming raw asset characteristics into a wide variety of nonlinear signals (buttressed by appropriate shrinkage). Such a large factor set improves the out-of-sample SDF Sharpe ratio and reduces out-of-sample pricing errors.

1.3 Literature Review

This paper is related to an emergent literature that shows machine learning asset pricing models achieve higher out-of-sample Sharpe ratios and smaller pricing errors than their parsimonious predecessors. Examples include Kozak et al. (2020), Gu et al. (2020a), Chen et al. (2023), Bryzgalova et al. (2020), Cong et al. (2022), Fan et al. (2022) and Preite et al. (2022), among others. It also relates to machine learning methods for analyzing factor models, including Connor et al. (2012), Fan et al. (2016), Kelly et al. (2020), Lettau and Pelger (2020), Giglio and Xiu (2021), Giglio et al. (2022), and He et al. (2023) (see Kelly and Xiu (2023) and Rapach and Zhou (2020) for more complete survey treatments of these topics). We extend this literature with detailed documentation of the counterintuitive phenomenon that out-of-sample asset pricing model performance appears to improve with ever richer model specifications.

[3] Jensen et al. (2023) reach a similar conclusion based on the rationale that the risk-return trade-off is difficult to measure and complexity manifests as an inability to find a single silver-bullet characteristic that pins down expected returns. Instead, researchers gradually expand and refine the set of noisy signals and conclude "a more positive take on the factor zoo is not as a collective exercise in data mining and false discovery, but rather as a natural outcome of a decentralized effort in which researchers make contributions that are correlated with, but incrementally improve on, the body of knowledge."

An understanding of this surprising empirical phenomenon is only beginning to take shape. KMZ theoretically analyze machine learning models for return prediction. We extend this work in a number of ways, including reorienting the statistical objective around minimizing pricing errors and maximizing SDF Sharpe ratios, and by moving from a single asset to a cross-section of assets.[4] Together with Martin and Nagel (2021), Da et al. (2022), Fan et al. (2022), and KMZ, our paper belongs to an emergent literature analyzing "limits to learning" in high-dimensional asset pricing models. Our paper also relates to the theoretical machine learning literature on topics surrounding "double descent" and "benign overfit" (e.g. Spigler et al., 2019; Hastie et al., 2019; Belkin et al., 2020; Bartlett et al., 2020).

Through its coupling with the literature on factor pricing models, our work also relates to the empirical literature surrounding the "factor zoo" and factor replicability, including Harvey et al. (2016), McLean and Pontiff (2016), Hou et al. (2020), Feng et al. (2020), Jensen et al. (2023), and Chen and Zimmermann (2021). Our theory helps rationalize the continued discovery of factors that are unspanned by simpler precedent models. Relatedly, the demonstrated success of machine learning models in predicting the cross-section of returns such as Chinco et al. (2019), Han et al. (2019), Freyberger et al. (2020), Rapach and Zhou (2020), Gu et al. (2020b), Avramov et al. (2023), and Guijarro-Ordonez et al. (2021), is additional evidence of the virtue of complexity in financial markets research.

The remainder of the paper proceeds by introducing our empirical design in Section 2 and our empirical findings in Section 3. Section 4 presents our theoretical analysis of large factor models, and Section 5 concludes.

[4]From a technical standpoint, we overcome a number of new theoretical hurdles relative to KMZ. In time series regressions of KMZ, the random matrix behavior of time series signal covariances dictates the market timing strategies. In the panel problem, behavior is determined not just by time series covariances but also by the covariance of signals across assets. While standard random matrix theory in KMZ requires dealing with double limits (number of observations and number of parameters), our analysis requires development of a non-trivial theoretical extension to deal with a third limiting dimension—the number of base assets. Importantly, we also remove the equal ex-ante predictive power assumption of KMZ and allow for a generic distribution of risk premia across factors.

2 Empirical Framework

We now present the details of our empirical framework for large factor pricing models and their associated SDF representations.

2.1 Background: Factor Models and the SDF

Hansen and Richard (1987) show that a true SDF, if one exists, is representable as a tradable portfolio of risky assets as introduced in Equation (1).

A factor pricing model is equivalent to an SDF that is linear in the factors. We can, therefore, analyze and compare asset pricing factor models as competing SDF specifications. Motivated by the APT and the principle of parsimony,[5] the literature has primarily investigated (1) with tightly constrained factor specifications of the SDF weight function $w(Z_t)$. A leading example is the Fama and French (1993) model, which restricts M_{t+1} to be a three-parameter model:

$$w_i^{\text{FF}} = c_1 \text{MKT}_i + c_2 \text{SIZE}_i + c_3 \text{VALUE}_i,$$

where the weight of asset i in the Fama-French SDF (w_i^{FF}) depends only on the stock's weight in the market portfolio (MKT_i) and on researcher-dictated functions of the stock's size and book-to-market ratio (SIZE_i and VALUE_i).

The Fama-French SDF portfolio can equivalently (and more familiarly) be viewed as a linear combination of three factors. The factors are portfolios whose individual stock weights are given by MKT_i, SIZE_i, and VALUE_i, respectively. The SDF combines these factors with three parameters, c_1, c_2, and c_3, corresponding to the Sharpe ratio maximizing combination

[5] See, e.g., Tukey (1961) and Box and Jenkins (1970). Economic theory also motivates functional restrictions on the SDF (e.g. Hansen and Singleton, 1982). However, restrictions derived from economic theory have had limited success to date in pricing cross-sections of assets such as stocks, bonds, and derivatives. The Fama-French SDF and many other factor models in the literature are motivated by empirical "anomalies" vis-a-vis the CAPM and not by a particular economic theory.

of the factors. These are the only three parameters that must be estimated in the Fama-French SDF, resulting in an extraordinarily parsimonious asset pricing model. Thus, the Fama-French model is a special case of (1) where

$$w(Z_t^{FF}) = Z_t^{FF}\lambda^{FF}.$$

The conditioning set is Z_t^{FF}, which has $D = 3$ columns containing stocks' Fama-French characteristics, and the weight function is linear in parameters $\lambda^{FF} = (c_1, c_2, c_3)'$. In this case, the SDF is

$$M_t^{\text{FF}} = 1 - \lambda^{FF\prime} Z_t^{FF\prime} R_{t+1},$$

where we recognize that $Z_t^{FF\prime} R_{t+1}$ is a vector of time t returns on the three Fama-French factors and $1 - M_t^{\text{FF}}$ is their mean-variance efficient combination.

2.2 Designing A Large Factor Model

The Fama-French model achieves its parsimony from rigid assumptions on the SDF weighting function—discarding all but a few conditioning variables and making detailed choices for the functional form that maps the selected variables into SDF weights.

The AIPT takes a different approach, founded on the philosophy of machine learning, and explores the benefits of flexible models that accommodate many potential conditioning variables and diverse functional forms. In our analysis, Z_t is an $N_t \times D$ matrix that collects a large number of D conditioning variables for each risky asset. We derive a complex factor pricing model by replacing detailed specification choices with a generic nonparametric approximation to the unknown SDF:

$$w(Z_t) \approx \sum_{p=1}^{P} \lambda_p S_p(Z_t) = S_t\lambda, \tag{3}$$

where each $S_p(Z_t)$ is an $N_t \times 1$ nonlinear basis function of the input data Z_t and λ_p is a scalar basis coefficient. This basis expansion explores the space of potential shapes for $w(Z_t)$, then combines these building blocks with coefficients (λ_p) that best reconstruct the unknown function $w(Z_t)$. Collecting the P basis terms into an $N_t \times P$ matrix of characteristics $S_t = [S_{1,t}, ..., S_{P,t}]$, and collecting parameters $\lambda = [\lambda_1, ..., \lambda_P]'$, we arrive at $S_t\lambda$ as the nonparametric approximation to the true SDF weighting function. The larger is P, the broader the set of nonlinearities considered, the larger the number of model parameters, and the more flexible the approximation.

The nonparametric approximation in (3) takes the same basic form as the Fama-French SDF with S_t in place of Z_t^{FF}. The basis terms $S_p(Z_t)$ merely transform conditioning variables into a large number of new nonlinear asset "characteristics" analogous to those underpinning the Fama-French model. Substituting the approximation in (3) into the SDF definition (1), we arrive at

$$M_{t+1} \approx 1 - \lambda' S_t' R_{t+1} = 1 - \lambda' F_{t+1} = 1 - R_{t+1}^M. \tag{4}$$

In other words, like the Fama-French model, the SDF portfolio (denoted $R_{t+1}^M = \lambda' F_{t+1}$) is a linear combination of factor portfolios,

$$F_{t+1} = S_t' R_{t+1}. \tag{5}$$

Each factor $F_{p,t+1} = S_p(Z_t)' R_{t+1}$ is a characteristic-managed portfolio of risky assets that uses the nonlinear asset "characteristic" $S_p(Z_t)$ as the vector of portfolio weights, and again λ is the mean-variance efficient combination of factors.[6] The difference vis-a-vis Fama-French (and other parsimonious models) is that the factors in (4) make flexible and nonlinear use of conditioning data, and there are a large number P of these factors.

[6]Because the size of the cross-section varies over time, our empirical analysis in fact scales the factors as $F_{t+1} = N_t^{-1/2} S_t' R_{t+1}$ to maintain a relatively consistent scale of factor returns over time.

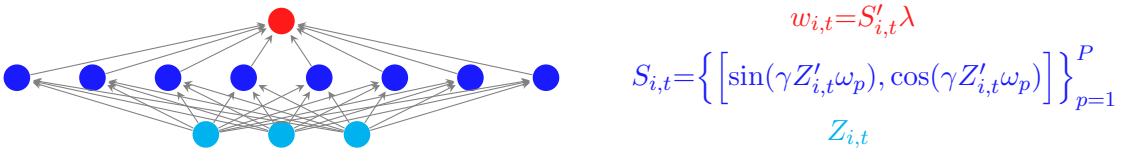

Figure 1: Neural Network Representation of Random Fourier Factors

Note. Illustration of nonlinear mapping from stock i's characteristics $Z_{i,t}$ into hidden layer neurons $S_{i,t}$ and then into the conditional SDF weight $w_{i,t}$ during the construction of random Fourier factors.

2.3 Random Fourier Factors

To operationalize this large factor model approximation of the SDF, our empirical approach adapts the method of random Fourier features, or RFF (Rahimi and Recht, 2007). While (Rahimi and Recht, 2007) propose RFF for predictive regression, our approach embeds RFF in an SDF. The nonlinear basis functions in this method are trigonometric transformations of the original signals Z_t:

$$[S_{2p-1,t}S_{2p,t}] \in \mathbb{R}^{N_t\times 2} = [\sin(\gamma Z_t\omega_p),\cos(\gamma Z_t\omega_p)]', \quad \omega_p \sim \text{i.i.d. } N(0,I), \quad p=1,...,P/2. \tag{6}$$

Each RFF basis function forms a random linear combination (ω_p) of the raw characteristics Z_t, then feeds this through sine and cosine activation functions.[7] The transformation is applied stock-by-stock, converting stock i's characteristics into new nonlinear characteristics that include not just transformations of each individual characteristic but general multi-way interactive effects involving all characteristics for stock i.

Upon closer inspection, we see that the SDF in (4) is, in fact, a two-layer neural network with input parameters given by ω_p, output parameters given by λ, and trigonometric activation neurons (see Figure 1). In a typical neural network, both the input and output parameters are estimated using computationally costly numerical methods. The fascinating

[7]The parameter γ controls the Gaussian kernel bandwidth in generating random Fourier features. Following Kelly et al. (2022), we randomly chose γ from the grid $[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$ for each ω_i that we generate. This embeds varying degrees of nonlinearity in the generated feature set S_t.

insight of (Rahimi and Recht, 2007) is that, by using randomization to fix the input parameters and only estimating the output parameters, random features regression can build nonparametric approximations in a computationally efficient manner.[8] Building on this insight, our SDF model is estimable in closed form using only least squares regression (as discussed below).

RFF is an ideal tool for our analysis because, from a fixed input data set Z_t we can create asset pricing models with any desired factor dimension, P. For a low-dimensional model of, say, $P = 2$, one generates a single pair of RFFs. For a high-dimensional model of, say, $P = 10{,}000$, one can instead draw many random weight vectors ω_p, $p = 1, ..., 5{,}000$. This gives us the ability to evaluate how asset pricing models are impacted by varying the number of factors *while holding the set of conditioning variables fixed*. It is important to emphasize that in our analysis, factor proliferation arises not from expanding the set of input data Z_t (which may or may not be low-dimensional) but instead from generating many nonlinear basis transformations of Z_t.

2.4 Estimator

The SDF coefficient vector λ represents the mean-variance efficient portfolio of factors. Note that the factor representation of the *conditional* SDF in essence produces an equivalent *unconditional* SDF. This allows us to estimate λ as the sample Markowitz portfolio of factors:

$$\arg\min_{\lambda} \left\{ \hat{E}[\lambda' F_t] - \frac{1}{2}\hat{E}\left[(\lambda' F_t)^2\right] \right\} = \hat{E}[F_t F_t']^{-1}\hat{E}[F_t] \tag{7}$$

[8] Indeed, RFF is a special case of kernel regression. For applications of kernel regression in asset pricing settings, see Kozak (2020) and Kelly et al. (2024a).

where $\hat{E}$ denotes the training sample mean. We denote the corresponding estimated SDF and SDF portfolio as

$$\hat{M}_t = 1 - \hat{R}_t^M \quad \text{and} \quad \hat{R}_t^M = \hat{\lambda}' F_t. \tag{8}$$

The AIPT explores large factor models in which the dimension of λ may exceed the number of time series observations, i.e., $P > T$. In this case, $\hat{E}[F_t F_t']^{-1}$ is rank-deficient, which we resolve by introducing a ridge penalty in the Markowitz problem:

$$\hat{\lambda}(z) = \arg\min_{\lambda} \left\{ \hat{E}[\lambda' F_t] - \frac{1}{2}\hat{E}\left[(\lambda' F_t)^2\right] + z\lambda'\lambda \right\} = \left(zI + \hat{E}[F_t F_t'] \right)^{-1} \hat{E}[F_t] \tag{9}$$

where z is the ridge parameter.

Kozak et al. (2020) point out that we can also interpret $\hat{\lambda}(z)$ in terms of "pricing errors." Since the factors F_t are tradable assets, the true SDF M_t prices these factors with zero error due to the marginal investor's first-order optimality condition:

$$E[M_t F_t] = 0. \tag{10}$$

Hansen and Jagannathan (1997) suggest a statistic for comparing the efficacy of SDF models in terms of their pricing error magnitudes. For a candidate SDF $\tilde{M}_t$, The Hansen-Jagannathan distance (HJD) is a weighted sum of squared pricing errors for the set of test assets F_t:[9]

$$\mathcal{D}^{HJ} = E[\tilde{M}_t F_t]' E[F_t F_t']^{-1} E[\tilde{M}_t F_t]. \tag{11}$$

Substituting the SDF approximation model (4) for $\tilde{M}$ in the HJD and adding a ridge penalty

[9] We discuss some attractive properties of the HJD in the following sections.

to handle rank deficiency when $P > T$, we have

$$\hat{\lambda}(z) = \arg\min_{\lambda} \left\{ \hat{E}[(1 - \lambda' F_t) F_t]' \hat{E}[F_t F_t']^{-1} \hat{E}[(1 - \lambda' F_t) F_t] + z\lambda'\lambda \right\}. \tag{12}$$

In other words, $\hat{\lambda}(z)$ is also the regularized SDF estimator that minimizes pricing error (we discuss HJD in further detail in Section 2.6).

2.5 Data and Training

To make the conclusions from this analysis as easy to digest as possible, we perform our analysis in a conventional setting with conventional data. We use a comprehensive and standardized sample of monthly US stock returns and 153 stock characteristics from 1963 to 2023, compiled by Jensen et al. (2023) (JKP henceforth).[10] As in JKP, our universe includes NYSE/AMEX/NASDAQ securities with CRSP share code 10, 11, or 12, excluding "nano" stocks (i.e., stocks with market capitalization below the first percentile of NYSE stocks).

Some of the JKP characteristics have low coverage, especially in the early parts of the sample. To ensure that characteristic composition is fairly homogeneous over time and to avoid purging a large number of stock-month observations due to missing data, we reduce the 153 characteristics to a smaller set of $D = 130$ characteristics with the fewest missing values. We drop stock-month observations for which more than 30% of the 130 characteristic values are missing and use N_t to denote the number of the remaining stock observations at time t. Next, we cross-sectionally rank-standardize each characteristic and map it to the $[-0.5, 0.5]$ interval, following Gu et al. (2020b). Ultimately, we obtain an $N_t \times D$ matrix of conditioning characteristics Z_t in each month.

In our empirical analyses, we study SDF performance as we vary two aspects of the model: the number of factors P and ridge penalty z. For each model of a given size and shrinkage, we conduct a rolling out-of-sample model performance analysis. Starting in January 1993, in

[10] To access the stock-level characteristic data and associated documentation, refer to `jkpfactors.com`.

each month t, we use the most recent 360 months of data to estimate the ridge SDF parameter in (9).[11] We then track the out-of-sample SDF portfolio return in the subsequent month. From the sequence of monthly out-of-sample SDF realizations, we calculate out-of-sample SDF performance metrics.

2.6 Performance Metrics

We evaluate our SDF models with two standard out-of-sample performance metrics. A true SDF is the mean-variance efficient portfolio of risky assets. Thus, our first performance metric is the SDF portfolio Sharpe ratio. For a candidate SDF $\tilde{M}_t = 1 - \tilde{R}_t^M$, the annualized out-of-sample Sharpe ratio is

$$\widehat{SR} = \sqrt{12}\frac{\hat{E}_{OS}[\tilde{R}_t^M]}{\hat{\sigma}_{OS}(\tilde{R}_t^M)}$$

where $\hat{E}_{OS}$ and $\hat{\sigma}_{OS}$ denote sample average and standard deviation among the T_{OS} out-of-sample observations (to differentiate versus the notation $\hat{E}$ and T used for training samples). Since the SDF is constructed from excess returns on risky assets, the numerator need not be adjusted for the risk-free rate.

Our second performance metric is the pricing error for a large set of test assets. The test assets in our main analysis are the 153 JKP anomaly factors constructed by those authors and downloadable at `jkpfactors.com` (robustness analyses report pricing errors for other test asset sets as well). To aggregate test asset pricing errors into a single metric, we calculate the normalized sum of squared pricing errors according to the Hansen-Jagannathan distance (HJD) referenced in Section 2.4. For a set of test asset returns $R_t^{\mathcal{T}}$, the out-of-sample HJD

[11] The stochastic nature of RFF means that there is inherent variability in the estimated SDF model when P is small. To mitigate this variability, we repeat the RFF-based estimation 20 times with different random seeds and report average performance metrics across seeds. This has little qualitative effect on the plots shown. See KMZ for further discussion of this point.

is

$$\hat{\mathcal{D}}^{HJ} = \hat{E}_{OS}\left[\tilde{M}_t R_t^{\mathcal{T}}\right]' \hat{E}_{OS}\left[R_t^{\mathcal{T}} R_t^{\mathcal{T}\prime}\right]^{+} \hat{E}_{OS}\left[\tilde{M}_t R_t^{\mathcal{T}}\right], \tag{13}$$

where $^{+}$ is the Moore-Penrose quasi-inverse (necessary due to the degeneracy of $\hat{E}_{OS}[R_t^{\mathcal{T}} R_t^{\mathcal{T}\prime}]$ when $P > T_{OS}$).

The HJD has a number of attractive model comparison properties.[12] It averages pricing errors among test assets using a common weighting matrix for all candidate models of M_t. This is important because it puts all models on equal footing for comparison, unlike other alpha or GMM-based comparisons. The weighting matrix is economically motivated since it gives the HJD a clear interpretation as the pricing error of the portfolio of test assets that is *most mispriced* by each model. Furthermore, the HJD is scale invariant.[13] While typically used for in-sample comparison, the HJD easily generalizes for out-of-sample evaluation because it avoids the need to estimate out-of-sample time series alphas and betas for each test asset. Finally, because our theoretical derivations explicitly characterize the expected out-of-sample HJD for complex SDF models, the empirical HJD can be directly compared to theoretical predictions.

2.7 Benchmark Models

We provide reference points for our analysis by comparing large factor model performance to five benchmark factor pricing models in the literature:[14]

FF6: This is a six-factor model that includes the five factors of Fama and French (2015) plus their UMD momentum.

[12] In addition to Hansen and Jagannathan (1997), the literature including Kan and Robotti (2009), Chen and Ludvigson (2009), and Kelly and Xiu (2023) further advocates HJD as a pricing error metric.

[13] Differences in volatilities among factors can skew test statistics like average absolute alpha or the GRS statistic, and as a result, the volatility scaling choice for test assets can lead to different conclusions. HJD is invariant to changes in test asset volatility.

[14] All benchmark factor data are conveniently available from the respective authors' websites.

SY: This is a four-factor model that includes the "mispricing" factors of Stambaugh and Yuan (2017).

HXZ: This is a five-factor model that includes the q-factor model specification of Hou et al. (2015) augmented to include the expected growth factor of Hou et al. (2021).

DHS: This is a three-factor model that includes the long-horizon and short-horizon behavioral factors of Daniel et al. (2020).

BS: This is a composite model that selects factors from the literature based on the Bayesian methodology of Barillas and Shanken (2018). Its six factors include the market factor, the investment and profitability factors from Hou et al. (2015), the size and momentum factors from Fama and French (2015), and the value factor (HML^m) from Asness et al. (2013).

To compare models, we use the SDF corresponding to each of these factor pricing models, which requires estimating the Markowitz portfolio of factors in each model. We do this in rolling 360-month training windows and track benchmark performance out-of-sample. In other words, we use the same training/evaluation design for the benchmarks that we use for our large factor pricing models.[15]

3 Empirical Results

3.1 The Virtue of Complexity in Factor Pricing Models

Our central empirical finding is that factor model performance is increasing in the number of factors. We illustrate this virtue of complexity by plotting out-of-sample model performance

[15] When building the Markowitz portfolio for each benchmark model, we do not use ridge shrinkage because the number of assets is far smaller than the number of time series observations. Even with infeasible (ex-post optimal) ridge shrinkage, we find the out-of-sample SDF performance of benchmark models changes negligibly.

as a function of the number of model parameters (which, in our setting, corresponds to the number of factors). KMZ refer to such plots as "VoC curves." Each curve applies a different ridge penalty $z = 10^{-5}, 10^{-3}, 10^{-1}, 1,$ or 10. Along each curve, we vary the complexity—i.e., the number of factors—in our SDF model. We consider specifications with $P = 36, ..., 360{,}000$ random Fourier factors, corresponding to $c = 0.1, ..., 1{,}000$. The VoC curves in Figure 2 show four out-of-sample properties of each SDF specification: expected return (Panel A), volatility (Panel B), Sharpe ratio (Panel C), and pricing error (Panel D).

We find the following main patterns associated with model complexity. As we increase the number of factors, i) the average SDF return rises, ii) SDF volatility rises, iii) the SDF return increases faster than volatility, resulting in a rising SDF Sharpe ratio, iv) pricing errors decline, and v) performance gains accrue more quickly with less ridge shrinkage. The exceptions to these patterns occur when ridge shrinkage is small, and complexity is close to one, in which case we see a spike in out-of-sample SDF volatility that induces an abrupt dip in the Sharpe ratio and a spike in pricing error. This "double ascent" in Sharpe ratio and "double descent" in pricing error is reminiscent of previously documented phenomena in the machine learning literature (e.g. Hastie et al., 2019; Bartlett et al., 2020).

Quantitatively, the differences in out-of-sample performance for high and low-complexity models are dramatic despite the fact that both model types use identical data inputs and differ only in the richness of their parameterizations. High-complexity models with light shrinkage earn out-of-sample Sharpe ratios near 3.7, roughly 2.6 times larger than the Sharpe ratios of low-complexity models with $c < 1$. Likewise, large factor models are much better positioned to price our demanding test assets (153 anomaly portfolios collected from decades of finance research). Pricing errors for low-complexity models are more than twice as large as those for high-complexity models (0.54 vs 0.22).

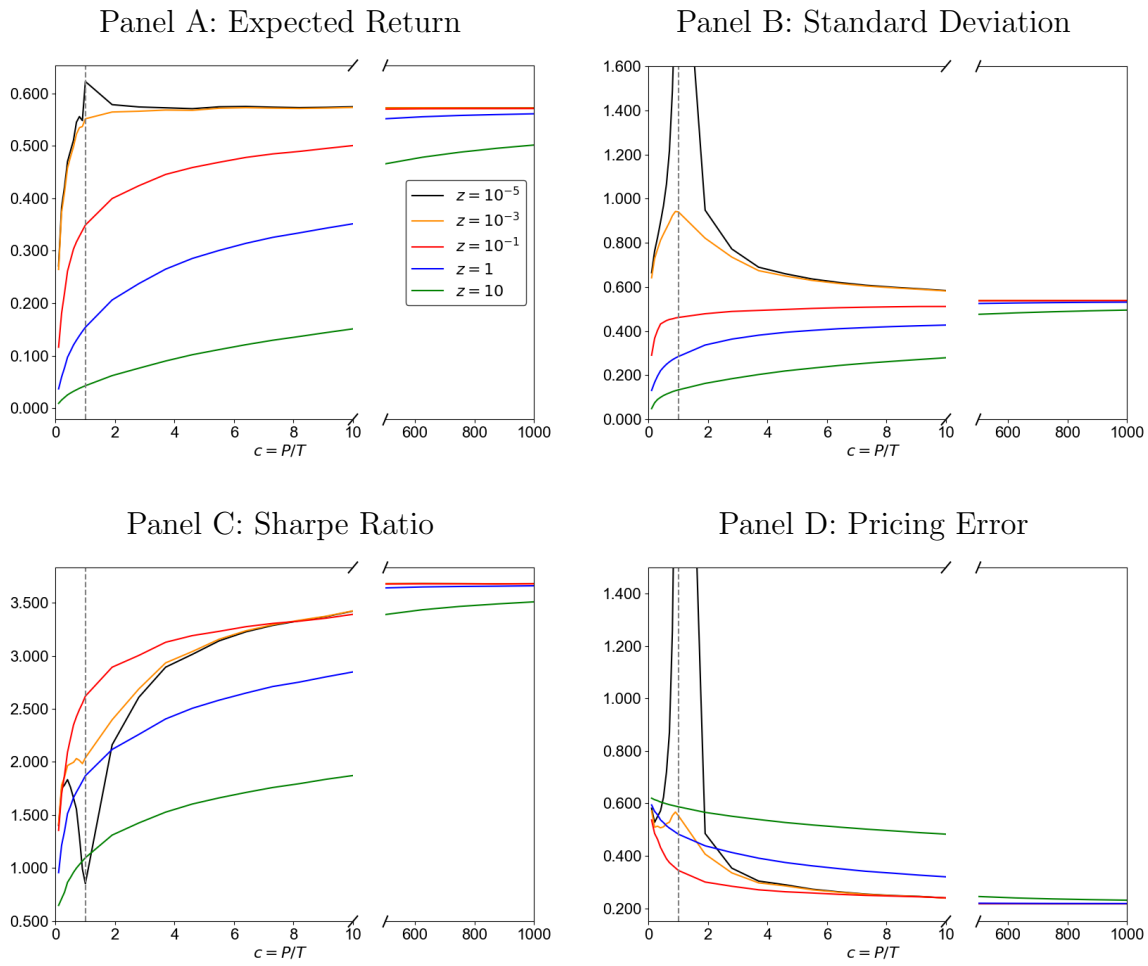

Figure 2: Out-of-sample Performance of Complex Factor Models

Note. Realized out-of-sample SDF portfolio average return, standard deviation, Sharpe ratio, and pricing error (HJD). The horizontal axis shows model complexity $c = P/T$, with P ranging from 36 to 360,000 and $T = 360$ months. Factors underlying the SDF are characteristic-managed portfolios constructed with random Fourier features derived from JKP stock characteristics. The test assets in Panel D are 153 anomaly factors from JKP.

3.2 Comparison With Benchmarks

In Figure 2, we compare large and small variants of factor models using random Fourier factors. However, the conclusion that large factor models outperform simple models is not specific to the nonparametric framework in Figure 2.

Figure 3 compares the high complexity factor model ($P = 360,000$ and $z = 10^{-5}$) to

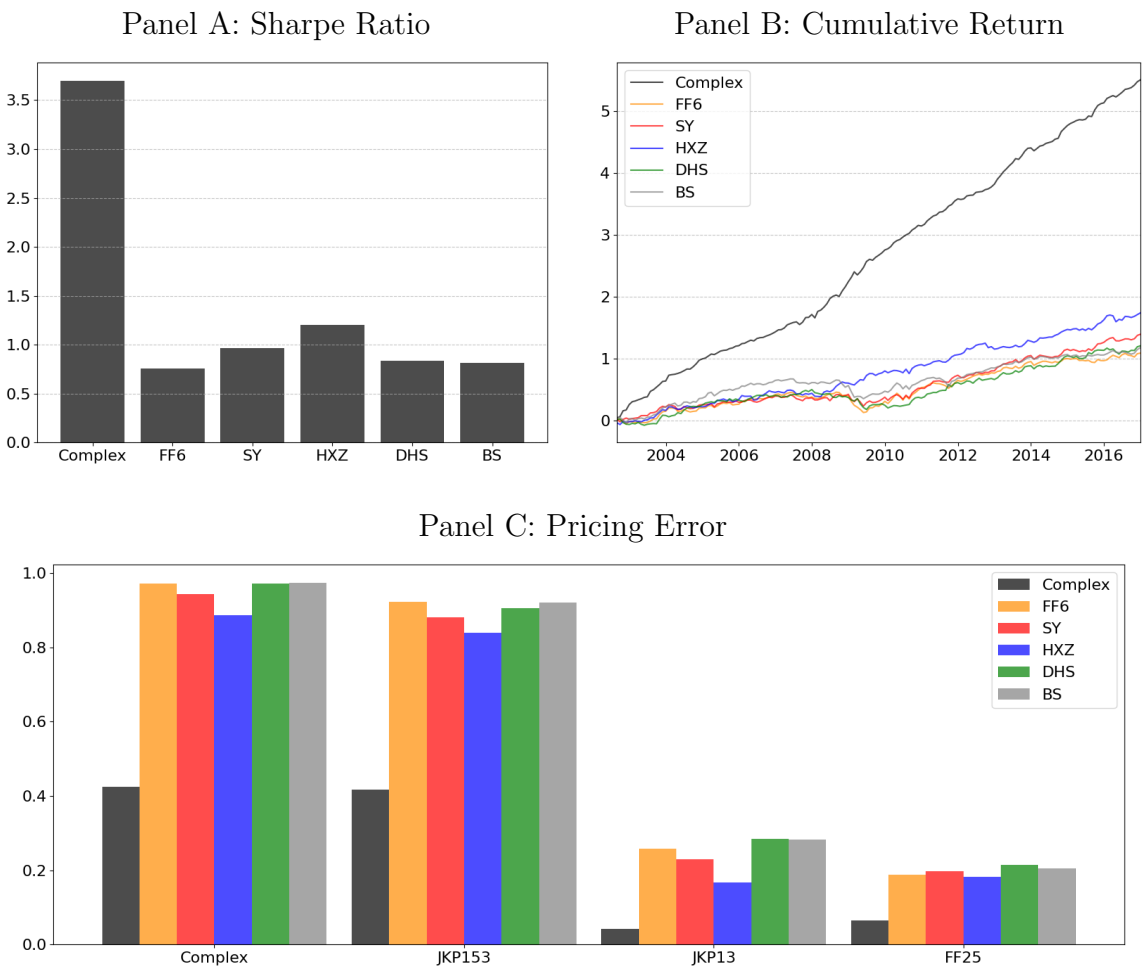

Figure 3: Performance Comparison of Complex and Benchmark Factor Models

Note. This figure reports the out-of-sample performance of SDF portfolios corresponding to various factor pricing models, including the complex model ($P = 360,000$ and $z = 10^{-5}$) as well as the simple benchmark models FF6, HXZ, SY, DHS, and BS. Panel A shows the out-of-sample Sharpe ratio for each model's factor Markowitz portfolio. Panel B shows the simple cumulative sum of out-of-sample returns for each model's factor Markowitz portfolio (we have standardized out-of-sample returns to 10% annualized volatility for all models in order to facilitate comparison in this panel). Panel C shows pricing errors (HJD) for four sets of test assets, including are the 360,000 random Fourier factors underlying the large factor model (denoted "complex"), the 153 JKP factors, the 13 JKP aggregate theme factors, and the 25 Fama-French size and book-to-market portfolios. Benchmark data sets are available for different sample periods, so we report statistics on the intersection of their out-of-sample ranges, which is from August 2002 to December 2016.

five of the leading factor models in the literature: FF6, HXZ, SY, DHS, and BS. Panel A

reports the out-of-sample Sharpe ratio for each model's SDF portfolio, Panel B shows the corresponding cumulative return plot, and Panel C reports out-of-sample pricing errors.[16]

Panel A shows that the benchmark factor models are all roughly on par with each other in terms of Sharpe ratio, while the complex model rather dramatically outperforms every simple model. Its Sharpe ratio is around 3-6 times higher than the benchmarks. The cumulative returns in Panel B show that the portfolio performance comparison is not driven by any particular subsample and that the complex model does not exhibit any periods of unusually turbulent behavior.

Panel C shows that the large factor model also dominates in terms of pricing error. When the 153 JKP factors are test assets, HJD for the benchmark models hovers around 0.9, roughly twice that of the large factor model. This finding is not specific to this set of test assets. When the test assets are the 360,000 random Fourier factors themselves, the HJD hovers around 0.95 for benchmark models, again roughly twice the HJD of the complex factor model. Third, we study the JKP theme factors, which collect the individual anomaly factors into 13 aggregate theme portfolios. The benchmark HJD for this test set is around 0.2, which is roughly four times larger than the complex model HJD. Finally, we use the classic Fama-French 25 size and value portfolios. This test set has been the target for pricing models for forty years, and one can argue that benchmark models have been honed to price test assets like these. Even in this case, benchmark models deliver an out-of-sample pricing error of around 0.2, roughly three times larger than those from the complex pricing model.

Table 1 extends the benchmark comparison by reporting pairwise alphas between SDF portfolios of the large factor pricing model and each of the benchmark models. For this, we use the out-of-sample SDF portfolios but renormalize them ex post to a 10% annualized volatility to make alphas comparable for all models (unlike HJD, alphas lack scale invariance, so volatility scaling is helpful for comparison). The complex model produces a large and

[16]Figure 3 shows small differences in performance metrics for the high complexity factor model compared to Figure 2. This is because Figure 3 is restricted to a shorter time sample in which data is available for all benchmark models.

Table 1: Performance Comparison of Complex and Benchmark Factor Models

This table reports a comparison of out-of-sample SDF portfolio performance for various factor pricing models, including the complex model ($P = 360,000$ and $z = 10^{-5}$) as well as the simple benchmark models FF6, HXZ, SY, DHS, and BS. The first two rows show the annualized percentage alpha and associated t statistic of the complex SDF portfolios versus each benchmark SDF portfolio. The next two rows show each benchmark SDF portfolio's alpha and t statistic versus the complex model. The last two rows show the ex-post tangency portfolio weights combining the SDF portfolios of all models with and without a non-negativity constraint (and constraining weights to sum to one). To facilitate alpha comparisons, out-of-sample SDF portfolio returns for all models are rescaled to have 10% annualized volatility. Benchmark data sets are available for different sample periods, so we report statistics on the intersection of their out-of-sample ranges, which is from August 2002 to December 2016.

	LHS					
	FF6	SY	HXZ	DHS	BS	Complex
Complex Alpha	42.8	40.7	39.4	43.5	44.4	–
vs. Benchmark	(13.9)	(13.6)	(13.0)	(13.7)	(13.8)	–
Benchmark Alpha	-2.8	-4.3	-2.3	1.50	3.90	–
vs. Complex	(-0.7)	(-1.2)	(-0.6)	(0.40)	(1.00)	–
Tangency Weights						
Unconstrained	-0.44	-0.24	-0.01	0.12	0.56	1.01
Non-negative	0.00	0.00	0.00	0.01	0.09	0.90

highly significant alpha versus all benchmarks, ranging from 39.4% to 44.4% per annum. The reverse is not true. Benchmark models all have small and statistically insignificant alphas versus the complex model. Finally, we estimate the "meta"-Markowitz portfolio that combines SDF portfolios of all individual models. The complex model dominates the meta portfolio, receiving a weight of 101% in the unconstrained portfolio and a weight of 90% in the portfolio that constrains weights on individual models to be non-negative.

3.3 The Nonlinear Fama-French Model

In the preceding analysis, we compare high complexity factor models derived from 130 JKP characteristics to the FF6 model and other benchmarks that are low complexity models derived from a small set of characteristics. Naturally, one may wonder about the benefits of complexity when we restrict the raw data inputs to match those used by a given benchmark

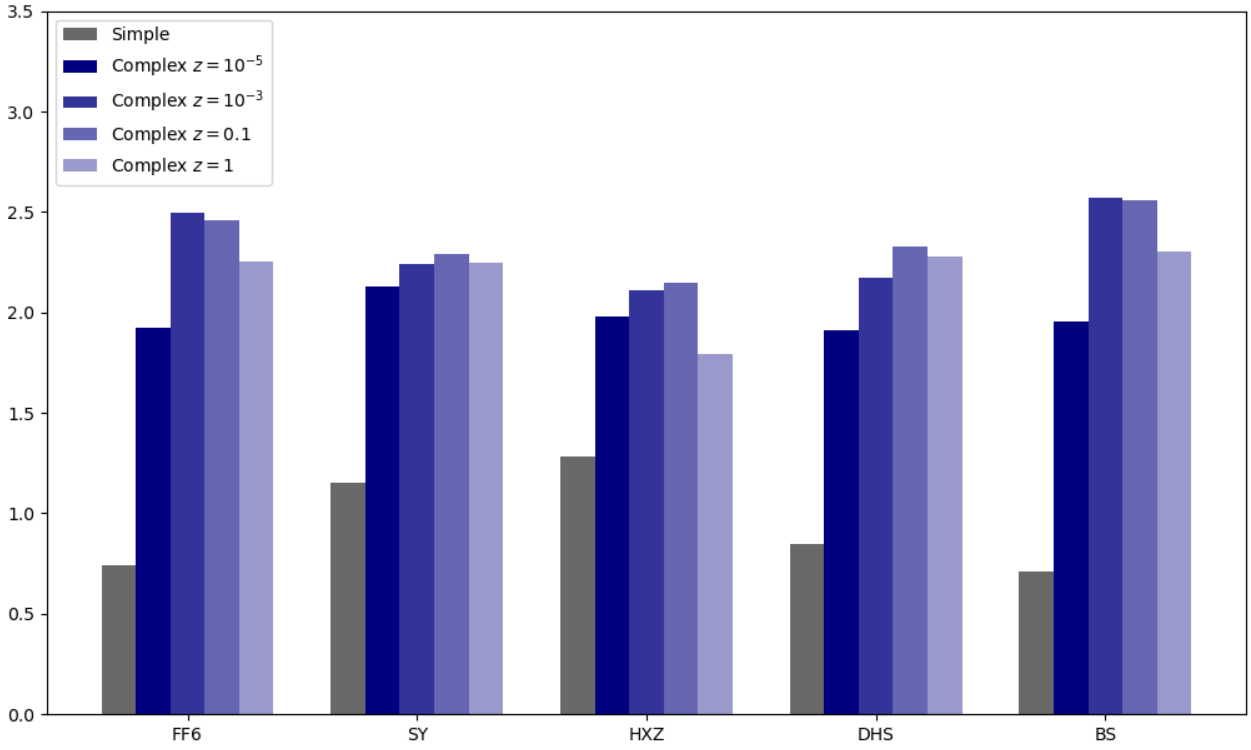

Figure 4: Complex Versions of Benchmark Models

Note. The figure reports out-of-sample Sharpe ratios of simple benchmark models compared to their nonlinear, high-complexity counterparts. For each benchmark model we construct its high complexity SDF with random Fourier factors derived from the characteristics underlying that model: size, value, investment, profitability, and momentum for FF6; size and two mispricing measures (MGMT and PERF) for SY; size, investment, profitability and expected growth for HXZ; size, PEAD, and two measures of issuance (CSI and NSI) for DHS; and asset growth, profitability, size, value and momentum for BS. Complex versions of each benchmark use $P = 360,000$ factors ($c = 1000$) and shrinkage of $z = 10^{-5}, 10^{-3}, 10^{-1}$ or 1.

model. To investigate this, we construct complex SDFs using random features derived from *only* the characteristics that underly each benchmark. For example, in the case of FF6, we construct a complex factor model with a large number of nonlinear factors derived from the size, value, investment, profitability, and momentum characteristics.

Figure 4 reports the out-of-sample Sharpe ratio of high complexity variations of each benchmark model using $P = 360{,}000$ ($c = 1{,}000$) random Fourier factors and ridge shrinkage of $z = 10^{-5}, 10^{-3}, 10^{-1}$ or 1. The original formulation of the FF6 model has an out-of-sample Sharpe ratio of 0.74, while its high complexity version ranges from 1.93 to 2.50

depending on the degree of shrinkage. All benchmarks show the same pattern—if we hold the stock characteristics of each simple model fixed but enrich the model specification to incorporate highly parameterized nonlinear transformations of those characteristics, the out-of-sample performance of the benchmarks improves by a factor of 1.4 to 3.6 depending on the benchmark and the amount of shrinkage.

The conclusion is that the benefits of large factor models accrue even when starting from a small conditioning information set. Complexity may be applied to any set of conditioning variables to more flexibly reflect patterns in the underlying return-generating process.

3.4 SDF Complexity or SDF Sparsity?

A recent line of financial machine learning research suggests that it is possible to estimate a successful factor pricing model through the imposition of sparsity.[17] The evidence indicates that an SDF model with a small number of factors can successfully price a wide variety of test assets. This is exemplified by Kozak et al. (2018), who study a collection of difficult-to-price anomaly portfolios that serve as their test assets. They then show that a simple linear SDF—comprised of just a few principal components of the anomaly portfolios—is powerful for pricing their entire anomaly cross-section.

The notion of SDF sparsity appears at odds with the benefits of complexity that we document above. While our results thus far demonstrate that a complex model can identify efficacious nonlinear pricing factors, is it possible that we also introduce unnecessary redundancy by using many thousands of factors? We investigate this possibility now.

In Figure 2, each point on each curve is a model with a particular number of factors, P, and a particular shrinkage parameter, z. To investigate the potential benefits of sparsity, we replace the P-factors in each model of Figure 2 with a small number K of principal

[17] This includes Gagliardini et al. (2016), Kelly et al. (2020), Kozak et al. (2020), Lettau and Pelger (2020), and Giglio and Xiu (2021), among others.

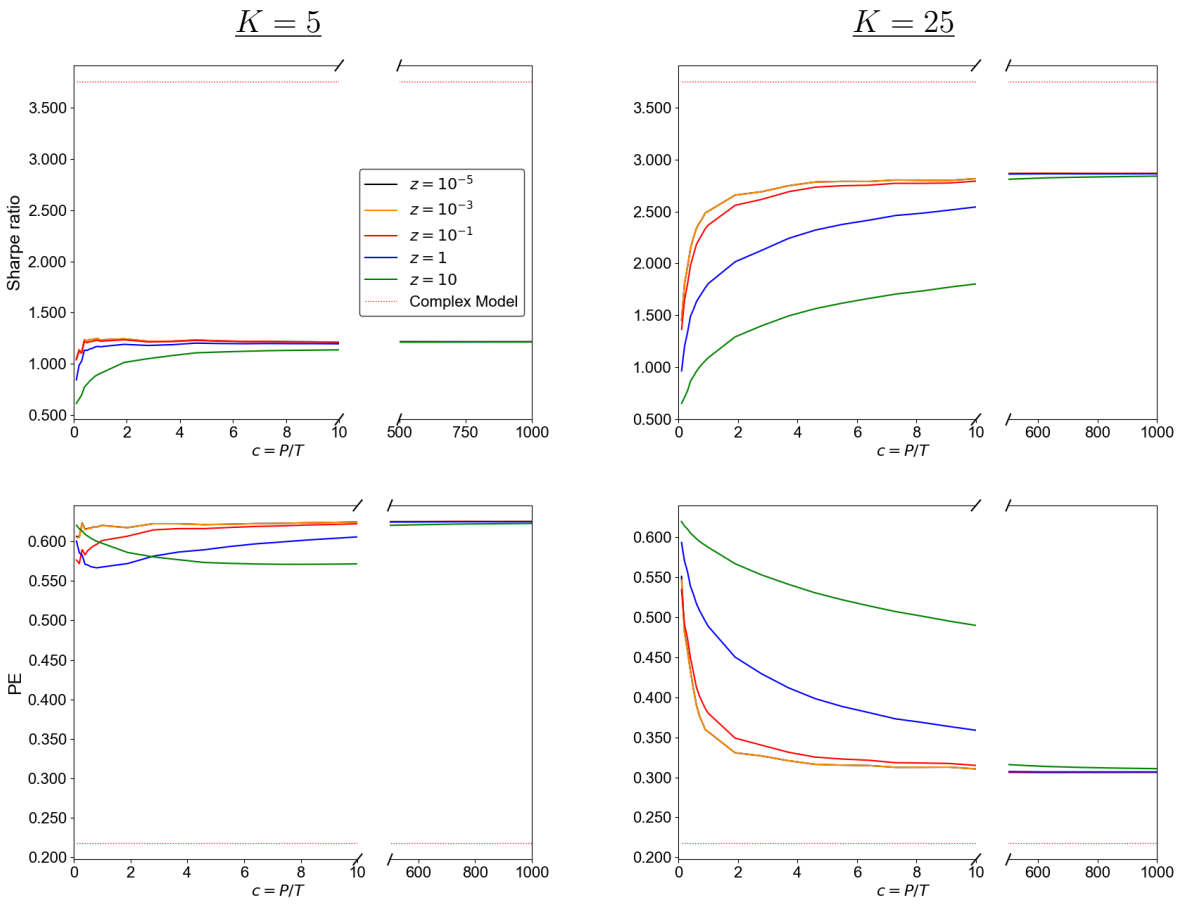

Figure 5: The Effect of Sparsity on Out-of-sample SDF Performance

Note. Realized out-of-sample SDF Sharpe ratio (top row) and pricing error (HJD, bottom row) for complex models with dimension reduction to $K = 5$ (left column) and $K = 25$ (right column) principal components. The horizontal axis shows model complexity $c = P/T$ with P ranging from 36 to 360,000 and $T = 360$ months. For ease of reference, "Complex Model" shows performance of the highest complexity model ($P =$ 360,000, $z = 10^{-5}$) from Figure 2 without dimension reduction.

components derived from those P factors. We then re-estimate the SDF coefficients on the K components and track the resulting out-of-sample factor model performance.

Figure 5 reports the results. In the left column, we consider a $K = 5$ component dimension reduction of each complex factor model (corresponding to the number of components in the main analysis of Kozak et al., 2018). The Sharpe ratio (top panel) of the dimension-reduced SDF is roughly 1.3, compared to 3.7 for the full complexity model. Likewise, pricing errors (bottom panel) are 0.57 for the $K = 5$ SDF versus 0.22 for the full complexity

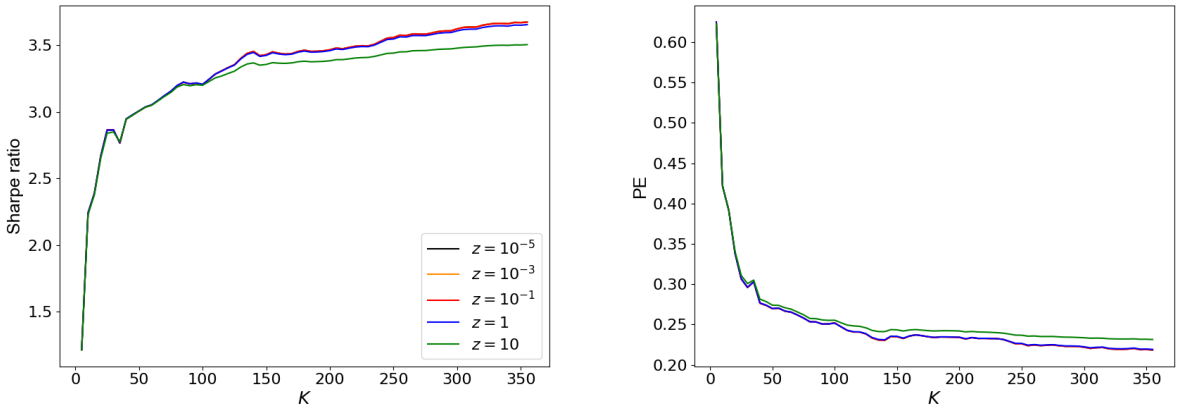

Figure 6: The Effect of Sparsity on Out-of-sample SDF Performance

Note. Realized out-of-sample SDF Sharpe ratio (top row) and pricing error (HJD, bottom row) for complex models with dimension reduction to $K = 5, ..., 360$ principal components (x-axis).

model. In other words, imposing sparsity on the SDF via dimension reduction inhibits SDF performance relative to the unreduced, high-complexity counterpart.

The right column of Figure 5 shows that increasing model complexity by using $K = 25$ rather than $K = 5$ principal components leads to a large improvement in out-of-sample performance. But even in this case, the Sharpe ratio remains well below that of the full complexity model (3.7 vs 2.9), and the pricing error remains 41% higher (0.31 vs 0.22).

Two properties of high-dimensional models drive this effect. First, even if the true (unobservable) factor covariance matrix has a few large eigenvalues and, hence, a strong factor structure, the factors become difficult to detect when complexity is high.[18] Second and more surprisingly, even low-variance components have a significant Sharpe ratio, so excluding them leads to a large drop in out-of-sample SDF performance. This fact is illustrated in Figure 6. In this analysis, we hold the number of factors in the underlying model fixed at $P = 360,000$, but vary the number of components extracted from the model from $K = 5, 10, ..., 360$. Adding low variance components typically helps, and never hurts, out-of-sample factor model performance. Based on this analysis, the AIPT's large factor model

[18] See, e.g., Lettau and Pelger (2020).

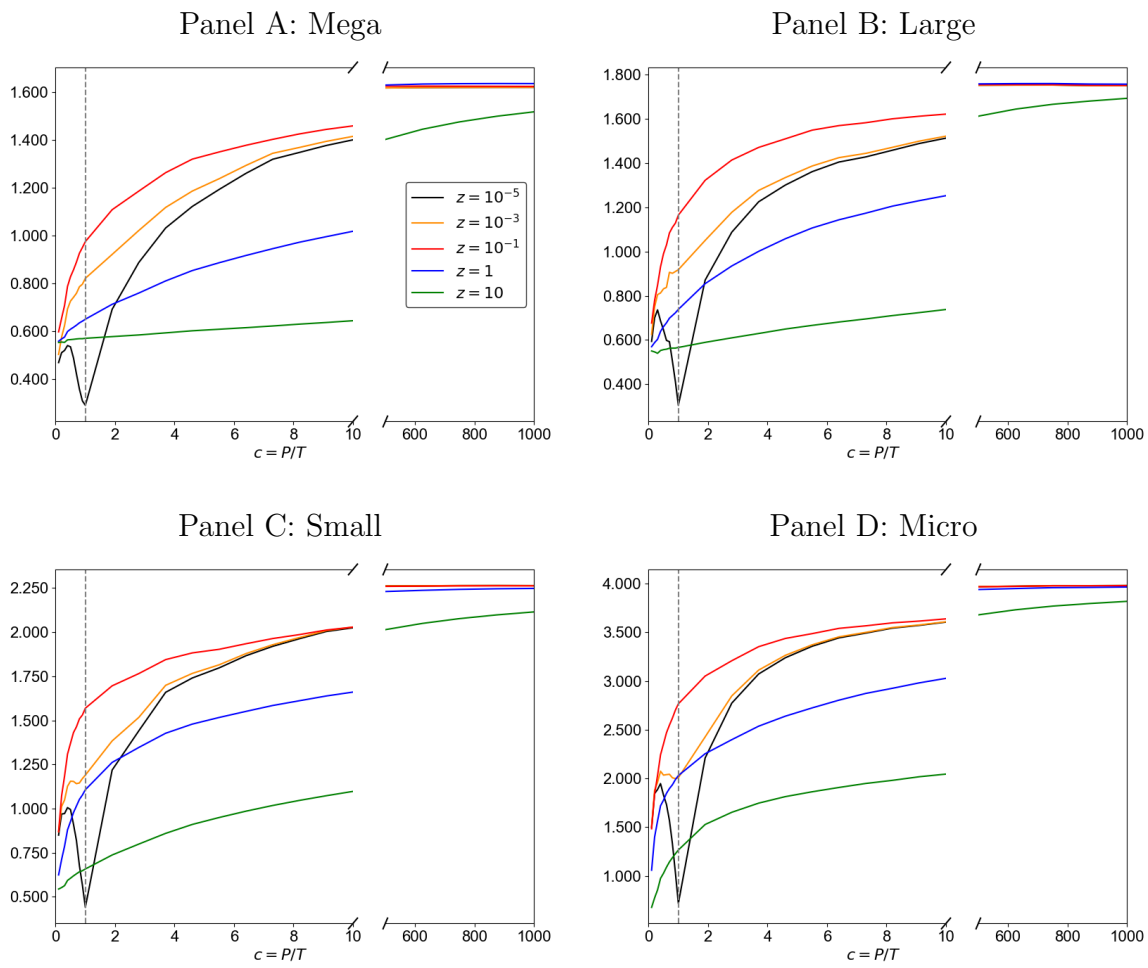

Figure 7: Out-of-sample Sharpe Ratios of Complex SDF Models By Size Group

Note. Realized out-of-sample SDF Sharpe ratio for SDFs estimated from subsamples based on market capitalization.

conjecture appears better suited for describing market behavior than does the parsimony conjecture of the APT.

3.5 Sensitivity to Liquidity

A Sharpe ratio of 3.7 for the high-complexity SDF model suggests that the model is exploiting inefficiencies associated with illiquidity. To understand the role of complexity in factor pricing models while abstracting from the question of asset liquidity, we conduct our empirical

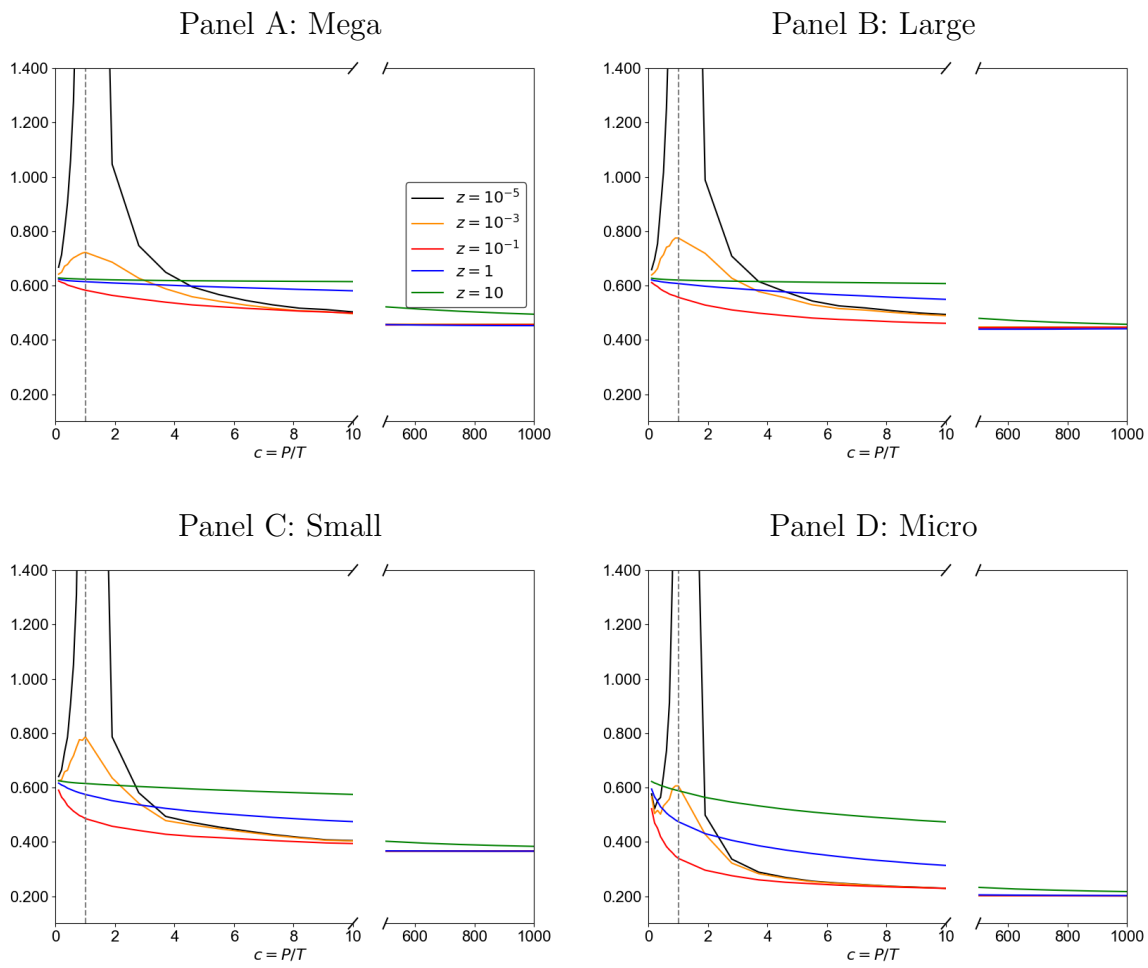

Figure 8: Out-of-sample Pricing Errors of Complex SDF Models By Size Group

Note. Realized out-of-sample SDF pricing error (HJD) for SDFs estimated from subsamples based on market capitalization.

analysis separately for different market capitalization groups. We study four size groups from JKP: mega (largest 20% of stocks based on NYSE breakpoints each period), large (between 80% and 50%), small (between 50% and 20%), and micro (between 20% and 1%).

Figure 7 plots out-of-sample Sharpe ratios for SDFs estimated separately within each size group, while Figure 8 plots pricing errors. The central conclusion from these figures is that the virtue of complexity arises in all stock size groups. In terms of magnitude, Figure 7 indeed suggests that the SDF Sharpe ratios in Figure 2 are heavily influenced by microcap

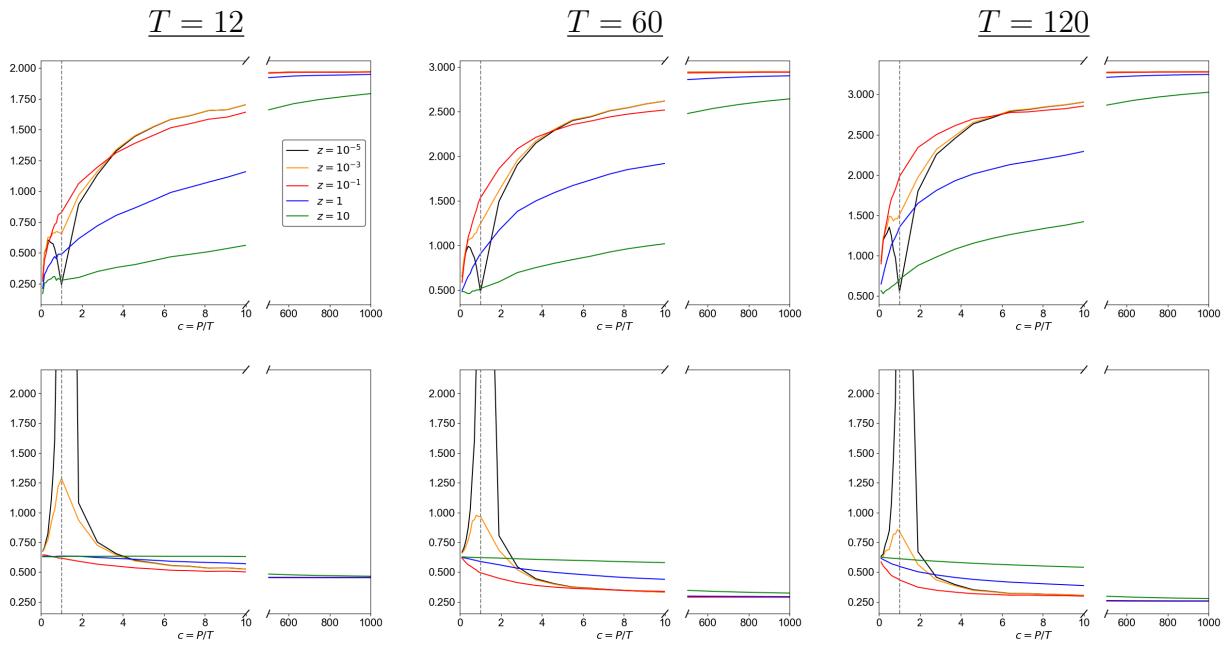

Figure 9: The Effect of Training Sample Size on Out-of-sample SDF Performance

Note. Realized out-of-sample SDF Sharpe ratio (top row) and pricing error (bottom row) training windows of $T = 12, 60$, and 120 months.

stocks. Yet the SDF Sharpe ratios based on mega or large stocks are on the order of 1.6 and 1.8, respectively. In other words, a complex factor model derived from mega stocks alone (roughly the 400 largest stocks in the US) produces an out-of-sample SDF Sharpe ratio well in excess of the FF6 model's Sharpe ratio of 0.74 (which uses the full cross-section of stocks). In summary, the patterns documented in our main analysis (Figure 2) do not appear to be driven by illiquidity and limits-to-arbitrage among the underlying assets. Instead, these findings suggest that models with a large number of factors are generally better suited to price assets in the cross-section, regardless of their liquidity.

3.6 Sensitivity to Sample Size

Our main analysis demonstrates the virtue of complexity when estimation is conducted in a rolling 360-month training sample. However, the benefits of complexity can accrue in much

smaller training samples. In Figure 9, we plot VoC curves for training samples of $T = 12, 60$ and 120 months. We find identical patterns in complex factor model behavior in training windows as short as a single year. However, for shorter training windows, model performance weakens as the training process has less information to learn from. The out-of-sample Sharpe ratio for our highest complexity factor model is 2.0, 2.9, 3.3, and 3.7 for $T = 12, 60, 120,$ and 360 months, respectively. Likewise, pricing errors reduce from roughly 0.45 for $T = 12$ months to 0.22 for $T = 360$.

4 AIPT: Theoretical Underpinnings of Complex Factor Models

This section presents artificial intelligence pricing theory. We first present assumptions that formalize the AIPT's core conjecture—that asset prices are determined by a high-dimensional factor structure. From these assumptions, we derive a statistical theory of large factor pricing models. Lastly, we show that the empirical behavior of factor pricing models in Section 3 bears a strikingly close correspondence to theoretically predicted behaviors of the AIPT.

4.1 Assets and Conditioning Information

Our theory begins with the following two assumptions:

Assumption 1 *There exist loadings $S_t \in \mathbb{R}^{N\times P}$, latent factors $\widetilde{F}_{t+1}$, and idiosyncratic shocks ε_{t+1} such that returns $R_{t+1} \in \mathbb{R}^N$ satisfy*

$$R_{t+1} = S_t \tilde{F}_{t+1} + \varepsilon_{t+1}, \tag{14}$$

where $E_t[\varepsilon_{t+1}] = 0$ and $E_t[\varepsilon_{t+1}\varepsilon_{t+1}'] = \Sigma_{\varepsilon,t}$, and $E[\varepsilon_{i,t}^4]$ are uniformly bounded. The latent factors satisfy $E_t[\tilde{F}_{t+1}] = \nu_F$, and $\Sigma_{F,t} = E_t[\tilde{F}_{t+1}\tilde{F}_{t+1}']$ satisfies $\mathrm{tr}(\Sigma_{F,t}) = O(1)$ as $P \to \infty$.

Assumption 2 *We have $S_t = \Sigma^{1/2} X_t \Psi^{1/2}$ for some positive definite matrices Σ, Ψ; here, the random variables $X_{i,k,t}$ satisfy $E[X_{i,k,t}] = 0$, $E[X_{i,k,t}^2] = 1$, and $X_{i,k,t}$ are independent*

and have uniformly bounded sixth moments. In the limit as $N,\ P \to \infty$, *both* Σ *and* Ψ *stay uniformly bounded,* $\text{tr}(\Sigma)$ *is uniformly bounded, and* $\lim_{N\to\infty} \text{tr}(\Sigma^2)/(\text{tr}(\Sigma))^2 = 0$. *Without loss of generality, everywhere in the sequel we assume* Σ *is normalized so that* $\text{tr}(\Sigma) = 1$.

Factor pricing models summarize the risk-return tradeoff among N risky assets in the economy. To this end, Assumption 1 describes the dependence structure of assets and expected returns. It is a standard conditional factor structure that allows for an arbitrary number of factors P.[19] Assumption 1 defines the relevant conditioning information in this economy, which amounts to the conditional factor loadings in the $N \times P$ matrix S_t. We refer to S_t as "characteristics" or "signals" in connection with the empirical asset pricing literature.[20] The P-vector of factor risk prices, denoted ν_F, together with the conditional loadings determine asset expected returns. Appendix B further discusses Assumption 1 as a generic statistical specification arising in the context of machine learning factor pricing models.

Next, our theoretical derivations require assumptions about the covariance structure of the signals S_t. By Assumption 2,

$$E[S_t'S_t] = \text{tr}(\Sigma)\Psi \in \mathbb{R}^{P\times P} \quad \text{and} \quad E[S_tS_t'] = \text{tr}(\Psi)\Sigma \in \mathbb{R}^{N\times N}, \tag{15}$$

thus the matrix Ψ captures the covariance of signals across factors and Σ captures their covariance across assets. The assumption of bounded $\text{tr}(\Sigma)$ is a no-arbitrage condition to ensure that the predictable variation in returns stays bounded. The last two limits in Assumption 2 ensure that characteristic-managed portfolios offer a meaningful amount of diversification across stocks.[21]

[19] The assumption $\text{tr}(\Sigma_{F,t}) = O(1)$ as $P \to \infty$ is the mathematical formalization of the idea of a factor structure. For example, if Σ_F has a finite rank K with bounded eigenvalues, then this condition is trivially satisfied. The assumption of constant conditional factor premia is without loss of generality since dynamic premia can be subsumed by S_t.

[20] One can think of the loadings as some function of other underlying conditioning characteristics, or the characteristics could be loadings themselves as in the BARRA model popular among industry professionals.

[21] For example, suppose that $\text{rank}(\Sigma) = 1$, so that $\Sigma^{1/2} = \pi\pi'$ for some $\pi \in \mathbb{R}^N$. Then, $S_t = \pi\pi' X_t \Psi^{1/2}$

4.2 Characteristic-managed Portfolios

From the equivalence of an SDF and the mean-variance efficient portfolio, Assumption 1 implies the following SDF representation.

Proposition 1 *A conditional stochastic discount factor is*

$$\tilde{M}_{t+1} \;=\; 1 \;-\; \tilde{w}(S_t)' R_{t+1}\,, \tag{17}$$

where

$$\tilde{w}(S_t) \;=\; (S_t \Sigma_{F,t} S_t' + \Sigma_{\varepsilon,t})^{-1} S_t \nu_F \tag{18}$$

is the conditional mean-variance efficient portfolio and

$$E_t[R_{i,t+1}\tilde{M}_{t+1}] \;=\; 0,\; i = 1, \cdots, N\,. \tag{19}$$

The SDF in (17) is stated as a conditional portfolio of the basic risky assets, R, as discussed in the empirical framework of Section 2. Estimating the conditional mean-variance portfolio of basic assets is challenging. Not only does it require estimates of means and covariances for a large number of assets, but it also requires these moments in *conditional* terms.

To avoid the difficult task of modeling the conditional distribution of basic assets, it is common in the empirical literature to instead study characteristic-managed portfolios, or

---

and therefore all factors are given by

$$F_{t+1} \;=\; \Psi^{1/2} X_t' \pi(\pi' R_{t+1})\,, \tag{16}$$

implying that all factor returns are proportional to returns on a single portfolio, $\pi' R_{t+1}$, with no idiosyncratic return. Assumption 2 ensures that such pathological situations cannot occur.

"factors,"[22]

$$F_{t+1} \;=\; S_t' R_{t+1}. \tag{20}$$

The conjecture is that by studying the *unconditional* properties of factors, we can learn about the conditional properties of asset markets. For example, Kozak et al. (2020) approximate the conditional SDF using the unconditional mean-variance efficient portfolio of managed portfolios.[23] Yet it is easy to see that the mean-variance portfolio of F_t,

$$\lambda = E[F_{t+1}F_{t+1}']^{-1}E[F_{t+1}] \tag{21}$$

is generally different from the conditionally efficient portfolio of basic assets that determines the true SDF in (17).

We prove in Proposition 2 that when the number of factors is large, the conjecture indeed holds, and the unconditional optimal portfolio of factors coincides with the true conditional SDF. While somewhat technical, the logic for this result is straightforward. Factors interact base asset returns with conditioning characteristics, and, in the large P limit, these interactions encode all available conditioning information. Therefore, we can rely on *unconditional* properties of factors to capture the conditional properties of the underlying assets.[24]

Proposition 2 (Characteristic-managed Portfolios and the Conditional SDF) *Suppose that in the limit, as $P \to \infty$, the vector of latent risk premia ν_F is uniformly bounded and*

[22]The literature often refers to the managed portfolios F_t as "factors," and we adhere to this slight abuse of nomenclature when the difference between F_t and the true factors $\tilde{F}_t$ is clear.

[23]Relatedly, to help justify the empirical approach of Kozak et al. (2020), Kozak and Nagel (2023) discuss conditions under which managed portfolios "span" the conditional SDF.

[24]See Appendix E for technical details.

satisfies

$$\nu_F' A \nu_F \;\to\; 0 \tag{22}$$

in probability, for any symmetric, positive definite A with uniformly bounded trace.[25] *Suppose also that $\Sigma_{\varepsilon,t} = I$*[26] *and let*

$$M_{t+1} = 1 - \lambda' F_{t+1} \;=\; 1 \;-\; w(S_t)' R_{t+1}, \;\; \textit{with } w(S_t) \;=\; \lambda' S_t \,, \tag{23}$$

be the factor approximation for the SDF with λ given by (21). Then, M_{t+1} converges to $\tilde{M}_{t+1}$ from (17) in probability and the Sharpe ratio of $w(S_t)'R_{t+1}$ converges to that of $\tilde{w}(S_t)'R_{t+1}$ as $N, P \to \infty$.

The surprising and very convenient implication of Proposition 2 is that model complexity, in fact, *simplifies* asset pricing. Much like continuous time limits simplify a variety of asset pricing calculations, large factor limits reduce conditional SDF modeling to an unconditional problem. Therefore, in the remaining theoretical development, we leave behind the base assets R_t and work directly with managed portfolios, F_t.[27]

4.3 Large Models as Approximations

As the George Box adage goes, "All models are wrong, but some are useful." In reality, the true data-generating process is unknown to the researcher. A core premise of modern artificial intelligence (and nonparametric statistics more broadly) is that large models are

[25] For example, this is the case when ν_F is sampled from $N(0, \Sigma_\lambda/P)$ for some bounded matrix Σ_λ. In this case, $E[\nu_F' A \nu_F] = \text{tr}(E[A\nu_F\nu_F']) = \text{tr}(AE[\nu_F\nu_F']) = \text{tr}(A\Sigma_\lambda)/P \le \text{tr}(A)\|\Sigma_\lambda\|/P \to 0$.

[26] When $\Sigma_{\varepsilon,t} \neq I$, Proposition 2 still holds true if we redefine managed portfolios as $F_{t+1} = S_t' \Sigma_{\varepsilon,t}^{-1/2} R_{t+1}$.

[27] Relatedly, Assumptions 1 and 2 govern the properties of R_t, which serves two purposes. First, the assumptions ensure that characteristic-managed portfolios span the SDF in the high-complexity limit, per Proposition 2. Second, they imply that characteristic-managed portfolios satisfy the technical conditions of random matrix theory (see Theorem 5 in the Appendix). Once these conditions are established, the origin of factors is no longer relevant, and the theory can treat them in the abstract when proving our main result in Theorem 3.

useful because they provide flexible approximations of the unknown truth. However, while added complexity brings better approximations, it also brings the usual statistical challenges of heavy parameterization in the face of limited data.

In this section, we design a framework to understand the out-of-sample behavior of factor pricing models as we gradually expand their parameterization. To embed this investigation in our theoretical environment, we consider a true SDF with a large number of factors, P^*. We then consider an empirical model that approximates the SDF with only a fraction $q = \frac{P}{P^*} \leq 1$ of those factors.

Assumption 3 *A misspecified empirical model of size $P < P^*$ ($q = \frac{P}{P^*}$) uses a subset of factors, $F_{t+1}(q) = (F_{i,t+1})_{i=1}^{P}$ with covariance matrix $E[F_t(q)F_t(q)'] \in \mathbb{R}^{P \times P}$.*

We are interested in characterizing out-of-sample factor model behavior as we expand the subset of factors from $P = 1, ..., P^*$, gradually reducing the degree of misspecification in the approximating model (when $P_1 = P$ the model is correctly specified) while increasing the number of parameters that must be estimated.

Our theoretical derivations are based on large P limits. The number of factors in the true, unattainable factor model is $P^* \to \infty$. We conceptualize the AIPT as a rich but nevertheless imperfect empirical model with $P \to \infty$ and $q \leq 1$. We now develop the limiting behavior of large empirical factor models using random matrix theory.

4.4 Feasible and Infeasible SDF Estimators

Proposition 2 directly motivates our empirical approximation to the SDF in equation (4) and the ridge estimator in equations (9) and (12). Following Assumption 3, the ridge estimator has access to a fraction q of the factors in the true model. Given a training sample of size T, the empirical model has complexity $c = P/T$. We define the *ridge SDF estimator* for an

empirical model of complexity c and specification q as

$$\hat{\lambda}(z;q;c) \;=\; (zI + \hat{E}[F_t(q)F_t(q)'])^{-1}\hat{E}[F_t(q)] \tag{24}$$

with corresponding portfolio return and SDF of

$$\hat{R}^M_{T+1}(z;q;c) \;=\; \hat{\lambda}(z;q;c)'F_{T+1}(q), \quad \hat{M}_{T+1}(z;q;c) = 1 - \hat{R}^M_{T+1}(z;q;c),$$

where $\hat{E}$ denotes sample average over the T training observations.

It will be useful to contrast $\hat{\lambda}(z;q;c)$ with the *infeasible ridge SDF estimator*

$$\lambda(z;q) = (zI + E[F(q)F(q)'])^{-1}E[F(q)] \tag{25}$$

and its return and SDF

$$R^M_{T+1}(z;q) \;=\; \lambda(z;q)'F_{T+1}(q), \quad M_{T+1}(z;q) = 1 - R^M_{T+1}(z;q). \tag{26}$$

This estimator is infeasible because it relies on the population mean and covariance of factors rather than their sample counterparts. The special case $\lambda = \lambda(0,1)$ corresponds to the true SDF in (23). Naturally, as z increases from zero or as q decreases from one, the Sharpe ratio of $R^M_{T+1}(z)$ declines. The portfolio $\lambda(z;q)$ is an intermediate object between the true SDF $\lambda(0,1)$ and the feasible estimator $\hat{\lambda}(z;q;c)$.

A remarkable aspect of the theoretical results below is that we can fully characterize the properties of $\hat{\lambda}(z;q;c)$ in terms of the infeasible estimator $\lambda(z;q)$. Our analysis focuses on the key properties that summarize the theoretical behavior of a factor-based SDF: its mean and variance (which in turn dictate its Sharpe ratio and pricing errors). Because they are central to our derivation below, we give special notation to these properties for the infeasible

ridge SDF:

$$\mathcal{E}(z;q) \equiv E[R^M_{T+1}(z;q)], \qquad \mathcal{V}(z;q) \equiv \text{Var}[R^M_{T+1}(z;q)]. \tag{27}$$

4.5 The Ridge SDF and Random Matrix Theory

It is perhaps easiest to understand our large P theory for factor model behavior through the calibrations in Section 4.7, and readers interested in a non-technical discussion of the theory may proceed there directly. For interested readers, this section provides a brief overview of the RMT concepts behind our analysis, and Section 4.6 gives a detailed presentation of our main result.

The central challenge to understanding $\hat{\lambda}(z;q;c)$ is the $P \times P$ matrix $\hat{E}[F_t(q)F_t(q)']$ whose dimension grows with the number of factors. The limiting properties of this object require the apparatus of random matrix theory (RMT).

One technical insight in our analysis is that incorporating ridge regularization in the SDF estimation problem allows us to extend core RMT results, such as the Marčenko and Pastur (1967) theorem, to asset pricing analysis. We derive the necessary extensions of Marčenko and Pastur (1967) to describe expected out-of-sample SDF behavior in terms of three objects: the eigenvalue distribution of the factor population covariance matrix, $E[FF']$, the number of parameters per training observation (i.e., complexity, c), and the extent of model misspecification (governed by q). Recall that our factors are defined as $F_{t+1} = S_t' R_{t+1}$, with signals and returns satisfy Assumption 1 and 2. As we show in the Appendix, the only properties of the data-generating process that matter for portfolio performance in the high complexity limit are Ψ and ν_F. Let $\Psi(q) = \Psi[1:P, 1:P]$ be the Ψ sub-matrix corresponding to the first P signals, and, similarly, let $\nu(q) = (\Psi\nu_F)[:P]$ to be the first P coordinates of the $\Psi\nu_F$ vector.

We need the following regularity condition on the covariance matrix of factors.

Assumption 4 *In the limit as* $P \to \infty$, *the eigenvalue distribution of* $\Psi(q) = U(q)D(q)U(q)'$, $D(q) = \text{diag}(D_i(q))$, *converges to a limit distribution* $dH(x;q)$ *supported on a bounded interval. Furthermore,* $\nu(q) = (\Psi\nu_F)[:P]$ *is such that*

$$\frac{1}{\|\nu(q)\|^2}\sum_i |U(q)'\nu(q)|^2(i)\mathbf{1}_{D_i(q)<x} \tag{28}$$

weakly converges to a probability distribution on $\mathbb{R}$ *as* $P \to \infty$. *Furthermore, in addition to Assumption 1, we have* $E[\varepsilon_{i,t}^3] = 0$ *for all* i, *and the latent factors satisfy* $E[\|\tilde{F}_{t+1}\|^4] < K$ *for some* $K > 0$.[28]

The central object that dictates the behavior of large factor models is the eigenvalue distribution of the factor covariance matrix, which our derivations represent as a Stieltjes transform. This transform for the eigenvalue distribution of $E[F_tF_t']$ in the large P limit is denoted

$$m(-z;q) \;=\; \lim_{P\to\infty}\frac{1}{P}\,\text{tr}\left((E[F_t(q)F_t(q)'] + zI)^{-1}\right), \tag{29}$$

and, similarly, the infeasible moments (27) converge to well-defined limits

$$\mathcal{E}(z;q)\ ,\ \mathcal{V}(z;q)\,.$$

We denote the limiting eigenvalue distribution of $\hat{E}[F_t(q)F_t(q)']$ as

$$m(-z;q;c) \;=\; \lim_{\substack{P\to\infty\\ P/T\to c\\ P/P^*\to q}}\frac{1}{P}\,\text{tr}\left(\left(zI + \hat{E}[F_t(q)F_t(q)']\right)^{-1}\right) \tag{30}$$

The main challenge of the AIPT is that when the factor model is heavily parameterized ($c > 0$), $\hat{E}[F_t(q)F_t(q)']$ converges not to $E[F(q)F(q)']$ but to a distortion of it. In Theorem

[28] See Lemma 13 which shows how this condition is essentially equivalent to Σ_F having a bounded trace and ν_F being bounded, as in Assumption 1, plus a condition on the fourth moments.

5 in Appendix D, we apply an extension of the Marčenko and Pastur (1967) theorem due to Bai and Zhou (2008) to our asset pricing environment in order to establish the explicit mapping between $\hat{E}[F_t(q)F_t(q)']$ and $E[F_tF_t']$. Namely,

$$m(z;q;c) = \frac{1}{1 \;-\; c \;-\; c\,z\,m(z;q;c)}\, m\left(\frac{z}{1 \;-\; c \;-\; c\,z\,m(z;q;c)};q\right). \tag{31}$$

The nonlinear master equation in (31) has a unique positive solution $m(z;c;q)$ for any $z<0$. This solution is a function of only $m(z;q)$ and c. Thus, equation (31) links $m(-z;q;c)$ in (30) to $m(-z;q)$ in (29). When complexity is low and misspecification is small, i.e. when $c \approx 0$ and $q \approx 1$, (31) implies $m(z;q;c) \approx m(z;q)$, as predicted by the standard law of large numbers. However, for $c>0$, the empirical Stieltjes transform and the "true" Stieltjes transform decouple in a *deterministic* manner. When complexity is high (e.g., $c>1$), large parts of information about $m(z)$ are lost in finite samples even in the correctly specified ($q=1$) case, and this information loss is, of course, further exacerbated by misspecification.

The last object we need to introduce is the particular formulation of the HJD that we use to characterize pricing errors. In the high complexity regime, exact details of computing the out-of-sample HJD are important. We assume that the data sample is split into two sets: in-sample data indexed as $t \in [1,T]$ and out-of-sample data indexed as $t \in (T+1, T+T_{OS}]$, where T_{OS} is the number of out-of-sample periods. For simplicity, we focus our theoretical analysis on the case in which the test assets are the factors $F_t(q)$ used to estimate the SDF.[29] We thus define the out-of-sample HJD as

$$\mathcal{D}^{HJ}_{OS}(z;q;c) \;=\; \mathscr{E}_{OS}(z;q;c)'\, B^{+}_{OS}\, \mathscr{E}_{OS}(z;q;c)\,,$$

Where $\mathscr{E}_{OS}(z;q;c) \;=\; \frac{1}{T_{OS}} \sum_{t\in(T,T+T_{OS}]} F_t(q)\hat{M}_t(z;q;c)$ is the out-of-sample pricing error

[29] The pricing error properties of the SDF are particularly tractable to derive when the test assets are the P factors $F_t(q)$, but this point can be generalized this at the cost of further notation and derivations.

vector and $B_{OS} = \frac{1}{T_{OS}} \sum_{t \in (T, T+T_{OS}]} F_t(q) F_t(q)'$ is the out-of-sample test asset second moment.

4.6 The Theoretical Behavior of Large Factor Models

We now state our main theoretical result, which describes the expected out-of-sample properties of large factor models.

Theorem 3 *In the limit as* $P,\ T \to \infty,\ P/T \to c, P/P^* \to q$, *the expected out-of-sample moments of the ridge SDF portfolio satisfy*

$$
\begin{aligned}
&i. \quad \lim E[\hat{R}^M_{T+1}(z;q;c)] \ = \ \mathcal{E}\big(Z^*(z;q;c);q\big) \quad \textit{where} \\
&\quad Z^*(z;q;c) \ = \ z\big(1+\xi(z;q;c)\big) \in (z, z+c) \ \ \textit{and} \ \ \xi(z;q;c) \ = \ \frac{c(1 \ - \ m(-z;c;q)z)}{1-c(1 \ - \ m(-z;c;q)z)}, \\
&ii. \quad \lim \mathrm{Var}[\hat{R}^M_{T+1}(z;q;c)] \ = \ \mathcal{V}\big(Z^*(z;q;c);q\big) \ + \ G(z;q;c)\,\mathcal{R}\big(Z^*(z;q;c);q\big) \quad \textit{where} \\
&\quad G(z;q;c) \ = \ (z\xi(z;q;c))' \ \in (0, cz^{-2}], \quad \mathcal{R}(z;q) \ \equiv \ (1-\mathcal{E}(z;q))^2 + \mathcal{V}(z;q) \\
&iii. \quad \lim \frac{\mathrm{Var}[\hat{R}^M_{T+1}(z;q;c)]}{E[\hat{R}^M_{T+1}(z;q;c)]^2} \ = \ (1+G(z;q;c))\frac{1}{\mathcal{S}^2(Z^*;q)} + \ G(z;q;c)\ \left(\frac{1-\mathcal{E}(Z^*;q)}{\mathcal{E}(Z^*;q)}\right)^2, \\
&iv. \quad \lim E[\mathcal{D}^{HJ}_{OS}(z;q;P;T)] \ = \ -(1-\mathcal{E}(0;1))\max(1-c_{OS},0) \\
&\quad + \ (1+G(z;q))\mathcal{R}(Z^*(z;q;c);q)\,.
\end{aligned}
$$

The central theme in Theorem 3 is that the large number of factors relative to the number of training observations limits the estimator's ability to learn the true parameters. When $c > 0$, there are too many parameters and too few data points for the estimator to converge to its population counterpart. This failure to fully hone in on the truth results in an asymptotic wedge between the out-of-sample performance of the trained model and that of the true model. We refer to this as "limits to learning." Perhaps surprisingly, Theorem 3 shows that we can explicitly quantify the effects of limits to learning using only knowledge of the factors' sample covariance in the training data.

Limits to learning manifest in two ways—implicit shrinkage and complexity risk—which impact the SDF's out-of-sample properties. We describe these properties now.

4.6.1 Expected Return

Part *i.* of Theorem 3 describes how complexity inhibits the expected return of the SDF portfolio. It describes the expected return in terms of the infeasible ridge portfolio's return. The key to this is the function $Z^*(z;q;c)$, which is the "implicit shrinkage" of the feasible estimator. Z^* is monotone increasing in c. In other words, high complexity imposes additional shrinkage on the SDF, above and beyond the explicit ridge shrinkage z.[30]

If the model is correctly specified ($q = 1$) and with a complexity of zero, then $Z^*(z;0;1) = z$ and the feasible SDF's expected return converges to the infeasible expected return, $E\left[\hat{\lambda}(z)'R_{T+1}\right] \to E[\lambda(z)'R_{T+1}] = \mathcal{E}(z)$. But holding z fixed, a rise in complexity to $c > 0$ induces additional bias in the estimator and drives down the expected return of the SDF portfolio. By how much? By the same amount that the expected return drops when the infeasible portfolio's shrinkage rises from z to $Z^*(z;1;c)$. In other words, the challenge of learning in a complex setting is equivalent to knowing the true factor moments but being forced to use an unduly large shrinkage. Remarkably, $Z^*(z;1;c)$ is available in closed form thanks to the expression for $\xi(z;q;c)$ from RMT.

The monotonicity of $Z^*(z;q;c)$ in z means that out-of-sample expected returns are highest with minimal shrinkage. But even in the ridgeless limit when $z \to 0$, Theorem 3 shows there are limits to learning. In particular, there is an unavoidable reduction in expected return because $Z^*(z;q;c)$ is uniformly bounded away from zero in the high complexity regime ($c > 1$).

Thus, limits to learning hold even if the model is correctly specified. But this case

[30] Complexity generates this additional implicit shrinkage in an intuitive way. Holding z fixed, if we increase P, we cannot raise $\|\lambda\|^2$ further due to the ridge penalty. By adding more parameters, we can only continue to satisfy the ridge constraint by shrinking the $\hat{\lambda}$ vector further.

is unrealistic; we can't ever expect to achieve an empirical model that nests all relevant conditioning information or uses that information in its proper nonlinear form. Furthermore, under correct specification, comparative statics for complexity change both the empirical *and the true model* as we vary c. Thus, these theoretical comparative statics cannot be taken to data.

The empirically relevant comparative statics must consider a single true DGP, in essence fixing the set of true factors and varying the number of those factors that the empirical model has access to. We can conceptualize these comparative statics by varying q. When the model is misspecified ($q < 1$), expected returns are hindered further because the model relies on incomplete information. But the effect of q is subtle. Holding T fixed, a higher q has two opposing effects on the expected SDF return. First, larger q means that there are more empirical factors, so c rises as well, which exacerbates the limits to learning. However, higher q also reduces specification bias and, therefore, improves the approximation power of the model. This shows up as a smaller implicit shrinkage, Z^*. Which effect dominates depends on the eigenvalue distribution of the factors. As we will see in the calibration below, when there is a concentrated eigenvalue distribution and thus a few dominant factors—as in the assumptions of the APT—the cost of complexity (i.e., more severe limits to learning) dominates the approximation benefits and high complexity may hurt expected SDF returns on net. But when the eigenvalue distribution is dispersed, and there are many relevant factors—i.e., under the conditions of the AIPT—approximation gains dominate, and high complexity leads to better out-of-sample SDF returns.

4.6.2 Variance

Part *ii.* of Theorem 3 regards the variance of the ridge SDF portfolio. On the right side of (32), the first term relates to the implicit shrinkage effect discussed above. In the large P limit, the feasible SDF portfolio with parameter z has the same volatility as the infeasible

portfolio with a larger ridge parameter $Z^*(z;q;c) > z$. Due to this implicit shrinkage and the monotonicity of Z^* in c, SDF portfolio variance decreases with model complexity when $c > 1$.

While the first term in (32) reflects the implicit shrinkage in large factor models, the second term represents a different phenomenon that we call "complexity risk." Complexity risk can be thought of as sampling variation that exists even in the large T limit. It is a pure-variance effect governed by the function $G(z;q;c)$, independent of expected factor returns, and only depends on the eigenvalue distribution of $E[F_t(q)F_t(q)']$. For a simple model, $c = 0$, there are infinitely more observations than parameters, so the SDF estimator converges to a non-random limit, thus $G(z;q;0) = 0$, and there is no complexity risk. However, when $c > 0$, sampling variation survives even in the large T limit because the number of parameters is too large to be accurately informed by the data. Complexity risk is a second-moment manifestation of limits to learning. The behavior of SDF portfolio variance is driven primarily by c and is relatively insensitive to the extent of model misspecification, q.

4.6.3 Sharpe Ratio and Pricing Error

Part *iii.* of Theorem 3 describes the SDF portfolio's limiting Sharpe ratio. First, focusing on the effect of complexity, consider the correct specification case, $q = 1$. Building on *i.* and *ii.*, a rise in complexity unambiguously decreases the expected SDF return due to implicit shrinkage. The effect on the variance is mixed—implicit shrinkage lowers variance, but complexity risk raises it. The net effect of complexity is an unambiguous decrease in the Sharpe ratio in the correctly specified setting.

As emphasized above, while the $q = 1$ case is helpful for developing intuition around complexity effects, our real interest lies in understanding SDF Sharpe ratios in the misspecified setting. In this case, the effect of complexity on the Sharpe ratio is ambiguous and depends on the eigenvalue distribution of the factor covariance. When eigenvalues are concentrated,

complexity has a small net effect on SDF Sharpe, and can potentially reduce it. However, when the eigenvalue distribution is relatively flat, the Sharpe ratio tends to increase with complexity, and the gains from complexity are potentially large.

In the high complexity limit, pricing error is the mirror image of Sharpe ratio. When conditions are such that Sharpe ratio increases with complexity, it is also the case that pricing errors are reduced by complexity.

In summary, Theorem 3 characterizes the subtle mechanisms through which complexity determines a model's out-of-sample performance. Complexity introduces a tradeoff between the quality of a model's approximation for the unknown truth on the one hand versus limits to learning (implicit shrinkage and complexity risk) on the other. In terms of the effect on the SDF expected return, a larger model reduces specification bias, which raises expected returns, but it also induces additional implicit shrinkage, which lowers returns. While complexity's implicit shrinkage has a dampening effect on SDF volatility, it at the same time raises volatility by producing sampling variation (i.e., complexity risk) that survives in the large T limit. When the marginal approximation benefits are large relative to the loss due to limits to learning, we find that the complexity raises the out-of-sample SDF Sharpe ratio and reduces pricing errors. Under these conditions, there is a virtue of complexity, and large factor models dominate simple, parsimonious models. Theorem 3 shows that the eigenvalue distribution among factors is critical for determining whether complexity is a virtue or a vice.

4.7 Illustrating the Theory

Theorem 3 derives foundational results for understanding the role of complexity in factor pricing models. However, the analytical expressions are dense and can be difficult to parse. In this section, we attempt to bring Theorem 3 to life by visualizing the role of complexity in different data environments. We present two DGP calibrations of our theory. In each

case, we evaluate the closed-form theoretical expressions in Theorem 3 to plot the key out-of-sample SDF properties as a function of model complexity given the assumed DGP. These illustrations are the direct theoretical counterparts to the empirical VoC curves presented in Section 3.

4.7.1 Calibration 1: The AIPT

As emphasized in the theory discussion above, the eigenvalue distribution of $E[F(q)F(q)']$ is a critical driver of out-of-sample SDF behavior. Our first calibration assumes a uniform eigenvalue distribution, $E[F(q)F(q)'] = I$. This calibration captures the spirit of the AIPT—that is, the conjecture that the underlying DGP is subject to a large number of distinct factors. To complete the calibration, we also assume $E[F(q)] \sim N(0, I/P)$, and we set the true complexity P^*/T to 10. To produce theoretical VoC curves, we gradually increase the fraction of observable factors in the empirical model ($q = P/P^*$) from 0 to 1, thus varying model complexity $c = P/T$ from 0 to 10.

Figure 10 plots the results. Each curve corresponds to a different choice of ridge penalty z, and each point on a curve corresponds to a different number of factors P. The four panels show how complexity (on the x-axis) affects the expected out-of-sample mean, variance, and Sharpe of the SDF portfolio and the SDF's expected out-of-sample pricing errors.

Panel A shows that the out-of-sample expected return of the SDF portfolio is increasing in model complexity. The intuition for this result is that higher model complexity allows the SDF model to approximate the unknown true SDF more accurately. As the approximation improves and specification error shrinks, the estimated SDF is able to achieve a higher expected portfolio return. When the true DGP has many factors, the gains from better approximation overwhelm performance loss due to limits to learning. This is true for all ridge penalty levels and throughout the full range of complexity. Expected return curves are

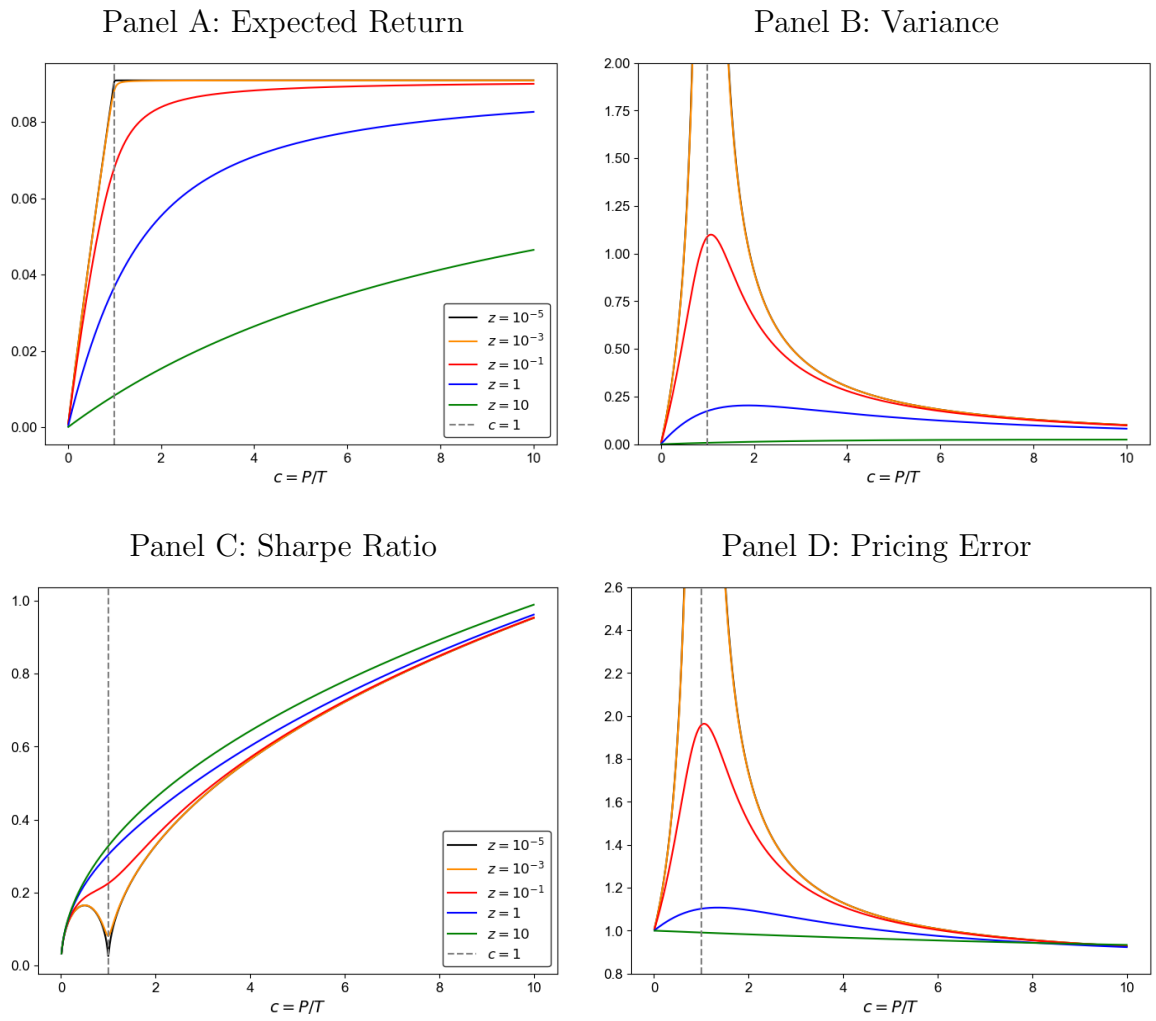

Figure 10: Out-of-sample Behavior of Large Factor Models in AIPT Calibration

Note. Limiting out-of-sample mean, variance, Sharpe ratio, and pricing error (HJD) of the SDF as a function of c and z from Theorem 3 assuming an identity factor covariance matrix.

flatter with higher explicit ridge shrinkage, z. This is because more shrinkage increases the estimator's bias, reducing the approximating power of the SDF, which eats into its returns.

In Panel B, we see that SDF volatility is highly sensitive to model complexity. When c approaches unity, the variance of the low z SDFs spikes. The logic for this behavior follows from arguments in KMZ. As $c \to 1$, the unregularized sample covariance matrix of factors becomes unstable, which causes the unregularized estimator to explode. Intuitively, when

$c = 1$, the number of model parameters equals the number of time series observations, so there is an SDF estimate with an infinite in-sample Sharpe ratio. Without regularization, this estimator badly overfits the training data and produces disastrous out-of-sample behavior.

When $c \gg 1$, the ridge SDF estimator experiences implicit shrinkage due to complexity, which reduces SDF volatility. As the other curves in Panel B show, SDF volatility can also be controlled by raising the explicit ridge shrinkage, z. This is the low variance benefit of the shrinkage-induced bias.

The out-of-sample SDF Sharpe ratio is shown in Panel C. The ridgeless SDF estimator demonstrates "double ascent," in analogy to the "double descent" MSE phenomenon studied in statistics literature.[31] At low complexity ($c \ll 1$), the Sharpe ratio rises with complexity as larger models show improved approximation power benefits. But near $c = 1$, the Sharpe ratio collapses to zero due to the explosion in SDF variance. Finally, at high complexity ($c \gg 1$), variance comes under control, and the benefits of improved approximation again dominate and lead to an increasing Sharpe ratio. Panel C also demonstrates that, with appropriate explicit shrinkage z, the complex SDF estimator exhibits "permanent ascent" with an increasing Sharpe ratio throughout the full range of complexity.

Finally, Panel D illustrates the behavior of out-of-sample SDF pricing errors (HJD) as a function of complexity. Pricing errors in Panel D are a mirror image of the patterns for the Sharpe ratio in Panel C. The more complex the SDF, the better its ability to price assets out-of-sample. As long as the number of true factors (P) is large, these pricing errors never go to zero, even for very high-complexity empirical models. Higher complexity means the empirical SDF includes more true factors, improving its pricing ability. But at the same time, higher complexity also means more stringent limits to learning, so out-of-sample pricing errors are bounded away from zero (this is true even for high complexity models that are *correctly* specified).

The most important point regarding Figure 10 is that it closely matches the qualitative

[31] See, for example, Spigler et al. (2019); Belkin et al. (2018, 2019, 2020); Bartlett et al. (2020).

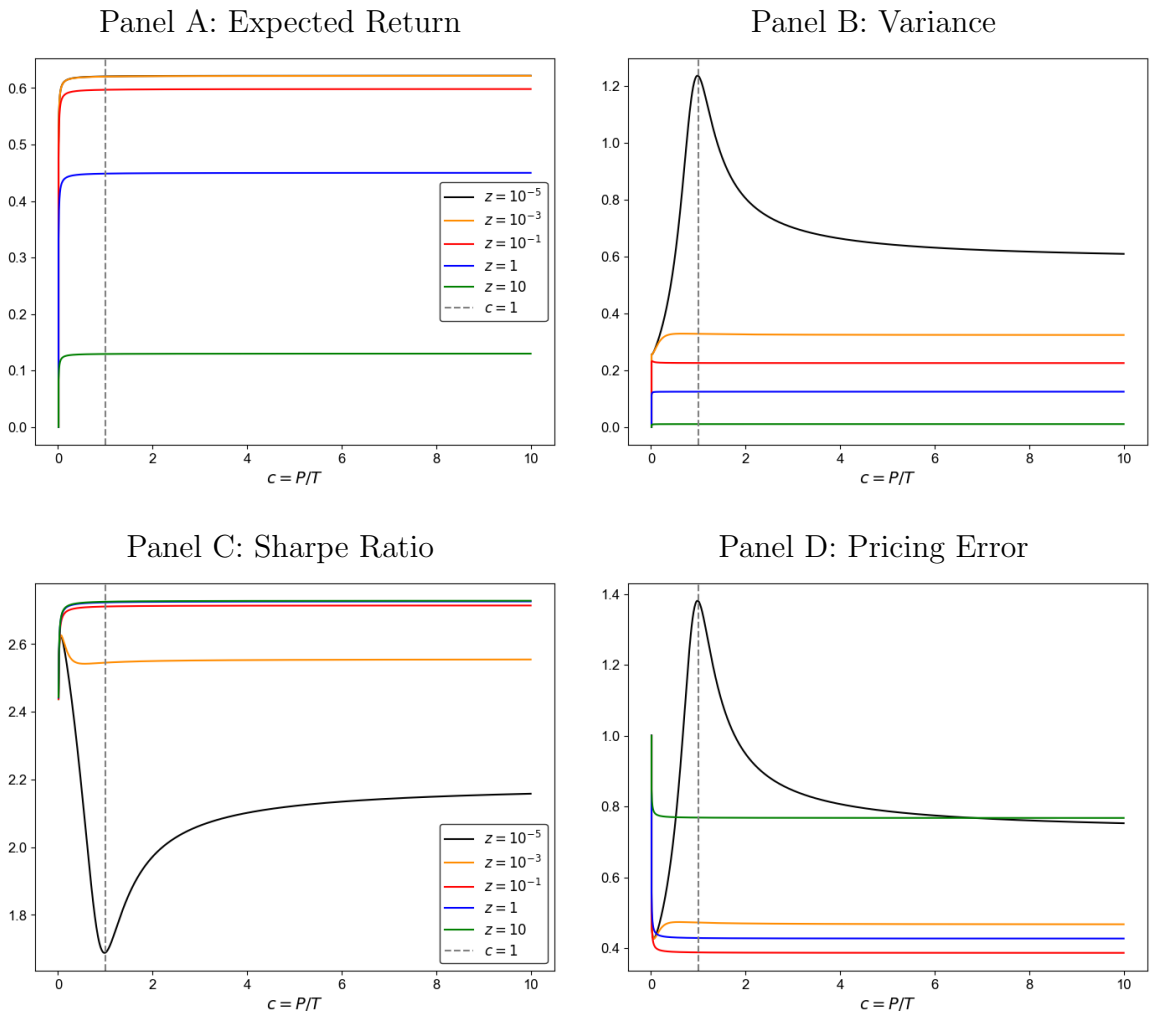

Figure 11: Out-of-sample Behavior of Large Factor Models in APT Calibration

Note. Limiting out-of-sample mean, variance, Sharpe ratio, and pricing error (HJD) of the SDF as a function of c and z from Theorem 3 assuming a factor covariance matrix with eigenvalues $(1, \cdots, P^*)^{-2}$.

patterns in the data, as documented in Figure 2. The out-of-sample expected return and Sharpe ratio of the ridge SDF are increasing in complexity while pricing errors are decreasing.

4.7.2 Calibration 2: The APT

To contrast with the AIPT in the previous calibration, we change the DGP assumption so that the cross-section of returns is dominated by only a few distinct factors, thus embodying

the APT conjecture. To do so, we change only one aspect of the calibration in Section 4.7.1 by assuming that eigenvalues of the true covariance matrix Ψ are given by $1, 2^{-2}, \cdots, P^{*-2}$. This assumption implies that $E[F(q)F(q)']$ has an effective rank $\text{EffRank}(\Psi) = (\text{tr}\,\Psi)^2/\,\text{tr}(\Psi^2)$ of 2.5. As (Bartlett et al., 2020) show, $\text{EffRank}(\Psi)$ is an intuitive way to quantify the number of significant eigenvalues of Ψ in the context of high dimensional models.

Figure 11 plots the results. The main conclusion conveyed by this figure is that there are no benefits to complexity when the eigenvalue distribution is very concentrated. When the ridge penalty is small (e.g., $z \leq 10^{-3}$), high-complexity models have lower expected returns and Sharpe ratios (and larger pricing errors) than models with complexity near zero. The fact that peak SDF performance occurs when $c \approx 0$ is at odds with the patterns in Figure 2, where performance peaks when $c \gg 1$ regardless of the value of z.[32]

In summary, while the AIPT conjecture aligns closely with the evidence in Figure 2, the APT appears counterfactual.

5 Conclusion

In this paper, we show that the out-of-sample performance of factor pricing models is increasing in terms of the number of factors. Larger models achieve higher risk-adjusted returns and lower pricing errors than smaller models, including standard low-dimensional factor models in the literature. Our novel machine learning empirical design allows us to compare models with varying degrees of statistical complexity while holding the raw data inputs fixed. More heavily parameterized models outperform simpler models because they afford a better approximation to the unknown data-generating process. The cost of these parameters is higher out-of-sample model volatility, but this cost is more than justified by the gains in expected returns from using rich parameterizations.

[32]Interestingly, while there are no gains to using large factor models in the APT calibration, there is surprisingly also *no cost* as long as moderate ridge shrinkage is used. This is because the overfit costs of adding additional (mostly redundant) parameters are offset by the implicit shrinkage that comes with complexity.

We then develop a theory of machine learning SDF estimators founded on the concept of statistical model complexity. Among our key theoretical findings is the virtue of asset pricing model complexity: In essence, the out-of-sample performance of factor pricing models generally improves with the number of factors. This result holds as long as the covariance matrix of factors is not too concentrated. We also characterize the limits to learning that arise from the application of highly parameterized prediction models amid relative data scarcity. While heavy parameterization precludes consistent estimation of the SDF, the virtue of complexity arises from the improved approximation power of complex models, overwhelming the countervailing effect of limits to learning. The empirical patterns that we document bear a strikingly close resemblance to the predictions of our theory. Meanwhile, the theoretical predictions of the APT are contradicted by our empirical evidence.

References

Asness, Clifford S, Tobias J Moskowitz, and Lasse Heje Pedersen, "Value and momentum everywhere," *The Journal of Finance*, 2013, *68* (3), 929–985.

Avramov, Doron, Si Cheng, and Lior Metzker, "Machine learning vs. economic restrictions: Evidence from stock return predictability," *Management Science*, 2023, *69* (5), 2587–2619.

Bai, Zhidong and Wang Zhou, "Large sample covariance matrices without independence structures in columns," *Statistica Sinica*, 2008, pp. 425–442.

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

Bartlett, Peter L, Philip M Long, Gábor Lugosi, and Alexander Tsigler, "Benign overfitting in linear regression," *Proceedings of the National Academy of Sciences*, 2020, *117* (48), 30063–30070.

Belkin, M, D Hsu, S Ma, and S Mandal, "Reconciling modern machine learning and the bias-variance trade-off. arXiv e-prints," 2018.

Belkin, Mikhail, Alexander Rakhlin, and Alexandre B Tsybakov, "Does data interpolation contradict statistical optimality?," in "The 22nd International Conference on Artificial Intelligence and Statistics" PMLR 2019, pp. 1611–1619.

___, Daniel Hsu, and Ji Xu, "Two models of double descent for weak features," *SIAM Journal on Mathematics of Data Science*, 2020, *2* (4), 1167–1180.

Box, George EP and Gwilym Jenkins, *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day, 1970.

Bryzgalova, Svetlana, Markus Pelger, and Jason Zhu, "Forest through the trees: Building cross-sections of stock returns," *Available at SSRN 3493458*, 2020.

Chen, Andrew Y and Tom Zimmermann, "Open source cross-sectional asset pricing," *Critical Finance Review, Forthcoming*, 2021.

Chen, Luyang, Markus Pelger, and Jason Zhu, "Deep learning in asset pricing," *Management Science*, 2023.

Chen, Xiaohong and Sydney C Ludvigson, "Land of addicts? an empirical investigation of habit-based asset pricing models," *Journal of Applied Econometrics*, 2009, *24* (7), 1057–1093.

Chinco, Alex, Adam D Clark-Joseph, and Mao Ye, "Sparse signals in the cross-section of returns," *The Journal of Finance*, 2019, *74* (1), 449–492.

Cong, Lin William, Guanhao Feng, Jingyu He, and Xin He, "Growing the efficient frontier on panel trees," *NBER Working Paper*, 2022, (w30805).

Connor, Gregory, Matthias Hagmann, and Oliver Linton, "Efficient semiparametric estimation of the Fama–French model and extensions," *Econometrica*, 2012, *80* (2), 713–754.

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

Daniel, Kent, David Hirshleifer, and Lin Sun, "Short-and long-horizon behavioral factors," *The review of financial studies*, 2020, *33* (4), 1673–1736.

Fama, Eugene F and Kenneth R French, "Common risk factors in the returns on stocks and bonds," *Journal of financial economics*, 1993, *33* (1), 3–56.

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

Fan, Jianqing, Yuan Liao, and Weichen Wang, "Projected principal component analysis in factor models," *Annals of statistics*, 2016, *44* (1), 219.

___ **, Zheng Tracy Ke, Yuan Liao, and Andreas Neuhierl**, "Structural Deep Learning in Conditional Asset Pricing," *Available at SSRN 4117882*, 2022.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu, "Taming the factor zoo: A test of new factors," *The Journal of Finance*, 2020, *75* (3), 1327–1370.

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber, "Dissecting characteristics nonparametrically," *The Review of Financial Studies*, 2020, *33* (5), 2326–2377.

Gagliardini, Patrick, Elisa Ossola, and Olivier Scaillet, "Time-varying risk premium in large cross-sectional equity data sets," *Econometrica*, 2016, *84* (3), 985–1046.

Gibbons, Michael R, Stephen A Ross, and Jay Shanken, "A test of the efficiency of a given portfolio," *Econometrica: Journal of the Econometric Society*, 1989, pp. 1121–1152.

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

___, Bryan Kelly, and Dacheng Xiu, "Factor models, machine learning, and asset pricing," *Annual Review of Financial Economics*, 2022, *14*, 337–368.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, "Autoencoder Asset Pricing Models," *Journal of Econometrics*, 2020.

___, ___, and ___, "Empirical asset pricing via machine learning," *The Review of Financial Studies*, 2020, *33* (5), 2223–2273.

Guijarro-Ordonez, Jorge, Markus Pelger, and Greg Zanotti, "Deep learning statistical arbitrage," *arXiv preprint arXiv:2106.04028*, 2021.

Han, Yufeng, Ai He, David Rapach, and Guofu Zhou, "Expected stock returns and firm characteristics: E-LASSO, assessment, and implications," *SSRN*, 2019.

Hansen, Lars Peter and Kenneth J Singleton, "Generalized instrumental variables estimation of nonlinear rational expectations models," *Econometrica: Journal of the Econometric Society*, 1982, pp. 1269–1286.

___ and Ravi Jagannathan, "Assessing specification errors in stochastic discount factor models," *The Journal of Finance*, 1997, *52* (2), 557–590.

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

Harvey, Campbell R, Yan Liu, and Heqing Zhu, "... and the cross-section of expected returns," *The Review of Financial Studies*, 2016, *29* (1), 5–68.

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

He, Ai, Dashan Huang, Jiaen Li, and Guofu Zhou, "Shrinking factor dimension: A reduced-rank approach," *Management science*, 2023, *69* (9), 5501–5522.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White, "Multilayer feedforward networks are universal approximators," *Neural networks*, 1989, *2* (5), 359–366.

Hou, Kewei, Chen Xue, and Lu Zhang, "Digesting anomalies: An investment approach," *The Review of Financial Studies*, 2015, *28* (3), 650–705.

___ **,** ___ **, and** ___ , "Replicating anomalies," *The Review of Financial Studies*, 2020, *33* (5), 2019–2133.

___ **, Haitao Mo, Chen Xue, and Lu Zhang**, "An augmented q-factor model with expected growth," *Review of Finance*, 2021, *25* (1), 1–41.

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

Kan, Raymond and Cesare Robotti, "Model comparison using the Hansen-Jagannathan distance," *The Review of Financial Studies*, 2009, *22* (9), 3449–3490.

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

___ **, Boris Kuznetsov, Semyon Malamud, and Teng Andrea Xu**, "Large (and Deep) Factor Models," *arXiv preprint arXiv:2402.06635*, 2024.

___ , **Semyon Malamud, and Kangying Zhou**, "The virtue of complexity in return prediction," *The Journal of Finance*, 2024, *79* (1), 459–503.

___ , **Seth Pruitt, and Yinan Su**, "Characteristics are Covariances: A Unified Model of Risk and Return," *Journal of Financial Economics*, 2020.

Kelly, Bryan T, Semyon Malamud, and Kangying Zhou, "The virtue of complexity everywhere," *Available at SSRN*, 2022.

Knowles, Antti and Jun Yin, "Anisotropic local laws for random matrices," *Probability Theory and Related Fields*, 2017, *169* (1), 257–352.

Kozak, Serhiy, "Kernel trick for the cross-section," *Available at SSRN 3307895*, 2020.

___ , **Stefan Nagel, and Shrihari Santosh**, "Interpreting Factor Models," *The Journal of Finance*, 2018, *73* (3), 1183–1223.

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

Kozak, Serhyi and Nagel, "When do cross-sectional asset pricing factors span the stochastic discount factor?," *Working Paper*, 2023.

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

Marčenko, Vladimir A and Leonid Andreevich Pastur, "Distribution of eigenvalues for some sets of random matrices," *Mathematics of the USSR-Sbornik*, 1967, *1* (4), 457.

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

McLean, R David and Jeffrey Pontiff, "Does academic research destroy stock return predictability?," *The Journal of Finance*, 2016, *71* (1), 5–32.

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

Rahimi, Ali and Benjamin Recht, "Random Features for Large-Scale Kernel Machines.," in "NIPS," Vol. 3 Citeseer 2007, p. 5.

Rapach, David E and Guofu Zhou, "Time-series and cross-sectional stock return forecasting: New machine learning methods," *Machine learning for asset management: New developments and financial applications*, 2020, pp. 1–33.

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

Santos, Tano and Pietro Veronesi, "Conditional betas," 2004.

Spigler, Stefano, Mario Geiger, Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart, "A jamming transition from under-to over-parametrization affects generalization in deep learning," *Journal of Physics A: Mathematical and Theoretical*, 2019, *52* (47), 474001.

Stambaugh, Robert F and Yu Yuan, "Mispricing factors," *The review of financial studies*, 2017, *30* (4), 1270–1315.

Tukey, John W, "Discussion, emphasizing the connection between analysis of variance and spectrum analysis," *Technometrics*, 1961, *3* (2), 191–219.

White, Halbert, *Estimation, inference and specification analysis* number 22, Cambridge university press, 1996.

Internet Appendix

A Road Map

The Appendix is organized as follows:

- Section B provides a Neural Network interpretation of the AIPT environment.
- Section C reports some useful properties of the infeasible portfolio.
- Section D presents some auxiliary results that we use in all of our proofs.
- Section E provides a proof of Proposition 2: The fact that, in the high complexity limit, characteristic-managed portfolios span the SDF.
- Section F presents some useful results from Random Matrix Theory.
- Section G contains the proof that managed portfolios satisfy the technical conditions of Random Matrix Theory, paving the road for all of our subsequent theoretical analyses.
- Section H contains more technical lemmas for computing moments involving managed portfolios.
- Section I computes $\lim E[\hat{R}^M_{T+1}(z;q;c)]$, proving item i. of Theorem 3.
- Section J develops mathematical techniques for computing RMT quantities arising in AIPT.
- Section K computes $\lim E[(\hat{R}^M_{T+1}(z;q;c))^2]$, proving item ii. of Theorem 3.
- Section L computing the limit of HJD distance, proving item iv. of Theorem 3.

B Neural Network Interpretation of the AIPT Environment

The structure of returns in Assumption 1 has a clear machine-learning interpretation. Imagine for a moment that returns on asset i are generated by a low-dimensional factor model like those common in economic theory,[33]

$$R_{i,t+1} = \beta(X_{i,t})' G_{t+1} + u_{i,t+1}, \tag{32}$$

where $X_{i,t}$ is a vector of J conditioning variables that determines i's conditional betas on a small number K of latent factors, G_{t+1}. If one has no knowledge of the specific functional form for the conditional beta function, one can use a machine learning model to approximate it. For example, a shallow neural network could replace the $K \times 1$ vector $\beta(X_{i,t})$ with the approximation

$$\beta(X_{i,t}) \approx \sum_{p=1}^{P} \xi_p S_{i,t,p} = \underbrace{\Xi}_{K \times P} \underbrace{S_{i,t}}_{P \times 1}, \tag{33}$$

where

$$S_{i,t} = A(\Omega X_{i,t}) = (A(\omega_p' X_{i,t}))_{p=1}^{P}. \tag{34}$$

The neural network model approximates the unknown beta function with a linear combination of "generated conditioning variables" denoted $S_{i,t,p}$. Specifically, each $S_{i,t,p}$ is a basis function that captures nonlinear predictive information in the raw conditioning variables $X_{i,t}$. To build the basis functions, the neural network first generates a $J \times P$ matrix $\Omega = (\omega_p)_{p=1}^{P}$ of weights with rows ω_p to combine the elements of $X_{i,t}$ into P different linear combinations of $\Omega X_{i,t} \in \mathbb{R}^P$. Next, these linear combinations are transformed by a nonlinear

[33] Santos and Veronesi (2004) is an example asset pricing theory that generates a conditional beta formulation along these lines.

activation function $A(x)$, so that we end up with nonlinear features $S_{i,t} = A(\Omega X_{i,t}) \in \mathbb{R}^P$. Then, equation (33) collects the P basis terms into a weighted sum in order to approximate $\beta(X_{i,t})$. The $K \times 1$ vectors ξ_p determine how each nonlinear basis term best contributes to the approximation of each of the K betas. We can write this sum in a matrix form by collecting the basis terms into a $P \times 1$ vector $S_{i,t}$ and the weights into the $K \times P$ matrix Ξ. Universal approximation theory such as Hornik et al. (1989) ensures that the formulation in (33) can accurately approximate the true conditional beta function under regularity conditions.[34]

To tie this back to Assumption 1, we may stack assets' beta coefficients into an $N \times K$ matrix and substitute (33) into (32) to deliver

$$\begin{aligned} &R_{t+1} \approx S_t \tilde{F}_{t+1} + u_{t+1}, \ with \\ &\tilde{F}_{t+1} = \Xi' G_{t+1}, \ \lambda_F = \Xi' E[G_{t+1}]\,. \end{aligned} \tag{35}$$

The key point of this neural network example is that, while Assumption 1 treats the factor loadings S_t as known and potentially high-dimensional, we interpret it as a generic statistical specification that arises from machine learning approximations to an unknown (and likely low-dimensional) factor pricing model.

C Properties of the Infeasible Portfolio

By a direct calculation,[35]

$$\lambda \ = \ E[FF']^{-1}E[F] \ = \ \frac{1}{1 + MaxSR^2} \operatorname{Var}[F]^{-1}E[F]\,, \tag{36}$$

[34] The approximating structure in (33) is analyzed by Gu et al. (2020a) and is a semi-nonparametric extension of the IPCA model in Kelly et al. (2020).

[35] See the Sherman-Morrison formula (77).

where $\text{Var}[F]$ is the covariance matrix of factors and where we have defined

$$MaxSR^2 \;=\; E[F]'\text{Var}[F]^{-1}E[F] \tag{37}$$

to be the maximal achievable unconditional squared Sharpe ratio. Most existing papers perform their analysis assuming that the population moments of the factors are directly observable and, hence, so is the vector of factor risk premia, λ. The corresponding portfolio satisfies

$$E[\lambda' F_{t+1}] \;=\; E[(\lambda' F_{t+1})^2] \;=\; E[F]'E[FF']^{-1}E[F] \;=\; \frac{MaxSR^2}{1+MaxSR^2}\,. \tag{38}$$

It will be instructive for our subsequent analysis to decompose the maximal Sharpe ratio into the contributions coming from the factor principal components. Given the eigenvalue decomposition $\text{Var}[F] \;=\; U\,\text{diag}(\mu)U'$, we can define PC_i to be the i-th column of $U'F$. In the sequel, we will use

$$\theta \;=\; U'E[F] \tag{39}$$

to denote the vector of mean returns of the PCs. Then, we can rewrite the maximal Sharpe ratio (37) as

$$MaxSR^2 \;=\; \sum_i \frac{\theta_i^2}{\mu_i} \;=\; \sum_i (SR(PC_i))^2\,. \tag{40}$$

We will now use this representation to understand the effect of ridge shrinkage on the performance of the *infeasible* efficient portfolio,

$$R_{t+1}^{infeas}(z) \;=\; E[F]'(zI+\text{Var}[F])^{-1}F_{t+1}\,. \tag{41}$$

We call this portfolio *infeasible* because, in the big data regime, when $P > T$, neither $E[F] \in \mathbb{R}^P$ nor $E[FF'] \in \mathbb{R}^{P\times P}$ can be efficiently estimated from only T observations. By construction, $R^{infeas}_{t+1}(0) = \lambda' F_{t+1}$ achieves the $MaxSR$, and

$$\mathcal{E}(z) \;=\; E[R^{infeas}(z)] \;=\; E[F]'(zI + E[FF'])^{-1}E[F] \;=\; \frac{A(z)}{1+A(z)}\,, \tag{42}$$

where we have defined

$$\begin{aligned} A(z) \;&=\; E[F]'(zI + \mathrm{Var}[F])^{-1}E[F] \\ &=\; \sum_i (SR(PC_i))^2 \frac{\mu_i}{\mu_i + z} \\ &=\; \sum_i (SR(PC_i))^2 \frac{1}{1 + z/\mu_i} \;\approx\; \sum_{i:\mu_i > z} (SR(PC_i))^2 \end{aligned} \tag{43}$$

and

$$A'(z) \;=\; -\sum_i \theta_i^2 \frac{1}{(\mu_i + z)^2}\,. \tag{44}$$

The function $A(z)$ will be important in understanding ridge-regularization in the high complexity case. It turns out that the risk of the efficient portfolio can be expressed in terms of the derivative of $A(z)$: Defining

$$(zA(z))' \;=\; \sum_i (SR(PC_i))^2 \left(\frac{\mu_i}{\mu_i + z}\right)^2\,, \tag{45}$$

a somewhat tedious calculation implies that

$$\mathrm{Var}[R^{infeas}(z)] = \;\frac{(zA(z))'}{(1+A(z))^2}\,. \tag{46}$$

and

$$
\begin{aligned}
&E[(R^{infeas}(z))^2] \\
&= \frac{1}{(1+A(z))^2} E[(E[F]'(zI+\Psi)^{-1}F_t)^2] \;=\; \frac{1}{(1+A(z))^2} E[E[F]'(zI+\Psi)^{-1}F_t F_t'(zI+\Psi)^{-1}E[F]] \\
&= \frac{1}{(1+A(z))^2} E[E[F]'(zI+\Psi)^{-1}F_t F_t'(zI+\Psi)^{-1}E[F]] \\
&= E[F]'(zI+\Psi)^{-1}\Psi(zI+\Psi)^{-1}E[F] \;+\; \mathcal{R}_1(z)^2 \\
&= \frac{1}{(1+A(z))^2}\sum_i \theta_i^2 (z+\mu_i)^{-2}\mu_i \;+\; \left(\frac{A(z)}{1+A(z)}\right)^2 \\
&= \frac{A(z)+zA'(z)+A^2(z)}{(1+A(z))^2} \\
&= \frac{(A(z)+zA'(z))(1+A(z))-zA(z)A'(z)}{(1+A(z))^2} \\
&= \frac{d}{dz}\left(\frac{zA(z)}{1+A(z)}\right).
\end{aligned}
\tag{47}
$$

Since the weights $\frac{\mu_i}{\mu_i+z}$ are monotone increasing in μ_i, we see that all that the ridge shrinkage does it re-weights principal components, giving a larger weight to higher-variance PCs. The following is a simple but important observation, implying that ridge shrinkage is always detrimental to performance.

Lemma 1 *The Sharpe ratio* $SR^{infeasible}(z) \;=\; SR(R^{infeasible}(z))$ *is monotone decreasing in* z.

D Auxilliary Results

D.1 Basic Inequalities

Definition 1 (Strongly uncorrelated variables) *We say that* $f_i,\; i \;=\; 1,\cdots,K$ *are strongly uncorrelated if* $E[f_{i_1}] = E[f_{i_1}f_{i_2}] = 0$ *for all* $i_1 \neq i_2$, $E[f_{i_1}f_{i_2}f_{i_3}] = 0$ *for any* i_1, i_2, i_3

and $E[f_{i_1} f_{i_2} f_{i_3} f_{i_4}] = 0$ *unless the set* $\{i_1, i_2, i_3, i_4\}$ *contains exactly two different elements. Furthemore,* $E[f_i^2 f_j^2] = E[f_i^2]E[f_j^2]$ *for* $i \neq j$.

Lemma 2 *Suppose that* $X = (X_i)_{i=1}^P$ *with* X_i *being strongly uncorrelated according to Definition 1. Suppose also that* $E[X_i^2] = 1, E[X_i^4] \leq k$, *and let* A_P *be random matrices independent of* X *and such that* $\|A_P\|_2 = o(1)$. *Let also*

$$Y_t \;=\; X_t' A_P X_t \,. \tag{48}$$

Then,

$$(1)\; Y_t = \mathrm{tr}(A_P X_t X_t') \tag{49}$$

$$(2)\; \lim_{P\to\infty} E[(Y_t - \mathrm{tr}(A_P))^2 | A_P] = 0 \tag{50}$$

In particular, If $A_P = B_P/P$ *where* $\|B_P\| \leq K$, *we have* $\|A_P\|_2^2 \leq P\|B_P\|^2/P^2 \leq K$, *and hence*

$$\lim_{P\to\infty} E[(X_t' B_P X_t - \mathrm{tr}(B_P))^2 | B_P]/P^2 \;=\; 0\,. \tag{51}$$

Proof of Lemma 2.

(1):

$$X_t' A X_t \in R \Rightarrow X_t' A X_t = \mathrm{tr}(X_t' A X_t)$$

$$\mathrm{tr}(AB) = \mathrm{tr}(BA) \Rightarrow \mathrm{tr}(X_t' A X_t) = \mathrm{tr}(A X_t X_t')$$

(2): Define $Y_t = X_t' A_P X_t$. We have

$$E[Y_t] \;=\; E[\mathrm{tr}(A_P(X_t X_t'))|A_P] = \mathrm{tr}(A_P E[X_t X_t']) = \mathrm{tr}(A_P)\,,$$

and hence

$$E[(Y_t - \mathrm{tr}(A_P))^2|A_P] \;=\; \mathrm{Var}[Y_t|A_P] \;=\; E[Y_t^2|A_P] - E[Y_t|A_P]^2 \tag{52}$$

and hence it suffices to prove that

$$E[Y_t^2|A_P] \;- (\mathrm{tr}(A_P))^2 \;\to\; 0 \tag{53}$$

For simplicity, we assume from now on that A_P is deterministic, and write $A_P = (A_{i,j})_{i,j=1}^P$. We also assume that A is *symmetric.* Then,

$$Y_t \;=\; \sum_{i,j} X_i X_j A_{i,j} \tag{54}$$

and therefore

$$Y_t^2 \;=\; \sum_{i_1,j_1,i_2,j_2} X_{i_1} X_{j_1} A_{i_1,j_1} A_{i_2,j_2} X_{i_2} X_{j_2} \tag{55}$$

Now, among all fourth-order moments, $E[X_{i_1} X_{j_1} X_{i_2} X_{j_2}]$, the only non-zero moments are those where either all are identical, $i_1 = i_2 = i_3 = i_4$, or when there are exactly two identical pairs. The latter can happen in exactly 3 ways. First, $(i_1 = i_2, j_1 = j_2)$, $(i_1 = j_2, j_1 = i_2)$ give rise to the terms A_{i_1,j_1}^2 because, by assumption, A is symmetric, so that $A_{i_1,j_1} = A_{j_1,i_1}$. Second, $(i_1 = j_1, i_2 = j_2)$ gives rise to $A_{i,i} A_{j,j}$. Note also that $E[X_i^2 X_j^2] \;= E[X_i^2] E[X_j^2] = 1$.

Thus,

$$
\begin{aligned}
E[Y_t^2] \;&=\; \sum_{i_1,j_1,i_2,j_2} A_{i_1,j_1} A_{i_2,j_2} E[X_{i_1} X_{j_1} X_{i_2} X_{j_2}] \\
&= \sum_i A_{i,i}^2 E[X_i^4] + \sum_{i,j,i\neq j} (2A_{i,j}^2 + A_{i,i} A_{j,j}) \\
&= \;\sum_i A_{i,i}^2 E[X_i^4] + \sum_{i,j,i\neq j} 2A_{i,j}^2 \;-\; \sum_i A_{i,i}^2 \;+\; (\sum_i A_{i,i})^2 \\
&= \;\sum_i A_{i,i}^2 E[X_i^4] - 2\sum_i A_{i,i}^2 + \sum_{i,j} 2A_{i,j}^2 \;-\; \sum_i A_{i,i}^2 \;+\; (\sum_i A_{i,i})^2 \\
&= \;\sum_i A_{i,i}^2 (E[X_i^4] - 3) \;+\; 2\|A\|_2^2 \;+\; (\mathrm{tr}(A))^2
\end{aligned}
\tag{56}
$$

Thus, since $E[Y_t] = \mathrm{tr}(A)$, we have

$$
E[Y_t^2] - E[Y_t]^2 \;=\; \sum_i A_{i,i}^2 (E[X_i^4] - 3) \;+\; 2\|A\|_2^2 \;\leq\; (k-1)\|A\|_2^2 \,. \tag{57}
$$

because

$$
\sum_i A_{i,i}^2 \;\leq\; \sum_{i,j} A_{i,j}^2 \;=\; \|A\|_2^2 \,, \tag{58}
$$

The proof is complete. □

We will need the following lemma, whose proof follows by direct calculation.

Lemma 3 *Suppose that $X_t \in \mathbb{R}^{N\times P}$ is a matrix with i.i.d. elements satisfying $E[X_{i,k} X_{j,l}] = \delta_{(i,k),(j,l)}$. Then,*

$$
E[X_t' \Sigma X_t] \;=\; \mathrm{tr}(\Sigma)\, I_{P\times P} \,.
$$

Lemma 4 *Let A be a symmetric matrix. Then, we have*

$$\begin{aligned} E[S_t' S_t A S_t' S_t] \;&=\; ((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2))\Psi A \Psi + \operatorname{tr}(\Sigma^2)\operatorname{tr}(\Psi A)\Psi \\ &+\; \operatorname{tr}(\Sigma \circ \Sigma)\, \Psi^{1/2}(\kappa - 3)\operatorname{diag}(\Psi^{1/2} A \Psi^{1/2})\Psi^{1/2} \end{aligned} \tag{59}$$

where $\operatorname{diag}(\Psi^{1/2} A \Psi^{1/2})$ *is the diagonal matrix with diagonal coinciding with that of* $\operatorname{diag}(\Psi^{1/2} A \Psi^{1/2})$, *and* $\Sigma \circ \Sigma$ *is the elementwise product of* Σ *with itself.*

Proof. Substituting $S_t = \Sigma^{1/2} X_t \Psi^{1/2}$, we get

$$E[S_t' S_t A S_t' S_t] \;=\; E[S_t' S_t A S_t' S_t] \;=\; \Psi^{1/2} E[X_t' \Sigma X_t \tilde{A} X_t' \Sigma X_t] \Psi^{1/2}\,, \tag{60}$$

where $\tilde{A} = \Psi^{1/2} A \Psi^{1/2}$. Thus, it suffices to consider the case $\Psi = I$ and just compute

$$E[X_t' \Sigma X_t \tilde{A} X_t' \Sigma X_t]\,, \tag{61}$$

in which case the desired identity takes the form

$$\begin{aligned} E[X_t' \Sigma X_t A X_t' \Sigma X_t] \;&=\; ((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2))A + \operatorname{tr}(\Sigma^2)\operatorname{tr}(A) I \\ &+\; \operatorname{tr}(\Sigma \circ \Sigma)\,(\kappa - 3)\operatorname{diag}(A)\,. \end{aligned} \tag{62}$$

Further, it suffices to prove the result for rank-one matrices because any symmetric matrix A can be written as

$$A \;=\; \sum_i \lambda_i \beta_i \beta_i'$$

Thus, suppose that $A = \beta\beta'$. Then,

$$\begin{aligned} & E[X'\Sigma X\beta\beta' X'\Sigma X]_{j,k} \\ &= E[\sum_{i_1,i_2,i_3,i_4,i_5,i_6} X_{i_1,j}\Sigma_{i_1,i_2}X_{i_2,i_3}\beta_{i_3}\beta_{i_4}X_{i_5,i_4}\Sigma_{i_5,i_6}X_{i_6,k}] \,. \end{aligned} \tag{63}$$

By assumption, all elements of X are i.i.d. and have mean zero. Thus, the only non-zero terms come from two identical pairs or when all terms are identical. For two identical pairs, they are determined by what coincides with (i_1, j). These are the posibilities: $(i_1, j) = (i_2, i_3)$ or $(i_1, j) = (i_5, i_4)$ or $(i_1, j) = (i_6, k)$. The latter can only happen when $j = k$.

- Suppose first that $j \neq k$. Then,

$$\begin{aligned} & E[\sum_{i_1,i_2,i_3,i_4,i_5,i_6} X_{i_1,j}\Sigma_{i_1,i_2}X_{i_2,i_3}\beta_{i_3}\beta_{i_4}X_{i_5,i_4}\Sigma_{i_5,i_6}X_{i_6,k}] \\ &= E[\sum_{i_1=i_2,j=i_3,i_5=i_6,i_4=k} X_{i_1,j}\Sigma_{i_1,i_2}X_{i_2,i_3}\beta_{i_3}\beta_{i_4}X_{i_5,i_4}\Sigma_{i_5,i_6}X_{i_6,k}] \\ &+ E[\sum_{i_1=i_5,j=i_4,i_2=i_6,i_3=k} X_{i_1,j}\Sigma_{i_1,i_2}X_{i_2,i_3}\beta_{i_3}\beta_{i_4}X_{i_5,i_4}\Sigma_{i_5,i_6}X_{i_6,k}] \\ &= (\mathrm{tr}(\Sigma)^2 + \mathrm{tr}(\Sigma^2))\beta_j\beta_k \;=\; (\mathrm{tr}(\Sigma)^2 + \mathrm{tr}(\Sigma^2))A_{j,k} \,. \end{aligned} \tag{64}$$

- When $j = k$, we get

$$
\begin{aligned}
& E[\sum_{i_1,i_2,i_3,i_4,i_5,i_6} X_{i_1,j}\Sigma_{i_1,i_2}X_{i_2,i_3}\beta_{i_3}\beta_{i_4}X_{i_5,i_4}\Sigma_{i_5,i_6}X_{i_6,j}] \\
&= E[\sum_{i_1,i_2,i_3,i_4,i_5} X^2_{i_1,j}\Sigma_{i_1,i_2}X_{i_2,i_3}\beta_{i_3}\beta_{i_4}X_{i_5,i_4}\Sigma_{i_5,i_1}] \\
&+ E[\sum_{i_1,i_2,i_3,i_4,i_5,i_6\neq i_1} X_{i_1,j}\Sigma_{i_1,i_2}X_{i_2,i_3}\beta_{i_3}\beta_{i_4}X_{i_5,i_4}\Sigma_{i_5,i_6}X_{i_6,j}] \\
&= E[\sum_{i_1,i_2,i_3} X^2_{i_1,j}\Sigma_{i_1,i_2}X^2_{i_2,i_3}\beta^2_{i_3}\Sigma_{i_2,i_1}] \\
&+ E[\sum_{i_1,i_2,i_3,i_4,i_5,i_6\neq i_1} X_{i_1,j}\Sigma_{i_1,i_2}X_{i_2,i_3}\beta_{i_3}\beta_{i_4}X_{i_5,i_4}\Sigma_{i_5,i_6}X_{i_6,j}] \\
&= \mathrm{tr}(\Sigma^2)\sum_i \beta_i^2 \ + \ (\kappa-1)\sum_i \Sigma^2_{i,i}\beta_j^2 \\
&+ E[\sum_{i_1,i_2,i_3,i_4,i_5,i_6\neq i_1} X_{i_1,j}\Sigma_{i_1,i_2}X_{i_2,i_3}\beta_{i_3}\beta_{i_4}X_{i_5,i_4}\Sigma_{i_5,i_6}X_{i_6,j}] \\
&= \mathrm{tr}(\Sigma^2)\sum_i \beta_i^2 + \ (\kappa-1)\sum_i \Sigma^2_{i,i}\beta_j^2 \\
&+ E[\sum_{i_1=i_2,i_3=j,i_4=j,i_5=i_6\neq i_1} X_{i_1,j}\Sigma_{i_1,i_2}X_{i_2,i_3}\beta_{i_3}\beta_{i_4}X_{i_5,i_4}\Sigma_{i_5,i_6}X_{i_6,j}] \\
&+ E[\sum_{i_1=i_5,i_3=j,i_4=j,i_2=i_6\neq i_1} X_{i_1,j}\Sigma_{i_1,i_2}X_{i_2,i_3}\beta_{i_3}\beta_{i_4}X_{i_5,i_4}\Sigma_{i_5,i_6}X_{i_6,j}] \\
&= \mathrm{tr}(\Sigma^2)\sum_i \beta_i^2 + \ (\kappa-1)\sum_i \Sigma^2_{i,i}\beta_j^2 \\
&+ \beta_j^2\sum_{i_1}\Sigma_{i_1,i_1}(\mathrm{tr}(\Sigma)-\Sigma_{i_1,i_1}) \\
&+ \beta_j^2\sum_{i_1}\sum_{i_2\neq i_1}\Sigma^2_{i_1,i_2} \\
&= \mathrm{tr}(\Sigma^2)\sum_i \beta_i^2 + \ (\kappa-3)\sum_i \Sigma^2_{i,i}\beta_j^2 + \beta_j^2(\mathrm{tr}(\Sigma))^2 \\
&+ \beta_j^2\sum_{i_1}(\sum_{i_2}\Sigma^2_{i_1,i_2} - \Sigma^2_{i_1,i_1}) \\
&= \mathrm{tr}(\Sigma^2)\sum_i \beta_i^2 + \ (\kappa-3)\sum_i \Sigma^2_{i,i}\beta_j^2 + \beta_j^2((\mathrm{tr}(\Sigma))^2 + \mathrm{tr}(\Sigma^2))
\end{aligned}
\tag{65}
$$

Since $j = k$, the latter can be rewritten as

$$\operatorname{tr}(\Sigma^2)\operatorname{tr}(A)\delta_{j,k} \;+\; (\kappa - 3)\operatorname{tr}(\Sigma \circ \Sigma)A_{j,k} + ((\operatorname{tr}(\Sigma))^2 + \operatorname{tr}(\Sigma^2))A_{j,k}\,. \tag{66}$$

□

Lemma 5 *Let ε be a random vector with i.i.d. coordinates, satisfying $E[\varepsilon] = 0$, and $E[\varepsilon\varepsilon'] = I$, and $E[\varepsilon_i^4] = \kappa_\varepsilon$. We have*

$$E[\varepsilon Z'\varepsilon] \;=\; Z$$

and

$$E[\varepsilon' Z\varepsilon'] \;=\; Z'$$

for any vector Z.

$$E[\varepsilon' A\varepsilon] \;=\; \operatorname{tr}(A)$$

for any matrix A. Furthermore, for any matrix B, we have

$$E[(\varepsilon' B\varepsilon)^2] \;=\; \sum_i \tilde{B}_{i,i}^2(\kappa_\varepsilon - 3) \;+\; 2\operatorname{tr}(\tilde{B}^2) \;+\; (\operatorname{tr}(\tilde{B}))^2 \tag{67}$$

where $\tilde{B} = 0.5(B + B')$, and

$$E[\varepsilon_t\varepsilon_t' B\varepsilon_t\varepsilon_t'] \;=\; 2\tilde{B} \;+\; \operatorname{diag}((\kappa_\varepsilon - 3)\operatorname{diag}(B) + \operatorname{tr}(B))$$

Similarly,

$$E[\varepsilon' B\varepsilon\varepsilon'] \;=\; \mathrm{diag}(B)E[\varepsilon_j^3]\,. \tag{68}$$

Proof. We have

$$E[\varepsilon Z'\varepsilon]_{i,j} \;=\; E[\varepsilon_i \sum_j Z_j\varepsilon_j] \;=\; \sum_j \Sigma_{\varepsilon,i,j} Z_j$$

and the first claim follows. The second claim follows because

$$E[\varepsilon' Z\varepsilon'] \;=\; E[\varepsilon Z'\varepsilon]'\,.$$

For the third claim, we have

$$E[\varepsilon' A\varepsilon] \;=\; \mathrm{tr}\, E[\varepsilon' A\varepsilon] = \mathrm{tr}\, E[A\varepsilon\varepsilon'] = \mathrm{tr}(A) \tag{69}$$

while (67) follows from (56). For the last claim, we make the observation that, for any matrix B,

$$\varepsilon' B\varepsilon \;=\; 0.5\varepsilon'(B+B')\varepsilon.$$

For $j \neq k$, we have

$$\begin{aligned} &E[\varepsilon\varepsilon'\tilde{B}\varepsilon\varepsilon']_{j,k} \\ &=\; E[\varepsilon_j \sum_{i_1,i_2} \varepsilon_{i_1}\varepsilon_{i_2}\tilde{B}_{i_1,i_2}\varepsilon_k] \;=\; 2\tilde{B}_{j,k} \end{aligned} \tag{70}$$

while, for $j = k$, we have

$$\begin{aligned} & E[\varepsilon\varepsilon'\tilde{B}\varepsilon\varepsilon']_{j,j} \\ &= E[\varepsilon_j^2 \sum_{i_1,i_2} \varepsilon_{i_1}\varepsilon_{i_2}\tilde{B}_{i_1,i_2}] \\ &= \kappa_\varepsilon \tilde{B}_{j,j} + \sum_{i\neq j} \tilde{B}_{i,i} = (\kappa_\varepsilon - 1)\tilde{B}_{j,j} + \mathrm{tr}(\tilde{B})\,. \end{aligned} \tag{71}$$

Similarly,

$$E[\varepsilon' B\varepsilon\varepsilon']_j = E[\sum_{i_1,i_2} B_{i_1,i_2}\varepsilon_{i_1}\varepsilon_{i_2}\varepsilon_j] = B_{j,j}E[\varepsilon_j^3] \tag{72}$$

□

D.2 Moments of Managed Portfolios

Recall that

$$F_{t+1} = S_t' R_{t+1} \tag{73}$$

Lemma 6 (Expected Factor Moments) *Suppose a normalization* $\mathrm{tr}(\Sigma) = 1$ *and let* $\sigma_* = \mathrm{tr}(\Sigma\Sigma_\varepsilon)$ *and* $E[X_{i,k}^4] = \kappa$ *for all* i, k. *We have*

$$E[S_t'\Sigma_\varepsilon S_t] = \mathrm{tr}(\Sigma\Sigma_\varepsilon)\Psi$$

and

$$\begin{aligned} E[F_{t+1}F_{t+1}'] &= ((\mathrm{tr}\,\Sigma)^2 + \mathrm{tr}(\Sigma^2))\Psi\Sigma_F\Psi \\ &+ \mathrm{tr}(\Sigma\circ\Sigma)\Psi^{1/2}(\kappa - 3)\,\mathrm{diag}(\Psi^{1/2}\Sigma_F\Psi^{1/2})\Psi^{1/2} + \Psi\Big(\mathrm{tr}(\Sigma\Sigma_\varepsilon) + \mathrm{tr}(\Psi\Sigma_F)\,\mathrm{tr}(\Sigma^2)\Big) \end{aligned} \tag{74}$$

Thus,

$$\|E[F_{t+1}F_{t+1}'] - (\Psi\Sigma_F\Psi + \sigma_*\Psi)\| \;\to\; 0 \tag{75}$$

when $P \to \infty$.

Proof of Lemma 6. Recall that, under the normalization $\operatorname{tr}(\Sigma) = 1$, by Assumption 2, $\operatorname{tr}(\Sigma^2) \to 0$ and $\operatorname{tr}(\Sigma \circ \Sigma) = \sum_i \Sigma_{i,i}^2 \le \sum_{i,j} \Sigma_{i,j}^2 = \operatorname{tr}(\Sigma^2) \to 0$. We have

$$E[F_{t+1}F_{t+1}'] \;=\; E[S_t'(S_t\widetilde{F} + \varepsilon)(S_t\widetilde{F} + \varepsilon)'S_t] \;=\; E[S_t'S_t\Sigma_F S_t'S_t] \;+\; E[S_t'\Sigma_\varepsilon S_t]\,,$$

and

$$E[S_t'\Sigma_\varepsilon S_t] \;=\; E[\Psi^{1/2}X_t'\Sigma^{1/2}\Sigma_\varepsilon\Sigma^{1/2}X_t\Psi^{1/2}] \;=\; \Psi^{1/2}E[X_t'\Sigma^{1/2}\Sigma_\varepsilon\Sigma^{1/2}X_t]\Psi^{1/2} \;=\; \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon) = \Psi\sigma_*\,,$$

while Lemma 4 implies that

$$\begin{aligned}
&E[S_t'(S_t\widetilde{F} + \varepsilon)(S_t\widetilde{F} + \varepsilon)'S_t] \\
&=\; ((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2))\Psi\Sigma_F\Psi \\
&+\; \operatorname{tr}(\Sigma\circ\Sigma)\Psi^{1/2}(\kappa - 3)\operatorname{diag}(\Psi^{1/2}\Sigma_F\Psi^{1/2})\Psi^{1/2} + \Psi + \operatorname{tr}(\Psi\Sigma_F)\operatorname{tr}(\Sigma^2)
\end{aligned} \tag{76}$$

The claim how follows because $\operatorname{tr}(\Sigma \circ \Sigma)$ and $\operatorname{tr}(\Sigma^2)$ converge to zero, and Ψ is uniformly bounded when $P \to \infty$ by assumption. The proof of Lemma 6 is complete.

□

E Proof of Proposition 2: Characteristic-managed Portfolios and the Conditional SDF

We will frequently be using the Sherman-Morrison formula

$$(A+xx')^{-1} \;=\; A^{-1}-A^{-1}xx'A^{-1}/(1+x'A^{-1}x),\; (A+xx')^{-1}x \;=\; A^{-1}x/(1+x'A^{-1}x) \tag{77}$$

for any matrix $A \in \mathbb{R}^{P\times P}$ and any vector $x \in \mathbb{R}^P$.

Lemma 7 *We have*

$$(A+B)^{-1} \;=\; B^{-1} - (A+B)^{-1}AB^{-1}\,, \tag{78}$$

and

$$(A+B)^{-1}AB^{-1} \;\leq\; A \tag{79}$$

in the sense of positive semi-definite order.

Proof of Lemma 7. Let $\hat{A} = B^{-1/2}AB^{-1/2}$. Then, we have

$$(A+B)^{-1}AB^{-1} \;=\; B^{-1/2}(\hat{A}+I)^{-1}\hat{A}B^{-1/2} \;\leq\; B^{-1/2}\hat{A}B^{-1/2} = B^{-1}AB^{-1}\,. \tag{80}$$

□

Proof of Proposition 2. Recall that, by Proposition 1,

$$\tilde{w}(S_t) \;=\; (S_t\Sigma_{F,t}S_t' + \Sigma_\varepsilon)^{-1}S_t\nu_F \tag{81}$$

is the conditionally efficient portfolio with the return

$$R'_{t+1}\tilde{w}(S_t) \;=\; F'_{t+1}(S_t\Sigma_{F,t}S'_t \;+\; \Sigma_\varepsilon)^{-1}S_t\nu_F\,. \tag{82}$$

For simplicity, in the sequal omit the t subindex for Σ_F and Σ_F^*. We have

$$((\Sigma_F)^{-1} + S'_tS_t)^{-1} \;\leq\; ((\Sigma_F)^{-1})^{-1}$$

Hence, defining

$$Q_t \;=\; (S_t\Sigma_F^*S'_t \;+\; \Sigma_\varepsilon)^{-1} \;=\; \Sigma_\varepsilon^{-1} \;-\; (S_t\Sigma_F^*S'_t \;+\; \Sigma_\varepsilon)^{-1}S_t\Sigma_F^*S'_t\Sigma_\varepsilon^{-1}\,, \tag{83}$$

we get

$$\begin{aligned}
&E[R'_{t+1}\tilde{w}(S_t)]\\
&=\; E[(S_t\widetilde{F}_{t+1} \;+\; \varepsilon_{t+1})'(S_t(\Sigma_F)S'_t \;+\; \Sigma_\varepsilon)^{-1}S_t\nu_F]\\
&=\; E[\nu'_FS'_t(S_t(\Sigma_F)S'_t \;+\; \Sigma_\varepsilon)^{-1}S_t\nu_F]\\
&=\; E[\nu'_FS'_t(S_t(\nu_F\nu'_F + \Sigma_F^*)S'_t \;+\; \Sigma_\varepsilon)^{-1}S_t\nu_F]\\
&=\; E[\nu'_FS'_t((S_t\nu_F)(S_t\nu_F)' + (S_t\Sigma_F^*S'_t \;+\; \Sigma_\varepsilon))^{-1}S_t\nu_F]\\
&\underbrace{=}_{(77)}\; E[\nu'_FS'_t(Q_t - Q_tS_t\nu_F\nu'_FS'_tQ_t(1 + \nu'_FS'_tQ_tS_t\nu_F)^{-1})S_t\nu_F]\\
&=\; E[Z_t - Z_t^2(1 + Z_t)^{-1}] \;=\; E[Z_t/(1 + Z_t)]\,,
\end{aligned} \tag{84}$$

where we have defined

$$Z_t \;=\; \nu'_FS'_tQ_tS_t\nu_F \;=\; \nu'_FS'_t\Sigma_\varepsilon^{-1}S_t\nu_F - q\,, \tag{85}$$

with

$$q \;=\; \nu_F' S_t' \Sigma_\varepsilon^{-1} S_t \nu_F - \nu_F' S_t' Q_t S_t \nu_F \,. \tag{86}$$

By Lemma 7,

$$(S_t \Sigma_F^* S_t' \;+\; \Sigma_\varepsilon)^{-1} S_t \Sigma_F^* S_t' \Sigma_\varepsilon^{-1} \;\leq\; \Sigma_\varepsilon^{-1} S_t \Sigma_F^* S_t' \Sigma_\varepsilon^{-1} \tag{87}$$

and hence

$$q \;=\; \nu_F' S_t' (S_t \Sigma_F^* S_t' \;+\; \Sigma_\varepsilon)^{-1} S_t \Sigma_F^* S_t' \Sigma_\varepsilon^{-1} S_t \nu_F \;\leq\; \nu_F' S_t' \Sigma_\varepsilon^{-1} S_t \Sigma_F^* S_t' \Sigma_\varepsilon^{-1} S_t \nu_F \,. \tag{88}$$

For simplicity, we will assume that $X_{i,k,t}$ all have the same fourth moment κ (otherwise, the identity needs to be replaced by an inequality). Then, we have that, by Lemma 4,

$$\begin{aligned}
E[\nu_F' S_t' \Sigma_\varepsilon^{-1} S_t A S_t' \Sigma_\varepsilon^{-1} S_t \nu_F] \;&=\; \nu_F' \Big(((\operatorname{tr} \hat{\Sigma})^2 + \operatorname{tr}(\hat{\Sigma}^2)) \Psi A \Psi + \operatorname{tr}(\hat{\Sigma}^2) \operatorname{tr}(\Psi A) \Psi \\
&+\; \operatorname{tr}(\hat{\Sigma}^2) \, (\kappa - 3) \Psi^{1/2} \operatorname{diag}(\Psi^{1/2} A \Psi^{1/2}) \Psi^{1/2} \Big) \nu_F \\
&=\; (\operatorname{tr} \hat{\Sigma})^2 \nu_F' \Big((1 + \frac{\operatorname{tr}(\hat{\Sigma}^2)}{(\operatorname{tr} \hat{\Sigma})^2}) \Psi A \Psi + \frac{\operatorname{tr}(\hat{\Sigma}^2)}{(\operatorname{tr} \hat{\Sigma})^2} \operatorname{tr}(\Psi A) \Psi \\
&+\; \frac{\operatorname{tr}(\hat{\Sigma} \circ \hat{\Sigma})}{(\operatorname{tr} \hat{\Sigma})^2} \, (\kappa - 3) \Psi^{1/2} \operatorname{diag}(\Psi^{1/2} A \Psi^{1/2}) \Psi^{1/2} \Big) \nu_F
\end{aligned} \tag{89}$$

with

$$A \;=\; \Sigma_F^* \tag{90}$$

and

$$\hat{\Sigma} \;=\; \Sigma^{1/2}\Sigma_\varepsilon^{-1}\Sigma^{1/2}\,. \tag{91}$$

If $\Sigma_\varepsilon = I$ then, by Assumption 2, $\frac{\mathrm{tr}(\hat{\Sigma}^2)+\mathrm{tr}(\hat{\Sigma}\circ\hat{\Sigma})}{(\mathrm{tr}\,\hat{\Sigma})^2} \to 0$ and, since ν_F and Ψ and A and $\mathrm{tr}(\hat{\Sigma})$ are uniformly bounded, we get that

$$E[\nu_F' S_t' \Sigma_\varepsilon^{-1} S_t A S_t' \Sigma_\varepsilon^{-1} S_t \nu_F] \;\approx (\mathrm{tr}\,\hat{\Sigma})^2 \nu_F' \Psi A \Psi \nu_F\,. \tag{92}$$

Since, by Assumption 1, $\mathrm{tr}(A)$ is uniformly bounded, we also get that $\mathrm{tr}(\Psi A \Psi) \le \|\Psi\|^2 \,\mathrm{tr}(A)$ is uniformly bounded and, hence, $\nu_F'\Psi A\Psi\nu_F \to 0$ by (22).

Thus, $E[q_t] \to 0$ and hence $q_t \to 0$ in probability. Now,

$$E[\nu_F' S_t' \Sigma_\varepsilon^{-1} S_t \nu_F] \;=\; \mathrm{tr}(\hat{\Sigma})\nu_F'\Psi\nu_F \tag{93}$$

whereas, by Lemma 4,

$$\begin{aligned} E[(\nu_F' S_t' \Sigma_\varepsilon^{-1} S_t \nu_F)^2] \;&=\; E[\nu_F' S_t' \Sigma_\varepsilon^{-1} S_t \nu_F \nu_F' S_t' \Sigma_\varepsilon^{-1} S_t \nu_F] \\ &=\; \nu_F' \Bigg(((\mathrm{tr}\,\hat{\Sigma})^2 + \mathrm{tr}(\hat{\Sigma}^2)) \Psi\nu_F\nu_F'\Psi + \mathrm{tr}(\hat{\Sigma}^2)\,\mathrm{tr}(\Psi\nu_F\nu_F')\Psi \\ &+\; \mathrm{tr}(\hat{\Sigma}\circ\hat{\Sigma})\,(\kappa-3)\Psi^{1/2}\,\mathrm{diag}(\Psi^{1/2}\nu_F\nu_F'\Psi^{1/2})\Psi^{1/2} \Bigg)\nu_F \end{aligned} \tag{94}$$

and the same argument as in (89) implies that

$$E[(\nu_F' S_t' \Sigma_\varepsilon^{-1} S_t \nu_F)^2] \;\approx\; \mathrm{tr}(\hat{\Sigma})^2 (\nu_F'\Psi\nu_F)^2\,. \tag{95}$$

Thus, $\mathrm{Var}[\nu_F' S_t' \Sigma_\varepsilon^{-1} S_t \nu_F] \to 0$ and, hence, $\nu_F' S_t' \Sigma_\varepsilon^{-1} S_t \nu_F \to \mathrm{tr}(\hat{\Sigma})\nu_F'\Psi\nu_F$ in probability.

As a result, $Z_t - \mathrm{tr}(\hat{\Sigma})\nu_F'\Psi\nu_F \to 0$ is probability, and hence

$$\frac{Z_t}{1+Z_t} \;-\; \frac{\mathrm{tr}(\hat{\Sigma})\nu_F'\Psi\nu_F}{1+\mathrm{tr}(\hat{\Sigma})\nu_F'\Psi\nu_F} \;\to\; 0 \tag{96}$$

in probability, and the dominated convergence theorem implies that the same holds in expectation. Similarly, for the second moment, we have

$$\begin{aligned}
&E[(\pi_t^{MV})' R_{t+1} R_{t+1}' \pi_t^{MV}] \\
&= E[\lambda' S_t'(S_t(\Sigma_F)S_t' \;+\; \Sigma_\varepsilon)^{-1}(S_t(\Sigma_F)S_t' \;+\; \Sigma_\varepsilon)(S_t(\Sigma_F)S_t' \;+\; \Sigma_\varepsilon)^{-1} S_t\lambda] \\
&= E[R_{t+1}'\pi_t^{MV}] \;\to\; \frac{\mathrm{tr}(\hat{\Sigma})\nu_F'\Psi\nu_F}{1+\mathrm{tr}(\hat{\Sigma})\nu_F'\Psi\nu_F}\,.
\end{aligned} \tag{97}$$

Now, for the factor portfolios, we have

$$\begin{aligned}
E[F_t] \;&=\; E[S_t'R_{t+1}] \;=\; E[S_t'(S_t\widetilde{F}_{t+1} \;+\; \varepsilon_{t+1})] \\
&= E[S_t'S_t\widetilde{F}_{t+1}] \;=\; E[\Psi^{1/2}X_t'\Sigma X_t\Psi^{1/2}\widetilde{F}_{t+1}] \\
&= E[\Psi^{1/2}X_t'\Sigma X_t\Psi^{1/2}]\nu_F \\
&= \mathrm{tr}(\Sigma)\, E[\Psi^{1/2}\Psi^{1/2}]\nu_F \\
&= \mathrm{tr}(\Sigma)\,\Psi\nu_F\,,
\end{aligned} \tag{98}$$

and, again by Lemma 4 and the same argument as in (89), we have

$$\begin{aligned}
E[F_tF_t'] \;&=\; E[S_t'(S_t\widetilde{F}_{t+1} \;+\; \varepsilon_{t+1})(S_t\widetilde{F}_{t+1} \;+\; \varepsilon_{t+1})'S_t|\lambda] \;=\; E[S_t'(S_t(\Sigma_F)S_t' \;+\; \Sigma_\varepsilon)S_t] \\
&\approx\; \mathrm{tr}(\Sigma\Sigma_\varepsilon)\Psi \;+\mathrm{tr}(\Sigma)^2\Psi(\Sigma_F)\Psi\,.
\end{aligned} \tag{99}$$

Then, defining

$$Q \;=\; (\mathrm{tr}(\Sigma\Sigma_\varepsilon)\Psi \;+\; \mathrm{tr}(\Sigma)^2\Psi\Sigma_F^*\Psi)^{-1}\,, \tag{100}$$

we get that the efficient portfolio of factors is given by

$$\begin{aligned}\pi_F &= (\mathrm{tr}(\Sigma\Sigma_\varepsilon)\Psi \ + \ \mathrm{tr}(\Sigma)^2\Psi\Sigma_F\Psi)^{-1}\Psi\nu_F \\ &= \ (\mathrm{tr}(\Sigma\Sigma_\varepsilon)\Psi \ + \ \mathrm{tr}(\Sigma)^2\Psi(\Sigma_F^* + \nu_F\nu_F')\Psi)^{-1}\Psi\nu_F \\ &\underbrace{=}_{(77)} \ \frac{1}{1+Z}Q\Psi\nu_F\,,\end{aligned} \tag{101}$$

where

$$Z \ = \ \mathrm{tr}(\Sigma)^2\nu_F'\Psi Q\Psi\nu_F\,. \tag{102}$$

By the same argument as above,

$$\nu_F'(\Psi Q\Psi - \Psi(\mathrm{tr}(\Sigma\Sigma_\varepsilon)\Psi)^{-1}\Psi)\nu_F \ \to 0 \tag{103}$$

by Assumption 22 because Σ_*^F has a bounded trace. Thus,

$$Z \ \approx \ \frac{\mathrm{tr}(\Sigma)^2}{\mathrm{tr}(\Sigma\Sigma_\varepsilon)}\nu_F'\Psi\nu_F \tag{104}$$

and

$$E[\pi_F'F_{t+1}] \ = \ E[\nu_F'\frac{1}{1+Z}\Psi Q\Psi\nu_F] \approx \ \frac{Z}{1+Z}\,, \tag{105}$$

while

$$E[\pi_F'F_{t+1}F_{t+1}'\pi_F] \ = \ E[\pi_F'F_{t+1}]\,, \tag{106}$$

and the proof is complete because

$$\mathrm{tr}(\Sigma\Sigma_\varepsilon^{-1})\nu_F'\Psi\nu_F \;=\; \frac{\mathrm{tr}(\Sigma)^2}{\mathrm{tr}(\Sigma\Sigma_\varepsilon)}\nu_F'\Psi\nu_F \tag{107}$$

when $\Sigma_\varepsilon = I$.

Finally, the fact that $E[(\pi_F'F_{t+1} - R_{t+1}'\tilde{w}(S_t))^2] \to 0$ follows because, otherwise, one could construct a better-diversified portfolio by combining the two, which is impossible. Namely, we know that

$$E[R_{t+1}'\tilde{w}(S_t) - 0.5((R_{t+1}'\tilde{w}(S_t))^2] \;\geq\; E[R_{t+1}'b_t - 0.5((R_{t+1}'b_t^2] \tag{108}$$

for any portfolio policy b_t because $\tilde{w}(S_t)$ is optimal for the agent with quadratic utility. Let $b_t \;=\; \tilde{w}(S_t) + \varepsilon S_t\pi_F$. Then,

$$\max_\varepsilon E[R_{t+1}'b_t - 0.5((R_{t+1}'b_t)^2] \tag{109}$$

is achieved with

$$\varepsilon \;=\; \frac{E[\pi_F'F_{t+1}] - E[\pi_F'F_{t+1}R_{t+1}'\tilde{w}(S_t)]}{E[(\pi_F'F_{t+1})^2]}. \tag{110}$$

Since the corresponding utility gain is zero, we must have

$$E[\pi_F'F_{t+1}] - E[\pi_F'F_{t+1}R_{t+1}'\tilde{w}(S_t)] \;=\; 0\,. \tag{111}$$

Thus,

$$\begin{aligned} E[(\pi_F' F_{t+1} - R_{t+1}' \tilde{w}(S_t))^2] &= E[(\pi_F' F_{t+1})^2] + E[(R_{t+1}' \tilde{w}(S_t))^2] - 2E[\pi_F' F_{t+1} R_{t+1}' \tilde{w}(S_t)] \\ &= E[(\pi_F' F_{t+1})^2] + E[(R_{t+1}' \tilde{w}(S_t))^2] - 2E[\pi_F' F_{t+1}] \\ &\approx \frac{Z}{1+Z} + \frac{Z}{1+Z} - 2\frac{Z}{1+Z} = 0\,. \end{aligned} \tag{112}$$

The proof is complete. □

F Random Matrix Theory: Auxiliary Results

Let

$$\hat{\lambda}(z) = (zI + B_T)^{-1} \frac{1}{T} \sum_{t=1}^{T} F_t \tag{113}$$

where

$$B_T = \frac{1}{T} \sum_{t=1}^{T} F_t F_t'\,, \tag{114}$$

while

$$\hat{R}_{T+1}^M(z) = \hat{\lambda}(z)' F_{t+1} = (S_t \hat{\lambda}(z))' R_{t+1}\,. \tag{115}$$

In the sequel, to simplify some expressions, we often assume that factor risk premia $\nu_F \sim N(0, \Sigma_\lambda/P)$ for some uniformly bounded sequence of matrices $\Sigma_\lambda = \Sigma_\lambda(P)$. The general case described in Assumption 4 follows from the following remarkable result from Knowles and Yin (2017).

Theorem 4 (Knowles and Yin (2017)) *For any bounded vector β, we have*

$$\underbrace{\lambda\,\beta'(\Psi^{1/2}ZZ'\Psi^{1/2}/t+\lambda I)^{-1}\beta}_{random\ \ (through\ \ Z)}\ \approx\ \underbrace{-\beta'(\Psi r_\Psi(\lambda;c)+I)^{-1}\beta}_{deterministic}, \tag{116}$$

where r is the unique solution to

$$\frac{1}{r}\ =\ -\lambda\ +\ \frac{1}{t}\operatorname{tr}(\Psi(I+r\Psi)^{-1})\,. \tag{117}$$

It is related to the Stieltjes transform

$$m(z;c)\ =\ (1-c+zr(z;c))/(cz), \tag{118}$$

which solves

$$m(z;c)\ =\ N^{-1}\operatorname{tr}((\Psi(1-c-czm)-zI)^{-1}). \tag{119}$$

See, also, (Hastie et al., 2019) for more detailed results that even hold for finite values of P.

In this case,

$$\nu_F'A\nu_F\ \approx P^{-1}\operatorname{tr}(A\Sigma_\lambda) \tag{120}$$

in probability (and in L_2). All our results hold under the more general condition (22), and all expressions can be rewritten without Σ_λ using (120).

Lemma 8 *We have*

$$(\tilde F_{t+1}'A_P\tilde F_{t+1}\ -\ \operatorname{tr}((\Sigma_{F,t}A_P)+P^{-1}\operatorname{tr}(A_P\Sigma_\lambda)))\ \to\ 0 \tag{121}$$

is L_2 and hence in probability, for any sequence of bounded matrices A_P.

Proof of Lemma 8. The proof follows directly from Lemma 2. □

We will also need the following Lemma from KMZ.

Lemma 9 *We have*

$$P^{-1}\operatorname{tr}(A_1(zI+B_T)^{-1}A_2) \;-\; P^{-1}\operatorname{tr}E[A_1(zI+B_T)^{-1}A_2] \;\to\; 0$$

almost surely for any bounded $A_1,\ A_2$ *that are independent of* F_t.

Lemma 10 *Let*

$$\frac{1}{T}\operatorname{tr}((zI+B_T)^{-1}\Psi\sigma_*) \;\to\; \xi(z;c) \tag{122}$$

almost surely and

$$\frac{1}{T}F_t'(zI+B_{T,t})^{-1}F_t \;\to\; \xi(z;c)\,, \tag{123}$$

in probability, where

$$\frac{c^{-1}\xi(z;c)}{1+\xi(z;c)} \;=\; 1 \;-\; m(-z;c)z \tag{124}$$

Proof. First, Lemma 14 implies that

$$\frac{1}{T}F_t'(zI+B_{T,t})^{-1}F_t \;-\; \frac{1}{T}\operatorname{tr}((zI+B_{T,t})^{-1}E[F_tF_t']) \;\to\; 0\,.$$

in probability. Next Lemma 9 applied to our setting implies that for any bounded matrix Q_T independent of $B_{T,t}$ we have

$$\frac{1}{T}\operatorname{tr}((zI+B_{T,t})^{-1}Q_T) \;-\; \frac{1}{T}E[\operatorname{tr}((zI+B_{T,t})^{-1}Q_T)] \;\to\; 0$$

almost surely. At the same time, by Lemma 6,

$$\begin{aligned} E[F_t F_t'] &= ((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2))\Psi\Sigma_F\Psi \\ &+ \operatorname{tr}(\Sigma^2)(\kappa-3)\Psi^{1/2}\operatorname{diag}(\Psi^{1/2}\Sigma_F\Psi^{1/2})\Psi^{1/2} + \Psi\Big(\operatorname{tr}(\Sigma\Sigma_\varepsilon) + \operatorname{tr}(\Psi\Sigma_F)\operatorname{tr}(\Sigma^2)\Big) \end{aligned} \tag{125}$$

We have

$$\frac{1}{T}\operatorname{tr}((zI + B_{T,t})^{-1}(\operatorname{tr}\Sigma)^2\Psi\Sigma_{F,t}\Psi) = O(1/T) \tag{126}$$

The same argument applies to the second term because the trace of

$$\operatorname{tr}(\Sigma^2)(\kappa-3)\Psi^{1/2}\operatorname{diag}(\Psi^{1/2}\Sigma_F\Psi^{1/2})\Psi^{1/2}$$

is also uniformly bounded. Thus, we get

$$\begin{aligned} &\frac{1}{T}F_t'(zI + B_{T,t})^{-1}F_t \sim \frac{1}{T}\operatorname{tr}((zI + B_{T,t})^{-1}E[F_tF_t']) \\ &\sim T^{-1}\operatorname{tr}[(zI + B_{T,t})^{-1}\Psi\sigma_*] \to \xi(z;c)\,. \end{aligned} \tag{127}$$

Now, we have

$$\begin{aligned} 1 &= P^{-1}\operatorname{tr}E[(zI + B_T)^{-1}(zI + B_T)] \\ &= zm(-z;c) + \frac{1}{P}\operatorname{tr}\frac{1}{T}\sum_t E[(zI + B_T)^{-1}F_tF_t'] \\ &= zm(-z;c) + \frac{1}{P}\operatorname{tr}E[(zI + B_T)^{-1}F_tF_t'] \end{aligned} \tag{128}$$

where we have used symmetry across t in the last step. Using the Sherman-Morrison formula, we get

$$\frac{1}{T}\operatorname{tr}E[(zI + B_T)^{-1}F_t'F_t] = E[\frac{\frac{1}{T}F_t'(zI + B_{T,t})^{-1}F_t}{1 + \frac{1}{T}F_t'(zI + B_{T,t})^{-1}F_t}]\,,$$

where

$$B_{T,t} \;=\; \frac{1}{T}\sum_{\tau\neq t} F_\tau F_\tau'.$$

Furthermore, since all functions involved are uniformly bounded, a standard argument implies that we can replace

$$\frac{1}{T}F_t'(zI + B_{T,t})^{-1}F_t$$

with

$$\xi(z;c)$$

by (127).[36] □

G A Proof that Managed Portfolio Returns Satisfy the Assumptions of RMT

The goal of this section is to prove the following theorem.

Theorem 5 *The eigenvalue distribution of* $E[F_tF_t']$ *converges to that of* $\Psi\sigma_*$ *where* $\sigma_* = \lim \mathrm{tr}(\Sigma\Sigma_\varepsilon)$ *in the limit as* $N, P, T \to \infty,\ P/T \to c$, *so that*

$$\frac{1}{P}\,\mathrm{tr}\left((zI + E[F_tF_t'])^{-1}\right) \;\to\; \sigma_*^{-1}m_\Psi(-z/\sigma_*) \;=\; m_{\sigma_*\Psi}(-z) \;=\; \frac{1}{P}\,\mathrm{tr}((zI+\sigma_*\Psi)^{-1})\,, \quad (129)$$

[36] Indeed, $E[\frac{Y_T}{1+Y_T} - \frac{Z_T}{1+Z_T}] \;=\; \frac{Y_T - Z_T}{(1+Y_T)(1+Z_T)}$ for any random variables $Y_T,\ Z_T$. If $Y_T, Z_T \geq 0$ then $\frac{|Y_T - Z_T|}{(1+Y_T)(1+Z_T)} \leq 1$ and hence convergence $Y_T - Z_T \to 0$ in probability implies convergence of expectations.

whereas

$$\frac{1}{P}\operatorname{tr}((zI + B_T)^{-1}) \;\to\; m(-z;c)\,, \tag{130}$$

where, for each $z < 0$*, we have that* $m(z;c)$ *is the unique positive solution to the nonlinear master equation*

$$m(z;c) = \frac{1}{1 \;-\; c \;-\; c\,z\,m(z;c)}\, m_{\sigma_*\Psi}\left(\frac{z}{1 \;-\; c \;-\; c\,z\,m(z;c)}\right). \tag{131}$$

This theorem's proof is non-trivial and based on techniques from the random matrix theory from (Bai and Zhou, 2008). Applying standard results from random matrix theory to F_t is not straightforward because of the complex cross-dependence in higher moments of F_t introduced by the signals. Namely, even if R_{t+1} are conditionally independent, $S_t' R_{t+1}$ have very strong cross-dependencies.

Once the theorem is proved, we can then directly establish another useful auxiliary result from KMZ.

Lemma 11 *Define* $\xi(z;c)$ *through*

$$\frac{c^{-1}\xi(z;c)}{1+\xi(z;c)} \;=\; 1 \;-\; m(-z;c)z\,. \tag{132}$$

Then,

$$\frac{1}{T}\operatorname{tr}((zI + B_T)^{-1}\Psi) \;\to\; \xi(z;c) \tag{133}$$

almost surely and

$$\frac{1}{T}F_{T+1}'(zI + B_T)^{-1}F_{T+1} \;\to\; \xi(z;c) \tag{134}$$

in probability. Furthermore, $\xi(z;c) < c/z$.

Define the effective shrinkage

$$Z^*(z;c) \;=\; z\,(1+\xi(z;c)) \;\in (z,\; z+c) \tag{135}$$

Then, $Z^*(z;c)$ *is monotone increasing in* z *and* c. *In the ridgeless limit as* $z \to 0$, *we have*

$$Z^*(z;c) \;\to\; \begin{cases} 0, & c < 1 \\ 1/\tilde{m}(c), & c > 1 \end{cases} \tag{136}$$

where $\tilde{m}(c) > 0$ *is the unique positive solution to*

$$c - 1 \;=\; \frac{\int \frac{dH(x)}{\tilde{m}(1+\tilde{m}\,x)}}{\int \frac{x dH(x)}{1+\tilde{m}\,x}} \tag{137}$$

Lemma 12 *Let* X_P *be a sequence of positive semi-definite matrices with* $\mathrm{tr}(X_P) \leq K$. *Then,*

$$\lim_{M\to\infty} (\frac{1}{P}\,\mathrm{tr}(zI + A_P + X_P)^{-1} - \frac{1}{P}\,\mathrm{tr}(zI + A_P)^{-1}) \;=\; 0$$

for any positive semi-definite matrices A_P.

Proof. We have

$$\frac{1}{P}\,\mathrm{tr}(zI + A_P + X_P)^{-1} - \frac{1}{P}\,\mathrm{tr}(zI + A_P)^{-1} \;=\; \frac{1}{P}\,\mathrm{tr}((zI + A_P + X_P)^{-1} - (zI + A_P)^{-1})$$

and the claim follows because

$$\frac{1}{P}\,\mathrm{tr}((zI + A_P + X_P)^{-1} - (zI + A_P)^{-1}) \;=\; -\frac{1}{P}\,\mathrm{tr}((zI + A_P + X_P)^{-1} X_P (zI + A_P)^{-1})$$

and

$$\begin{aligned} &\mathrm{tr}((zI + A_P + X_P)^{-1} X_P (zI + A_P)^{-1}) \;=\; \mathrm{tr}(X_P (zI + A_P)^{-1} (zI + A_P + X_P)^{-1}) \\ &\leq\; \mathrm{tr}(X_P) \|(zI + A_P)^{-1} (zI + A_P + X_P)^{-1}\| \;\leq\; K z^{-2} \end{aligned} \tag{138}$$

Thus, the difference is bounded in absolute value by Kz^{-2}/M. □

The technical condition in Assumption 4 that $E[\|\tilde{F}_t\|^4]$ is bounded holds, for example, under the following setup.

Lemma 13 *Suppose that $\tilde{F}_{t+1} = \nu_F + \Sigma_{F,t}^{1/2} \tilde{X}_{t+1}$, where the coordinates of $\tilde{X}_{t+1}$ are independent and $E[\tilde{X}_{t+1}] = 0$, $E[\tilde{X}_{t+1} \tilde{X}_{t+1}'] = I$, and have uniformly bounded forth moments. Then, $E[\|\tilde{F}_{t+1}\|^4]$ is uniformly bounded.*

Proof of Lemma 13. Using Lemma 2, we get

$$\begin{aligned} &E[\|\tilde{\beta}\|^4] \;=\; E[(\tilde{F}_t' \Psi \tilde{F}_t)^2] \;=\; E[(\nu_F' \Psi \nu_F + \Sigma_{F,t}^{1/2} \tilde{X}_t' \Psi \tilde{X}_t)^2 \\ &=\; E[(\nu_F' \Psi \nu_F)^2 \;+\; 2\nu_F' \Psi \nu_F \,\mathrm{tr}(\Sigma_F \Psi) \;+\; (\mathrm{tr}(\Sigma_F \Psi))^2 \;+\; \sum_i (\Sigma_F^{1/2} A \Sigma_F^{1/2})_{i,i}^2 (E[\tilde{X}_i^4] - 3) \;+\; 2\|\Sigma_F^{1/2} A \Sigma_F^{1/2}\|_2^2] \\ &\leq\; K(\|\nu_F\|^2 \;+\; E[\mathrm{tr}(\Sigma_F)]) \end{aligned} \tag{139}$$

for some $K > 0$ and is therefore uniformly bounded. □

Lemma 14 (Managed Portfolios Satisfy The RMT Conditions) *Suppose that $P, T \to \infty, P/T \to c > 0$. Let A_P be a sequence of symmetric $P \times P$ matrices such that $\|A_P\| \leq K$ and A_P are independent of F_t. Then, $E[F_t F_t']$ is uniformly bounded and*

$$\mathrm{Var}[\frac{1}{T} F_t' A_P F_t] \;\to\; 0\,, \tag{140}$$

so that

$$\frac{1}{T}\left(F_t' A_P F_t \;-\; \mathrm{tr}\left(A_P\,\sigma_*\Psi\right)\right) \;\to\; 0$$

in probability. That is, averaging across P factors leads to constant risk, no matter which matrix A we use to measure it.

An important observation is that by Lemma 6,

$$\frac{1}{T}\,\mathrm{tr}(A_P E[F_t F_t']) \;\approx\; \frac{1}{T}\,\mathrm{tr}(A_P(\Psi\Sigma_F\Psi + \sigma_*\Psi))\,. \tag{141}$$

However, since Σ_F has a uniformly bounded trace norm, we have

$$\frac{1}{T}\,\mathrm{tr}(A_P(\Psi\Sigma_F\Psi + \sigma_*\Psi)) \;\approx\; \frac{1}{T}\,\mathrm{tr}(A_P(\sigma_*\Psi)) \tag{142}$$

Note that $\mathrm{tr}\left(A_P\,F_t F_t'\right) \;=\; F_t' A_P F_t$.

Proof of Lemma 14. For simplicity, we will assume that A_P is deterministic.[37] We can also assume that A_P is symmetric because $F_t' A_P F_t = F_t 0.5(A_P + A_P')F_t$. We need to prove that

$$\frac{1}{T^2}E[F_t' A_P F_t F_t' A_P F_t] \;-\; \left(\frac{1}{T}E[F_t' A_P F_t]\right)^2 \;\to\; 0$$

We have by Lemma 6 that

$$\begin{aligned} E[F_t F_t'] \;&=\; ((\mathrm{tr}\,\Sigma)^2 + \mathrm{tr}(\Sigma^2))\Psi\Sigma_F\Psi \\ &+ \mathrm{tr}(\Sigma\circ\Sigma)\Psi^{1/2}(\kappa-3)\,\mathrm{diag}(\Psi^{1/2}\Sigma_F\Psi^{1/2})\Psi^{1/2} \;+\Psi\Big(\,\mathrm{tr}(\Sigma\Sigma_\varepsilon) + \mathrm{tr}(\Psi\Sigma_F)\,\mathrm{tr}(\Sigma^2)\Big) \end{aligned} \tag{143}$$

[37] Otherwise, we replace all expectations below by expectations conditional on A_P.

and, with Σ_F having uniformly bounded traces and Assumption 2, we get

$$\begin{aligned}
\frac{1}{T}E[F_t' A_P F_t] &= \frac{1}{T}\operatorname{tr} E[A_P F_t F_t'] \\
&\approx \frac{1}{T}\operatorname{tr}\Bigg(A_P\Bigg((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \operatorname{tr}(\Sigma\circ\Sigma)\Psi^{1/2}(\kappa-3)\operatorname{diag}(\Psi^{1/2}\Sigma_F\Psi^{1/2})\Psi^{1/2} \\
&+ \Psi\Big(\operatorname{tr}(\Sigma\Sigma_\varepsilon) + \operatorname{tr}(\Psi\Sigma_F)\operatorname{tr}(\Sigma^2)\Big)\Bigg)\Bigg) \\
&\approx T^{-1}\operatorname{tr}(A_P\Psi)\sigma_*,
\end{aligned} \tag{144}$$

since

$$\frac{1}{TP}\operatorname{tr}(\Psi A_P\Psi\Sigma_F) = O(1/T).$$

Similarly, the kurtosis term does not matter because it has a uniformly bounded trace. Throughout the proof, we will use the notation

$$\beta \equiv \tilde{F}_t\,, \tag{145}$$

so that

$$F_t = S_{t-1}' R_t = S_{t-1}'(S_{t-1}\beta + \varepsilon_t). \tag{146}$$

Note that, by Assumption 4,

$$E[\|\beta\|^4] \tag{147}$$

is uniformly bounded.

Then, we have

$$\begin{aligned} F_t F_t' &= S_{t-1}'(S_{t-1}\beta\beta' S_{t-1}' + \varepsilon_t \beta' S_{t-1}' + S_{t-1}\beta\varepsilon_t' + \varepsilon_t\varepsilon_t')S_{t-1} \\ &= Z_t\beta\beta' Z_t + S_{t-1}'\varepsilon_t\beta' Z_t + Z_t\beta\varepsilon_t' S_{t-1} + S_{t-1}'\varepsilon_t\varepsilon_t' S_{t-1}\,. \end{aligned} \tag{148}$$

with $Z_t = S_{t-1}'S_{t-1}$. Then, Lemma 5 implies that

$$
\begin{aligned}
&\frac{1}{T^2}E[F_t'AF_tF_t'AF_t] \;=\; \frac{1}{T^2}\operatorname{tr}E[F_tF_t'AF_tF_t'A] \\
&=\; \frac{1}{T^2}\operatorname{tr}E[(Z_t\beta\beta'Z_t + S_{t-1}'\varepsilon_t\beta'Z_t + Z_t\beta\varepsilon_t'S_{t-1} + S_{t-1}'\varepsilon_t\varepsilon_t'S_{t-1})A \\
&(Z_t\beta\beta'Z_t + S_{t-1}'\varepsilon_t\beta'Z_t + Z_t\beta\varepsilon_t'S_{t-1} + S_{t-1}'\varepsilon_t\varepsilon_t'S_{t-1})A] \\
&=\; \frac{1}{T^2}\operatorname{tr}E[Z_t\beta\beta'Z_tAZ_t\beta\beta'Z_tA] \\
&+\; \frac{1}{T^2}2\operatorname{tr}E[Z_t\beta\beta'Z_tAS_{t-1}'\varepsilon_t\varepsilon_t'S_{t-1}A] \\
&+ \frac{1}{T^2}2\operatorname{tr}E[S_{t-1}'\varepsilon_t\beta'Z_tAS_{t-1}'\varepsilon_t\beta'Z_tA] \\
&+ \frac{1}{T^2}2\operatorname{tr}E[S_{t-1}'\varepsilon_t\beta'Z_tAZ_t\beta\varepsilon_t'S_{t-1}A] \\
&+ \frac{1}{T^2}\operatorname{tr}E[S_{t-1}'\varepsilon_t\varepsilon_t'S_{t-1}AS_{t-1}'\varepsilon_t\varepsilon_t'S_{t-1}A] \\
&=\; \frac{1}{T^2}\operatorname{tr}E[Z_t\beta\beta'Z_tAZ_t\beta\beta'Z_tA] \\
&+\; \frac{1}{T^2}2\operatorname{tr}E[Z_t\beta\beta'Z_tAZ_tA] \\
&+ \frac{1}{T^2}2\operatorname{tr}E[Z_tAZ_t\beta\beta'Z_tA] \\
&+ \frac{1}{T^2}2\operatorname{tr}E[(\beta'Z_tAZ_t\beta)Z_tA] \\
&+ \frac{1}{T^2}(2\operatorname{tr}E[Z_tAZ_tA] \;+\; E[(\operatorname{tr}(Z_tA))^2]) \\
&+ E[\sum_i (S_{t-1}AS_{t-1}')_{i,i}^2(\kappa_\varepsilon - 3)] \\
&=\; \frac{1}{T^2}\operatorname{tr}E[Z_t\beta\beta'Z_tAZ_t\beta\beta'Z_tA] \\
&+\; \frac{1}{T^2}4\operatorname{tr}E[Z_t\beta\beta'Z_tAZ_tA] \\
&+ \frac{1}{T^2}2\operatorname{tr}E[(\beta'Z_tAZ_t\beta)Z_tA] \\
&+ \frac{1}{T^2}(2\operatorname{tr}E[Z_tAZ_tA] \;+\; E[(\operatorname{tr}(Z_tA))^2]) \\
&+ \frac{1}{T^2}E[\sum_i (S_{t-1}AS_{t-1}')_{i,i}^2(\kappa_\varepsilon - 3)] \\
&=\; Term1 \;+\; Term2 \;+\; Term3 \;+\; Term4 \;+\; Term5 \;+\; Kurt,
\end{aligned}
\tag{149}
$$

where in the last term, we have used Lemma 5 to show that, by formula (67), we have

$$\begin{aligned}
&\operatorname{tr} E[S'_{t-1}\varepsilon_t\varepsilon'_t S_{t-1}AS'_{t-1}\varepsilon_t\varepsilon'_t S_{t-1}A] \\
&= \operatorname{tr} E[\varepsilon'_t S_{t-1}AS'_{t-1}\varepsilon_t\varepsilon'_t S_{t-1}AS'_{t-1}\varepsilon_t] \\
&= E[(\varepsilon'_t S_{t-1}AS'_{t-1}\varepsilon_t)^2] \\
&= E[\sum_i (S_{t-1}AS'_{t-1})^2_{i,i}(\kappa_\varepsilon - 3) \;+\; 2\operatorname{tr}((S_{t-1}AS'_{t-1})^2) \;+\; (\operatorname{tr}(S_{t-1}AS'_{t-1}))^2]
\end{aligned} \tag{150}$$

and

$$\operatorname{tr}(S_{t-1}AS'_{t-1}) \;=\; \operatorname{tr}(AS'_{t-1}S_{t-1}) \;=\; \operatorname{tr}(AZ_t) \tag{151}$$

whereas

$$\operatorname{tr}((S_{t-1}AS'_{t-1})^2) \;=\; \operatorname{tr}(S_{t-1}AS'_{t-1}S_{t-1}AS'_{t-1}) \;=\; \operatorname{tr}(AZ_tAZ_t)\,. \tag{152}$$

Furthermore, we have used the fact that all terms that are cubic in ε_t vanish because of Assumption 1 and (68). Note also that

$$Kurt \;=\; T^{-2}E[\sum_i (S_{t-1}AS'_{t-1})^2_{i,i}(\kappa_\varepsilon - 3)] \;\leq\; T^{-2}(\kappa_\varepsilon - 3)Term4. \tag{153}$$

Below, we show that $Term4$ is negligible and, hence, so it $Kurt$.

In our proofs, we will be using Newton's identities.

Lemma 15 (Newton's identities) *Let A be a matrix with eigenvalues λ_i. Then,*

$$\begin{aligned}
\sum_{i_1,i_2,i_1\neq i_2} \lambda_{i_1}\lambda_{i_2} &= (\operatorname{tr} A)^2 - \operatorname{tr}(A^2) \\
\sum_{i_1,i_2,i_3 \ \ all \ \ different} \lambda_{i_1}\lambda_{i_2}\lambda_{i_3} &= (\operatorname{tr} A)^3 - 3\operatorname{tr}(A)\operatorname{tr}(A^2) \ + \ 2\operatorname{tr}(A^3) \\
\sum_{i_1,i_2,i_3,i_4 \ \ all \ \ different} \lambda_{i_1}\lambda_{i_2}\lambda_{i_3}\lambda_{i_4} & \\
= \ (\operatorname{tr} A)^4 - 6(\operatorname{tr}(A))^2\operatorname{tr}(A^2) \ + \ 3(\operatorname{tr}(A^2))^2 + 8(\operatorname{tr} A)(\operatorname{tr}(A^3)) - \ 6\operatorname{tr}(A^4)\,. &
\end{aligned} \tag{154}$$

We also note that Assumption 2 implies

$$\operatorname{tr}(\Sigma^3) \ \leq \ \operatorname{tr}(\Sigma^2)\operatorname{tr}(\Sigma) \ = \ o((\operatorname{tr}\Sigma)^3), \ \operatorname{tr}(\Sigma^4) \leq (\operatorname{tr}(\Sigma^2))^2 \ = \ o((\operatorname{tr}\Sigma)^4) \tag{155}$$

For simplicity, we, throughout the proof below, assume that X_t is Gaussian. Without this assumption, excess kurtosis terms with the factor $(\kappa - 3)$ appear, but our calculations below imply that they are all negligible.

G.1 *Term*1 in (149)

We start with the first term. We have

$$\frac{1}{T^2}\operatorname{tr} E[Z_t\beta\beta' Z_t A Z_t\beta\beta' Z_t A] \ = \ \frac{1}{T^2}E[(\beta' Z_t A Z_t\beta)^2]\,. \tag{156}$$

Writing

$$Z_t = S'_{t-1}S_{t-1} \ = \ \Psi^{1/2}X'_{t-1}\Sigma X_{t-1}\Psi^{1/2}$$

and defining

$$\tilde{\beta} = \Psi^{1/2}\beta\,,$$

and

$$\tilde{A} \;=\; \Psi^{1/2} A \Psi^{1/2}\,,$$

and then using rotational invariance of all moments up to eight, we may assume that $\tilde{A}$ is diagonal and Σ is diagonal and $\tilde{\beta} = e_1 \|\tilde{\beta}\| = (1, 0, \cdots, 0)\|\tilde{\beta}\|$. Note that

$$\|\tilde{\beta}\|^2 \;=\; \beta' \Psi \beta \;=\; \tilde{F}_t' \Psi \tilde{F}_t\,.$$

Then,

$$\begin{aligned}
&\frac{1}{T^2} \operatorname{tr} E[Z_t \beta \beta' Z_t A Z_t \beta \beta' Z_t A] \;=\; \frac{1}{T^2} E[(\beta' Z_t A Z_t \beta)^2] \\
&\leq\; \|A\|^2 T^{-2} E[\|\beta' Z_t\|^2] \\
&=\; \|A\|^2 T^{-2} E[\tilde{\beta}' X_t' \Sigma X_t X_t' \Sigma X_t \tilde{\beta}] \\
&=\; \|A\|^2 T^{-2} E[\|\tilde{\beta}\|^4 (X_t' \Sigma X_t X_t' \Sigma X_t)_{1,1}] \\
&=\; \|A\|^2 \frac{1}{T^2} E[\|\tilde{\beta}\|^4] E[\left(\sum_{i_1, j_1, k_1} X_{i_1,1} \lambda_{i_1}(\Sigma) X_{i_1,k_1} \lambda_{k_1} X_{j_1,k_1} \lambda_{j_1}(\Sigma) X_{j_1,1} \right)^2] \\
&=\; \|A\|^2 \frac{1}{T^2} E[\|\tilde{\beta}\|^4] E[\left(\sum_{i_1, j_1, k_1} X_{i_1,1} \lambda_{i_1}(\Sigma) X_{i_1,k_1} \lambda_{k_1} X_{j_1,k_1} \lambda_{j_1}(\Sigma) X_{j_1,1} \right)^2] \\
&=\; \|A\|^2 \frac{1}{T^2} E[\|\tilde{\beta}\|^4] \\
&\times E[\sum_{i_2, j_2, k_2} \sum_{i_1, j_1, k_1} X_{i_1,1} \lambda_{i_1}(\Sigma) X_{i_1,k_1} \lambda_{k_1} X_{j_1,k_1} \lambda_{j_1}(\Sigma) X_{j_1,1} X_{i_2,1} \lambda_{i_2}(\Sigma) X_{i_2,k_2} \lambda_{k_2} X_{j_2,k_2} \lambda_{j_2}(\Sigma) X_{j_2,1}]
\end{aligned} \tag{157}$$

Here, we have used that, under Gaussianity of X_t, we have that a random, independent rotation of X_t is still Gaussian and is independent of $\|\tilde{\beta}\|$.

- First, consider the terms with $k_1 = k_2$ in (157):

$$\frac{1}{T^2} E[\|\tilde{\beta}\|^4] E[\sum_{i_2,j_2} \sum_{i_1,j_1,k_1} X_{i_1,1} \lambda_{i_1}(\Sigma) X_{i_1,k_1} \lambda_{k_1} X_{j_1,k_1} \lambda_{j_1}(\Sigma) X_{j_1,1} X_{i_2,1} \lambda_{i_2}(\Sigma) X_{i_2,k_1} \lambda_{k_1} X_{j_2,k_1} \lambda_{j_2}(\Sigma) X_{j_2,1}] \tag{158}$$

Using Newton's identities, we get that the contribution of terms with $k_1 = 1$ is given

by

$$
\begin{aligned}
& E[\|\tilde{\beta}\|^4]\frac{1}{T^2}E[\sum_{i_2,j_2}\sum_{i_1,j_1}X_{i_1,1}^2\lambda_{i_1}(\Sigma)\lambda_1^2X_{j_1,1}^2\lambda_{j_1}(\Sigma)X_{i_2,1}^2\lambda_{i_2}(\Sigma)X_{j_2,1}^2\lambda_{j_2}(\Sigma)] \\
&= E[\|\tilde{\beta}\|^4]\frac{1}{T^2}\Bigg(E[\sum_{i_2,j_2,i_1,j_1\ all\ different}X_{i_1,1}^2\lambda_{i_1}(\Sigma)X_{j_1,1}^2\lambda_{j_1}(\Sigma)X_{i_2,1}^2\lambda_{i_2}(\Sigma)X_{j_2,1}^2\lambda_{j_2}(\Sigma)] \\
&+ E[\sum_{i_2,j_2,i_1,j_1\ only\ two\ are\ equal}X_{i_1,1}^2\lambda_{i_1}(\Sigma)X_{j_1,1}^2\lambda_{j_1}(\Sigma)X_{i_2,1}^2\lambda_{i_2}(\Sigma)X_{j_2,1}^2\lambda_{j_2}(\Sigma)] \\
&+ E[\sum_{i_2,j_2,i_1,j_1\ only\ three\ are\ equal}X_{i_1,1}^2\lambda_{i_1}(\Sigma)X_{j_1,1}^2\lambda_{j_1}(\Sigma)X_{i_2,1}^2\lambda_{i_2}(\Sigma)X_{j_2,1}^2\lambda_{j_2}(\Sigma)] \\
&+ E[\sum_{i_2,j_2,i_1,j_1\ all\ four\ are\ equal}X_{i_1,1}^2\lambda_{i_1}(\Sigma)X_{j_1,1}^2\lambda_{j_1}(\Sigma)X_{i_2,1}^2\lambda_{i_2}(\Sigma)X_{j_2,1}^2\lambda_{j_2}(\Sigma)]\Bigg) \\
&= E[\|\tilde{\beta}\|^4]\frac{1}{T^2}\Bigg((\operatorname{tr}\Sigma)^4-6(\operatorname{tr}\Sigma)^2(\operatorname{tr}(\Sigma^2))\ +\ 8(\operatorname{tr}\Sigma)(\operatorname{tr}(\Sigma^3))+3(\operatorname{tr}(\Sigma^2))^2-6\operatorname{tr}(\Sigma^4) \\
&+ \binom{4}{2}E[X^4]\sum_j\lambda_j(\Sigma)^2\sum_{i_1,j_1\neq j,i_1\neq j_1}\lambda_{i_1}(\Sigma)\lambda_{j_1}(\Sigma) \\
&+ 4E[X^6]\sum_j\lambda_j(\Sigma)^3\sum_{i_1\neq j}\lambda_{i_1}(\Sigma) \\
&+ E[X^8]\operatorname{tr}(\Sigma^4)\Bigg) \\
&= E[\|\tilde{\beta}\|^4]\frac{1}{T^2}\Bigg((\operatorname{tr}\Sigma)^4-6(\operatorname{tr}\Sigma)^2(\operatorname{tr}(\Sigma^2))\ +\ 8(\operatorname{tr}\Sigma)(\operatorname{tr}(\Sigma^3))+3(\operatorname{tr}(\Sigma^2))^2-6\operatorname{tr}(\Sigma^4) \\
&+ \binom{4}{2}E[X^4]\sum_j\lambda_j(\Sigma)^2((\operatorname{tr}(\Sigma)-\lambda_j)^2-(\operatorname{tr}(\Sigma^2)-\lambda_j^2) \\
&+ 4E[X^6](\operatorname{tr}(\Sigma)\operatorname{tr}(\Sigma^3)-\operatorname{tr}(\Sigma^4))+E[X^8]\operatorname{tr}(\Sigma^4)\Bigg) \\
&= O(1/T^2)
\end{aligned}
\tag{159}
$$

because, by the assumed normalization, $\operatorname{tr}(\Sigma)=1$ and $\sum_i\lambda_i(\Sigma)^2\ =\operatorname{tr}(\Sigma^2)\to 0$, and $E[\|\tilde{\beta}\|^4]$ is bounded by Assumption 4.

The remaining terms with $k_1 = k_2 \neq 1$ must have i_1, i_2, j_1, j_2 have at least two identical pairs. The first contribution would be

$$\begin{aligned} & T^{-2} E[\|\tilde{\beta}\|^4] E[\sum_{i_1=i_2 \neq j_1=j_2; k_1} X_{i_1,1}^2 \lambda_{i_1}^2(\Sigma) X_{i_1,k_1}^2 X_{j_1,k_1}^2 \lambda_{j_1}^2(\Sigma) X_{j_1,1}^2] \\ & \leq E[\|\tilde{\beta}\|^4] P(\mathrm{tr}(\Sigma^2))^2 \,, \end{aligned} \tag{160}$$

there will be *three* contributions like this, corresponding to the three cases: $i_1 = i_2$, $i_1 = j_1$, and $i_1 = j_2$.

In the case when more than two out of i_1, i_2, j_1, j_2 are identical, they would all have to be identical. This contribution would be negligible because it would give

$$T^{-2} E[\|\tilde{\beta}\|^4] E[X^4] P\left(\mathrm{tr}(\Sigma^4)\right) \;=\; T^{-2} o(P),$$

which is negligible.

- We can now focus on the case $k_1 \neq k_2$ in (157). First, consider the terms with $k_1 = 1$. By symmetry, terms with $k_2 = 1$ give the same contribution. Since $k_2 \neq 1$, Newton's

identities imply that

$$
\begin{aligned}
&\frac{1}{T^2}E[\|\tilde{\beta}\|^4]E[\sum_{i_2,j_2,k_2\neq 1}\sum_{i_1,j_1}X^2_{i_1,1}\lambda_{i_1}(\Sigma)X^2_{j_1,1}\lambda_{j_1}(\Sigma)X_{i_2,1}\lambda_{i_2}(\Sigma)X_{i_2,k_2}\lambda_{k_2}X_{j_2,k_2}\lambda_{j_2}(\Sigma)X_{j_2,1}] \\
&\sim \frac{1}{T^2}E[\|\tilde{\beta}\|^4]E[\sum_{i_2,k_2}\sum_{i_1,j_1}X^2_{i_1,1}X^2_{j_1,1}\lambda_{i_1}(\Sigma)\lambda_{j_1}(\Sigma)X^2_{i_2,1}\lambda_{i_2}(\Sigma)^2X^2_{i_2,k_2}] \\
&\sim E[\|\tilde{\beta}\|^4]\,\frac{1}{T^2}P\left(E[\sum_{i_2}\sum_{i_1,j_1}X^2_{i_1,1}X^2_{j_1,1}\lambda_{i_1}(\Sigma)\lambda_{j_1}(\Sigma)X^2_{i_2,1}\lambda_{i_2}(\Sigma)^2]\right) \\
&= E[\|\tilde{\beta}\|^4]\,\frac{1}{T^2}P\Bigg(\sum_{i_2,i_1,j_1\ \ all\ \ different}\lambda_{i_1}(\Sigma)\lambda_{j_1}(\Sigma)\lambda_{i_2}(\Sigma)^2 \\
&+ \sum_{i_1=j_1\neq i_2}E[X^4]\lambda_{i_1}(\Sigma)^2\lambda_{i_2}(\Sigma)^2 \\
&+ 2\sum_{i_1\neq j_1=i_2}E[X^4]\lambda_{i_1}(\Sigma)\lambda_{i_2}(\Sigma)^3 \\
&+ E[X^6]\,\mathrm{tr}(\Sigma^4)\Bigg) \\
&= E[\|\tilde{\beta}\|^4]\,\frac{1}{T^2}P\Bigg(\sum_{i_2}\lambda_{i_2}(\Sigma)^2((\mathrm{tr}(\Sigma)-\lambda_{i_2})^2-(\mathrm{tr}(\Sigma^2)-\lambda^2_{i_2})) \\
&+ E[X^4]((\mathrm{tr}(\Sigma^2))^2-\mathrm{tr}(\Sigma^4)) \\
&+ 2E[X^4]\sum_{i_2}\lambda_{i_2}(\Sigma)^3(\mathrm{tr}(\Sigma)-\lambda_{i_2}) \\
&+ E[X^6]\,\mathrm{tr}(\Sigma^4)\Bigg) \\
&= E[\|\tilde{\beta}\|^4]\,\frac{1}{T^2}P\Bigg((\mathrm{tr}(\Sigma)^2)\,\mathrm{tr}(\Sigma^2)-2(\mathrm{tr}\,\Sigma)(\mathrm{tr}(\Sigma^3))+2\,\mathrm{tr}(\Sigma^4)-(\mathrm{tr}(\Sigma^2))^2 \\
&+ E[X^4]((\mathrm{tr}(\Sigma^2))^2-\mathrm{tr}(\Sigma^4)) \\
&+ 2E[X^4]((\mathrm{tr}\,\Sigma)(\mathrm{tr}(\Sigma^3))-\mathrm{tr}(\Sigma^4))+\ E[X^6]\,\mathrm{tr}(\Sigma^4)\Bigg)
\end{aligned}
\tag{161}
$$

because the rest terms are zero. And this term gets multiplied by 2 when we add the

contribution of the $k_2 = 1$ case. As above, all these terms are

$$O(P/T^2)$$

and hence are negligible.

- Now, in the case when $k_1 \neq k_2$ and both are different from 1 in (157), we immediately get that (i_1, i_2, j_1, j_2) must either be all identical, or come in two identical pairs. The first case gives a contribution of

$$E[\|\tilde{\beta}\|^4]E[\sum_{i,k_1 \notin \{k_2,1\}} X_{i,1}^4 X_{i,k_1}^2 X_{i,k_2}^2 \lambda_i(\Sigma)^4] \sim E[\|\tilde{\beta}\|^4]E[X^4]\,(P^2-P)\,\mathrm{tr}(\Sigma^4) = o(P^2)\,.$$

The second one ought to have $i_1 = j_1, i_2 = j_2$ because $k_1 \neq k_2$ and both are not equal to 1, giving

$$\begin{aligned}
&\frac{1}{T^2}E[\|\tilde{\beta}\|^4]E[\sum_{i_2,k_2}\sum_{i_1,k_1} X_{i_1,1}^2 X_{i_1,k_1}^2 \lambda_{i_1}^2(\Sigma)\lambda_{i_2}^2(\Sigma) X_{i_2,1}^2 X_{i_2,k_2}^2] \\
&\leq \frac{1}{T^2}E[\|\tilde{\beta}\|^4]P^2\left(E[\sum_{i_2}\sum_{i_1} X_{i_1,1}^2 \lambda_{i_1}^2(\Sigma)\lambda_{i_2}^2(\Sigma) X_{i_2,1}^2]\right) \\
&= \frac{1}{T^2}E[\|\tilde{\beta}\|^4]P^2((\mathrm{tr}(\Sigma^2))^2 - \mathrm{tr}(\Sigma^4)) \\
&\leq \frac{1}{T^2}E[\|\tilde{\beta}\|^4]P^2(\mathrm{tr}(\Sigma^2))^2 = o(P^2)\frac{1}{T^2}
\end{aligned} \tag{162}$$

because $\mathrm{tr}(\Sigma^2) = o(1)$.

Summarizing, $Term1$ is negligible.

G.2 $Term2$ in (149)

We now proceed with the second term (note that *it comes with a factor of four*). As above, we, for simplicity, work under the assumption that X_t are Gaussian so that we could rotate

them and assume that $\tilde{\beta}$ is proportional to e_1. Then,

$$E[\beta' Z_t A Z_t A Z_t \beta] \;\leq\; \|A\| E[\|\tilde{\beta}\|^2] E[\sum X_{i_1,1} \lambda_{i_1}(\Sigma) X_{i_1,k_1} X_{i_2,k_1} \lambda_{i_2}(\Sigma) X_{i_2,k_2} X_{i_3,k_2} \lambda_{i_3}(\Sigma) X_{i_3,1}] \,. \tag{163}$$

- Suppose first that $k_1 = k_2 \neq 1$ in (163). The respective contribution is

$$E[\sum X_{i_1,1} \lambda_{i_1}(\Sigma) X_{i_1,k_1} X_{i_2,k_1}^2 \lambda_{i_2}(\Sigma) X_{i_3,k_1} \lambda_{i_3}(\Sigma) X_{i_3,1}] \,, \tag{164}$$

  and hence $i_1 = i_3$ for non-zero terms, so that this contribution becomes

$$\begin{aligned} & E[\sum X_{i_1,1}^2 \lambda_{i_1}(\Sigma)^2 X_{i_1,k_1}^2 X_{i_2,k_1}^2 \lambda_{i_2}(\Sigma)] \\ &= \left(\sum_{i_1 \neq i_2, k_1 \neq 1} \lambda_{i_1}(\Sigma)^2 \lambda_{i_2}(\Sigma) \;+\; E[X^4] \sum_{i_1, k_1 \neq 1} \lambda_{i_1}(\Sigma)^3 \right) \\ &\leq P((E[X^4] - 1) \operatorname{tr}(\Sigma^3) + \operatorname{tr}(\Sigma) \operatorname{tr}(\Sigma^2)) \;=\; O(P) \end{aligned} \tag{165}$$

- The terms with $k_1 = k_2 = 1$ in (163) give

$$\begin{aligned} & E[\sum X_{i_1,1}^2 \lambda_{i_1}(\Sigma) X_{i_2,1}^2 \lambda_{i_2}(\Sigma) X_{i_3,1}^2 \lambda_{i_3}(\Sigma)] \\ &\sim \left(\sum_{i_1,i_2,i_3 \;\; pairwise \;\; different} \lambda_{i_1}(\Sigma) \lambda_{i_2}(\Sigma) \lambda_{i_3}(\Sigma) \right. \\ &\left. + \; 3 \sum_{i_1,i_2 \;\; different} E[X^4] \lambda_{i_1}^2(\Sigma) \lambda_{i_2}(\Sigma) \;+\; E[X^6] \operatorname{tr}(\Sigma^3) \right) \\ &\leq \left((\operatorname{tr} \Sigma)^3 - 3 (\operatorname{tr} \Sigma) \operatorname{tr}(\Sigma^2) \;+\; 2 \operatorname{tr}(\Sigma^3) \right. \\ &\left. + \; 3 E[X^4] ((\operatorname{tr} \Sigma) \operatorname{tr}(\Sigma^2) - \operatorname{tr}(\Sigma^3)) + E[X^6] \operatorname{tr}(\Sigma^3) \right) \;=\; O(1) \end{aligned} \tag{166}$$

by Newton's identities, where $3\sum_{i_1,i_2\ different}$ appears because there are three possibilities for a coincidence of pair among i_1, i_2, i_3.

- For the terms with $k_1 \neq k_2$ and none of them equal to 1 in in (163), we must have $i_1 = i_2 = i_3$ for them to be non-zero, giving

$$\begin{aligned} & E[\sum X_{i_1,1}^2 \lambda_{i_1}(\Sigma)^3 X_{i_1,k_1}^2 X_{i_1,k_2}^2] \ \sim \ \|\tilde{\beta}\|^2((P)^2 - P)\,\mathrm{tr}(\Sigma^3) \\ & = \ o(P^2) \end{aligned} \tag{167}$$

since $((P)^2 - P) = O(P^2)$.

- If $k_1 \neq k_2 = 1$ in (163), then we get the contribution

$$\begin{aligned} & E[\sum X_{i_1,1}\lambda_{i_1}(\Sigma)X_{i_1,k_1}X_{i_2,k_1}\lambda_{i_2}(\Sigma)X_{i_2,1}\lambda_{i_3}(\Sigma)X_{i_3,1}^2] \\ & = E[\sum X_{i_1,1}\lambda_{i_1}(\Sigma)X_{i_1,k_1}X_{i_2,k_1}\lambda_{i_2}(\Sigma)X_{i_2,1}\lambda_{i_3}(\Sigma)X_{i_3,1}^2] \\ & = \ \{only\ terms\ with\ i_1 = i_2\ survive\} \\ & = \ E[\sum X_{i_1,1}^2\lambda_{i_1}^2(\Sigma)X_{i_1,k_1}^2\lambda_{i_3}(\Sigma)X_{i_3,1}^2] \\ & \sim \ \left(\mathrm{tr}(\Sigma)(\mathrm{tr}(\Sigma^2)) \ + \ (E[X^4] - 1)\,\mathrm{tr}(\Sigma^3)\right) \ = \ O(P(\mathrm{tr}(\Sigma))^3) \ = \ O(P) \end{aligned} \tag{168}$$

and there is an identical contribution with $k_1 = 1 \neq k_2$.

Thus,

$$Term2 \ \sim T^{-2}o(T^2) \tag{169}$$

is negligible.

G.3 *Term*3 in (149)

We now proceed with the third term. As above, we perform calculations under the Gaussianity assumption and rotate signals so that $\tilde{\beta}$ is proportional to e_1 and $\beta' Z_t A Z_t \beta = (Z_t A Z_t)_{1,1} = \|\tilde{\beta}\|^2 (X'_{t-1}\Sigma X_{t-1}\tilde{A}X'_{t-1}\Sigma X_{t-1})_{1,1}$. We have

$$
\begin{aligned}
& 2\frac{1}{T^2}E[\mathrm{tr}(AZ_t)\,\beta' Z_t A Z_t \beta] \\
&= \; 2E[\|\tilde{\beta}\|^2]\frac{1}{T^2}E[\sum_k \lambda_k(\tilde{A})\sum_i \lambda_i(\Sigma)X_{i,k}^2 \sum_{i_1,k_1,i_2} X_{i_1,1}\lambda_{i_1}(\Sigma)X_{i_1,k_1}\lambda_{k_1}(\tilde{A})X_{i_2,k_1}\lambda_{i_2}(\Sigma)X_{i_2,1}]
\end{aligned}
\tag{170}
$$

- First consider the terms with $k_1 = 1$ in (170). This gives

$$
\begin{aligned}
& 2E[\|\tilde{\beta}\|^2]\frac{1}{T^2}E[\sum_k \lambda_k(\tilde{A})\sum_i \lambda_i(\Sigma)X_{i,k}^2 \sum_{i_1,i_2} X_{i_1,1}^2\lambda_{i_1}(\Sigma)\lambda_1(\tilde{A})\lambda_{i_2}(\Sigma)X_{i_2,1}^2] \\
&\le 2P\|A\|^2 E[\|\tilde{\beta}\|^2]\frac{1}{T^2}(\mathrm{tr}\,\Sigma)\,E[\sum_{i_1,i_2} X_{i_1,1}^2\lambda_{i_1}(\Sigma)\lambda_{i_2}(\Sigma)X_{i_2,1}^2] \;+\; T^{-2}O(1) \\
&= \; 2P\|A\|^2 E[\|\tilde{\beta}\|^2]\frac{1}{T^2}(\mathrm{tr}\,\Sigma)\,((\mathrm{tr}(\Sigma))^2 + (E[X^4]-1)\,\mathrm{tr}(\Sigma^2)) \;+\; T^{-2}O(1) \\
&= \; O(PT^{-2}),
\end{aligned}
\tag{171}
$$

  where the $O(1)$ term comes from the contribution of $k = 1$ terms:

$$
2E[\|\tilde{\beta}\|^2]\frac{1}{T^2}E[\lambda_1(\tilde{A})\sum_i \lambda_i(\Sigma)X_{i,1}^2 \sum_{i_1,i_2} X_{i_1,1}^2\lambda_{i_1}(\Sigma)\lambda_1(\tilde{A})\lambda_{i_2}(\Sigma)X_{i_2,1}^2] \;=\; T^{-2}O(1)\,.
\tag{172}
$$

  because $\mathrm{tr}(\Sigma) = 1$.

- If $k_1 \neq 1$ in in (170), the only non-zero terms are with $i_1 = i_2$ and they give

$$
\begin{aligned}
& 2E[\|\tilde{\beta}\|^2]\frac{1}{T^2}E[\sum_k \lambda_k(\tilde{A})\sum_i \lambda_i(\Sigma)X_{i,k}^2 \sum_{i_1,k_1\neq 1} X_{i_1,1}^2\lambda_{i_1}^2(\Sigma)X_{i_1,k_1}^2\lambda_{k_1}(\tilde{A})] \\
&= \; 2E[\|\tilde{\beta}\|^2]\frac{1}{T^2}E[\sum_{k\neq 1} \lambda_k(\tilde{A})\sum_i \lambda_i(\Sigma)X_{i,k}^2 \sum_{i_1,k_1\neq 1} X_{i_1,1}^2\lambda_{i_1}^2(\Sigma)X_{i_1,k_1}^2\lambda_{k_1}(\tilde{A})] + T^{-2}O(P) \\
&= \; 2E[\|\tilde{\beta}\|^2]\frac{1}{T^2}\,E[\sum_{k\neq 1} \lambda_k(\tilde{A})\sum_i \lambda_i(\Sigma)X_{i,k}^2 \sum_{i_1,k_1\neq 1} \lambda_{i_1}^2(\Sigma)X_{i_1,k_1}^2\lambda_{k_1}(\tilde{A})] + T^{-2}O(P) \\
&= \; 2E[\|\tilde{\beta}\|^2]\frac{1}{T^2}\left(E[\sum_{k\neq 1} \lambda_k^2(\tilde{A})\sum_i \lambda_i(\Sigma)X_{i,k}^2 \sum_{i_1} \lambda_{i_1}^2(\Sigma)X_{i_1,k}^2]\right. \\
&+ \; \left. E[\sum_{k\neq 1} \lambda_k(\tilde{A}) \sum_{i,k_1\neq 1,k} \lambda_i(\Sigma)X_{i,k}^2 \sum_{i_1} \lambda_{i_1}^2(\Sigma)X_{i_1,k_1}^2\lambda_{k_1}(\tilde{A})]\right) + T^{-2}O(P) \\
&\leq \; \|A\|^2 2E[\|\tilde{\beta}\|^2]\frac{1}{T^2}\left(E[X^4]P\,\mathrm{tr}(\Sigma^3) + \sum_{k\neq 1}\sum_i \lambda_i(\Sigma)\sum_{i_1\neq i}\lambda_{i_1}^2(\Sigma)\right. \\
&+ \; \left. \sum_{k\neq 1}\sum_{i,k_1\neq 1,k} \lambda_i(\Sigma)\sum_{i_1}\lambda_{i_1}^2(\Sigma)\right) + T^{-2}O(P) \\
&\leq 2\|A\|^2 E[\|\tilde{\beta}\|^2]\frac{1}{T^2}\left(P\Big((E[X^4]-1)\,\mathrm{tr}(\Sigma^3) + \mathrm{tr}(\Sigma)\,\mathrm{tr}(\Sigma^2)\Big)\right. \\
&+ \; \left. (P^2-P)\,\mathrm{tr}(\Sigma)\,\mathrm{tr}(\Sigma^2)\right) + T^{-2}O(P) \;\; \leq 2\|A\|^2 E[\|\tilde{\beta}\|^2]\frac{1}{T^2}P^2\,\mathrm{tr}(\Sigma)\,\mathrm{tr}(\Sigma^2) \;\; + T^{-2}O(P)\,.
\end{aligned}
\tag{173}
$$

Thus,

$$
Term3 \;=\; o(P^2/T^2) \tag{174}
$$

and, hence, is negligible.

G.4 $Term4$ and $Term5$ in (149)

As above, we are only considering the Gaussian case, for simplicity. We have

$$\begin{aligned} & E[2\,\mathrm{tr}(AZ_tAZ_t) \;+\; (\mathrm{tr}(AZ_t))^2] \\ &= \; 2E[\sum \lambda_k(\tilde{A})X_{i,k}\lambda_i(\Sigma)X_{i,k_1}\lambda_{k_1}(\tilde{A})X_{i_1,k_1}\lambda_{i_1}(\Sigma)X_{i_1,k}] \\ &+ E[(\sum_k \lambda_k(\tilde{A})\sum_i \lambda_i(\Sigma)X_{i,k}^2)^2] \end{aligned} \tag{175}$$

We have

$$\begin{aligned} & E[(\sum_k \lambda_k(\tilde{A})\sum_i \lambda_i(\Sigma)X_{i,k}^2)^2] \\ &= \; E[\sum_{k,k_1,i,i_1} \lambda_k(\tilde{A})\lambda_{k_1}(\tilde{A})\lambda_{i_1}(\Sigma)X_{i_1,k_1}^2\lambda_{i_2}(\Sigma)X_{i_2,k_2}^2] \\ &= \; E[\sum_k \lambda_k^2(\tilde{A})\sum_{i_1,i_2}\lambda_{i_1}\lambda_{i_2}X_{i_1,k}^2X_{i_2,k}^2] \;+\; \sum_{k_1\neq k_2}\lambda_{k_1}(\tilde{A})\lambda_{k_2}(\tilde{A})(\mathrm{tr}(\Sigma))^2 \\ &= \; E[X^4]\sum_k \lambda_k^2(\tilde{A})\sum_{i_1=i_2}\lambda_{i_1}\lambda_{i_2} + \sum_k \lambda_k^2(\tilde{A})\sum_{i_1\neq i_2}\lambda_{i_1}\lambda_{i_2} + \sum_{k_1\neq k_2}\lambda_{k_1}(\tilde{A})\lambda_{k_2}(\tilde{A})(\mathrm{tr}(\Sigma))^2 \\ &= \; E[X^4]\,\mathrm{tr}(\tilde{A}^2)\,\mathrm{tr}(\Sigma^2) \;+\; \mathrm{tr}(\tilde{A}^2)(\mathrm{tr}(\Sigma)^2 - \mathrm{tr}(\Sigma^2)) + (\mathrm{tr}(\tilde{A})^2 - \mathrm{tr}(\tilde{A}^2))\,\mathrm{tr}(\Sigma)^2 \end{aligned} \tag{176}$$

Thus,

$$Term5 \;=\; T^{-2}E[(\mathrm{tr}(AZ_t))^2] \;=\; T^{-2}\,\mathrm{tr}(\tilde{A})^2 \;+\; O(P/T^2)\,. \tag{177}$$

Similarly,

$$
\begin{aligned}
E[2\,\mathrm{tr}(AZ_tAZ_t)] \;&=\; 2E[\sum \lambda_k(\tilde{A})X_{i,k}\lambda_i(\Sigma)X_{i,k_1}\lambda_{k_1}(\tilde{A})X_{i_1,k_1}\lambda_{i_1}(\Sigma)X_{i_1,k}] \\
&=\; 2E[\sum_{k_1=k} \lambda_k(\tilde{A})^2X_{i,k}^2\lambda_i(\Sigma)\lambda_{i_1}(\Sigma)X_{i_1,k}^2] \\
&+\; 2E[\sum_{k\neq k_1}\sum_i \lambda_k(\tilde{A})X_{i,k}^2\lambda_i^2(\Sigma)X_{i,k_1}^2\lambda_{k_1}(\tilde{A})] \\
&\leq \|A\|^2(2P((E[X^4]-1)\,\mathrm{tr}(\Sigma^2)+(\mathrm{tr}\,\Sigma)^2) \;+\; 2((P)^2-P)\,\mathrm{tr}(\Sigma^2)) \;=\; o(T^2)\,.
\end{aligned}
\tag{178}
$$

Thus, the only term that is non-negligible is $Term5$ in (177). The proof of Lemma 14 is complete. □

Proof of Theorem 5. The first claim follows because, by Lemma 12, the other contributions do not impact eigenvalue distribution.

To prove the claim about the eigenvalue distribution of B_T, we use a Theorem of (Bai and Zhou, 2008). According to (Bai and Zhou, 2008), defining $F_{t+1} = S_t'R_{t+1}$, we need to verify the following technical conditions:

(1) $E[F_{t+1}F_{t+1}'] = A_P$ for some matrix A_P

(2) $E[(F_{t+1}'BF_{t+1} - \mathrm{tr}(A_PB_P))^2] \;=\; o(T^2)$ for any bounded matrix sequence $B_P,\ P>0$.

(3) The norm of A_P is uniformly bounded, and its eigenvalue distribution converges as $P\to\infty$.

The only non-trivial claim here is item (2), which in turn follows from Lemma 14. The proof of Theorem 5 is complete. □

H Technical Lemmas for Computing Higher Moments

The following lemma is a direct consequence of (149) and the polarization identity

$$ab \;=\; 0.25((a+b)^2-(a-b)^2)\,.$$

Lemma 16 *Let $Z_t = S_{t-1}'S_{t-1}$. Recall also that*

$$R_{t+1} \;=\; S_t\beta \;+\; \varepsilon_{t+1}, \tag{179}$$

where, for brevity, we omit the time index for $\beta \;=\; \tilde{F}_{t+1}$. Thus,

$$F_t \;=\; Z_t\beta \;+\; S_{t-1}'\varepsilon_t\,. \tag{180}$$

For any two matrices A, B with A being symmetric, we have

$$\begin{aligned}
&\frac{1}{T}E[F_t'AF_tF_t'BF_t] \\
&= \;\frac{1}{T}\operatorname{tr}E[Z_t\beta\beta'Z_tAZ_t\beta\beta'Z_tB] \\
&+ \;\frac{1}{T}2\operatorname{tr}(E[\beta'Z_tAZ_tBZ_t\beta]+E[\beta'Z_tBZ_tAZ_t\beta]) \\
&+ \frac{1}{T}\operatorname{tr}(E[(\beta'Z_tAZ_t\beta)Z_tB]+E[(\beta'Z_tBZ_t\beta)Z_tA]) \\
&+ \frac{1}{T}((\kappa_\varepsilon-1)\operatorname{tr}E[Z_tAZ_tB]+\;E[\operatorname{tr}(Z_tA)\operatorname{tr}(Z_tB)]) \\
&= \;Term1 \;+\; Term2 \;+\; Term3 \;+\; Term4 \;+\; Term5\,.
\end{aligned} \tag{181}$$

Proof. When A, B are symmetric, (149) implies

$$\begin{aligned}
&\frac{1}{T}E[F_t'AF_tF_t'BF_t] \\
&= \frac{1}{T}\operatorname{tr} E[Z_t\beta\beta' Z_tAZ_t\beta\beta' Z_tB] \\
&+ \frac{1}{T}2\operatorname{tr}(E[Z_t\beta\beta' Z_tAZ_tB] + E[Z_t\beta\beta' Z_tBZ_tA]) \\
&+ \frac{1}{T}\operatorname{tr}(E[(\beta' Z_tAZ_t\beta)Z_tB] + E[(\beta' Z_tBZ_t\beta)Z_tA]) \\
&+ \frac{1}{T}((\kappa_\varepsilon - 1)\operatorname{tr} E[Z_tAZ_tB] + E[\operatorname{tr}(Z_tA)\operatorname{tr}(Z_tB)])
\end{aligned} \tag{182}$$

The general case follows because

$$\begin{aligned}
&\frac{1}{T}E[F_t'AF_tF_t'BF_t] = \frac{1}{T}E[F_t'0.5(A+A')F_tF_t'0.5(B+B')F_t] \\
&= \frac{1}{T}\operatorname{tr} E[Z_t\beta\beta' Z_t0.5(A+A')Z_t\beta\beta' Z_t0.5(B+B')] \\
&+ \frac{1}{T}2\operatorname{tr}(E[Z_t\beta\beta' Z_t0.5(A+A')Z_t0.5(B+B')] + E[Z_t\beta\beta' Z_t0.5(B+B')Z_t0.5(A+A')]) \\
&+ \frac{1}{T}\operatorname{tr}(E[(\beta' Z_t0.5(A+A')Z_t\beta)Z_t0.5(B+B')] + E[(\beta' Z_t0.5(B+B')Z_t\beta)Z_t0.5(A+A')]) \\
&+ \frac{1}{T}((\kappa_\varepsilon - 1)\operatorname{tr} E[Z_t0.5(A+A')Z_t0.5(B+B')] + E[\operatorname{tr}(Z_t0.5(A+A'))\operatorname{tr}(Z_t0.5(B+B'))]) \\
&= \frac{1}{T}\operatorname{tr} E[Z_t\beta\beta' Z_tAZ_t\beta\beta' Z_tB] \\
&+ \frac{1}{T}\operatorname{tr}(E[\beta' Z_tAZ_tBZ_t\beta] + E[\beta' Z_tBZ_tAZ_t\beta] + E[\beta' Z_tA'Z_tBZ_t\beta] + E[\beta' Z_tAZ_tB'Z_t\beta]) \\
&+ \frac{1}{T}\operatorname{tr}(E[(\beta' Z_tAZ_t\beta)Z_tB] + E[(\beta' Z_tBZ_t\beta)Z_tA]) \\
&+ \frac{1}{T}((\kappa_\varepsilon - 1)0.5\operatorname{tr}(E[Z_tAZ_tB] + E[Z_tA'Z_tB]) + E[\operatorname{tr}(Z_tA)\operatorname{tr}(Z_tB)])
\end{aligned} \tag{183}$$

□

Lemma 17 *For any two matrices A, B, we have*

$$
\begin{aligned}
&\frac{1}{T}\operatorname{tr} E[Z_t\beta\beta' Z_t A Z_t\beta\beta' Z_t B] \\
&\sim \Big((\tilde\beta'\tilde A\tilde\beta)\operatorname{tr}(\tilde B) + (\tilde\beta'\tilde B\tilde\beta)P\Big)\|\tilde\beta\|^2\ \operatorname{tr}(\Sigma^2)(\operatorname{tr}(\Sigma))^2\frac{1}{T} \\
&+ E[\|\tilde\beta\|^4]((\operatorname{tr}\tilde A)(\operatorname{tr}\tilde B) + 2\operatorname{tr}(\tilde A\tilde B))(\operatorname{tr}(\Sigma^2))^2\frac{1}{T} \\
&+ E[\|\tilde\beta\|^4]E[X^4]\,P\operatorname{tr}(\tilde B)\operatorname{tr}(\Sigma^4)\frac{1}{T} \\
&\frac{1}{T}2\operatorname{tr}(E[Z_t\beta\beta' Z_t A Z_t B] + E[Z_t\beta\beta' Z_t B Z_t A]) \\
&\sim \frac{1}{T}4\|\tilde\beta\|^2\operatorname{tr}(\tilde A\tilde B)\operatorname{tr}(\Sigma)\operatorname{tr}(\Sigma^2) \\
&+ \frac{1}{T}4\|\tilde\beta\|^2(P\operatorname{tr}(\tilde B) - \operatorname{tr}(\tilde A\tilde B))\operatorname{tr}(\Sigma^3) \\
&+ \frac{1}{T}4\Big(\tilde\beta'\tilde A\tilde\beta\,(\operatorname{tr}\tilde B) + \tilde\beta'\tilde B\tilde\beta\,(\operatorname{tr}\tilde A)\Big)\operatorname{tr}(\Sigma)(\operatorname{tr}(\Sigma^2)) \\
&\frac{1}{T}\operatorname{tr}(E[(\beta' Z_t A Z_t\beta)Z_t B] + E[(\beta' Z_t B Z_t\beta)Z_t A]) \\
&\sim \frac{1}{T}\Big(\tilde\beta'\tilde A\tilde\beta\,(\operatorname{tr}\tilde B) + \tilde\beta'\tilde B\tilde\beta\,(\operatorname{tr}\tilde A)\Big)(\operatorname{tr}\Sigma)^3 + 2E[\|\tilde\beta\|^2]\frac{1}{T}(\operatorname{tr}\tilde A)(\operatorname{tr}\tilde B)\operatorname{tr}(\Sigma)\operatorname{tr}(\Sigma^2) \\
&\frac{1}{T}((\kappa_\varepsilon - 1)\operatorname{tr} E[Z_t A Z_t B] + E[\operatorname{tr}(Z_t A)\operatorname{tr}(Z_t B)]) \\
&\sim \Big((\operatorname{tr}\tilde A)(\operatorname{tr}\tilde B) + 2\operatorname{tr}(\tilde A\tilde B)\Big)(\operatorname{tr}\Sigma)^2\frac{1}{T}
\end{aligned}
\tag{184}
$$

with $\tilde A = \Psi^{1/2}A\Psi^{1/2}$ *and* $\tilde B = \Psi^{1/2}B\Psi^{1/2}$.

Proof of Lemma 17. Using (**??**), (169), (174), and (**??**) , we get the following result:

$$
\begin{aligned}
&\frac{1}{T}\operatorname{tr} E[Z_t\beta\beta' Z_t A Z_t\beta\beta' Z_t B] \;\sim\; 3E[\|\tilde\beta\|^4]\operatorname{tr}(\tilde A\tilde B)\,(\operatorname{tr}(\Sigma^2))^2\frac{1}{T} \;+\; E[\|\tilde\beta\|^4]E[X^4]\operatorname{tr}(\tilde A\tilde B)\,(\operatorname{tr}(\Sigma^4))\frac{1}{T}\\
&+\;\Big((\tilde\beta'\tilde A\tilde\beta)\operatorname{tr}(\tilde B)\;+\;(\tilde\beta'\tilde B\tilde\beta)P\Big)\,\|\tilde\beta\|^2\Bigg(\operatorname{tr}(\Sigma^2)(\operatorname{tr}(\Sigma))^2-2(\operatorname{tr}\Sigma)(\operatorname{tr}(\Sigma^3))+2\operatorname{tr}(\Sigma^4)-(\operatorname{tr}(\Sigma^2))^2\\
&+\;E[X^4]((\operatorname{tr}(\Sigma^2))^2-\operatorname{tr}(\Sigma^4))\\
&+\;2E[X^4]((\operatorname{tr}\Sigma)(\operatorname{tr}(\Sigma^3))-\operatorname{tr}(\Sigma^4))+\;E[X^6]\operatorname{tr}(\Sigma^4)\Bigg)\frac{1}{T}\\
&+\;E[\|\tilde\beta\|^4]E[X^4]\,(P\operatorname{tr}(\tilde B)-\operatorname{tr}(\tilde A\tilde B))\;\operatorname{tr}(\Sigma^4)\frac{1}{T}\\
&+\;E[\|\tilde\beta\|^4]((\operatorname{tr}\tilde A)\operatorname{tr}(\tilde B)-\operatorname{tr}(\tilde A\tilde B))(\operatorname{tr}(\Sigma^2))^2\frac{1}{T}\\
&\frac{1}{T}2\operatorname{tr}(E[Z_t\beta\beta' Z_t A Z_t B]+E[Z_t\beta\beta' Z_t B Z_t A])\\
&\sim\;\frac{1}{T}4\|\tilde\beta\|^2\operatorname{tr}(\tilde A\tilde B)((E[X^4]-1)\operatorname{tr}(\Sigma^3)+\operatorname{tr}(\Sigma)\operatorname{tr}(\Sigma^2))\\
&+\frac{1}{T}4\|\tilde\beta\|^2(P\operatorname{tr}(\tilde B)-\operatorname{tr}(\tilde A\tilde B))\operatorname{tr}(\Sigma^3)\\
&+\;\frac{1}{T}4\Big(\tilde\beta'\tilde A\tilde\beta\,(\operatorname{tr}\tilde B)+\tilde\beta'\tilde B\tilde\beta\,(\operatorname{tr}\tilde A)\Big)\Bigg(\operatorname{tr}(\Sigma)(\operatorname{tr}(\Sigma^2))\;+\;(E[X^4]-1)\operatorname{tr}(\Sigma^3)\Bigg)\\
&\frac{1}{T}\operatorname{tr}(E[(\beta' Z_t A Z_t\beta)Z_t B]+E[(\beta' Z_t B Z_t\beta)Z_t A])\\
&\sim\;\frac{1}{T}\Big(\tilde\beta'\tilde A\tilde\beta\,(\operatorname{tr}\tilde B)+\tilde\beta'\tilde B\tilde\beta\,(\operatorname{tr}\tilde A)\Big)\,(\operatorname{tr}\Sigma)^3+2E[\|\tilde\beta\|^2]\frac{1}{T^2}(\operatorname{tr}\tilde A)(\operatorname{tr}\tilde B)\operatorname{tr}(\Sigma)\operatorname{tr}(\Sigma^2)\\
&\frac{1}{T}((\kappa_\varepsilon-1)\operatorname{tr} E[Z_t A Z_t B]\;+\;E[\operatorname{tr}(Z_t A)\operatorname{tr}(Z_t B)])\\
&\sim\;\Big((\operatorname{tr}\tilde A)(\operatorname{tr}\tilde B)\;+\;2\operatorname{tr}(\tilde A\tilde B)\Big)(\operatorname{tr}\Sigma)^2\frac{1}{T}\\
&+2\Big((\operatorname{tr}\tilde A)(\operatorname{tr}\tilde B)\;-\;\operatorname{tr}(\tilde A\tilde B)\Big)\operatorname{tr}(\Sigma^2)\frac{1}{T^2}
\end{aligned}
\tag{185}
$$

where we have used that

$$\begin{aligned}
&\Big(\operatorname{tr}(\Sigma^2)(\operatorname{tr}(\Sigma))^2 - 2(\operatorname{tr}\Sigma)(\operatorname{tr}(\Sigma^3)) + 2\operatorname{tr}(\Sigma^4) - (\operatorname{tr}(\Sigma^2))^2 \\
&+ \; E[X^4]((\operatorname{tr}(\Sigma^2))^2 - \operatorname{tr}(\Sigma^4)) \\
&+ \; 2E[X^4]((\operatorname{tr}\Sigma)(\operatorname{tr}(\Sigma^3)) - \operatorname{tr}(\Sigma^4)) + \; E[X^6]\operatorname{tr}(\Sigma^4)\Big) \;\sim\; \operatorname{tr}(\Sigma^2)(\operatorname{tr}(\Sigma))^2
\end{aligned} \tag{186}$$

□

Lemma 18 *Define $\psi_{*,1}$ through the equation*

$$b_*\psi_{*,1} \;=\; \operatorname{tr}((\Sigma_{F,t}\Psi) + \nu_F'\Psi\nu_F)) \,. \tag{187}$$

Then, we have

$$\frac{1}{T}\operatorname{tr} E[\beta\beta' F_{t_1}F_{t_1}'F_{t_1}F_{t_1}'Q] \;\sim\; \frac{1}{T}\operatorname{tr}(\Psi)\,(\operatorname{tr}(\Sigma))^2(b_*\operatorname{tr}\Sigma\psi_{*,1}+1)E[\beta'\Psi Q\beta]$$

for any uniformly bounded Q that is independent of F.

Proof of Lemma 18. We have

$$\frac{1}{T}\operatorname{tr} E[\beta\beta' F_{t_1}F_{t_1}'F_{t_1}F_{t_1}'Q] \;=\; \frac{1}{T}\operatorname{tr} E[F_{t_1}'F_{t_1}F_{t_1}'Q\beta\beta'F_{t_1}] \tag{188}$$

and hence we are in a position to apply Lemmas 16 and 17 with the two matrices given by $A = I$ and $B = \Psi^{1/2}Q\beta\beta'\Psi^{1/2}$ so that $\tilde{A} = \Psi$ and $\tilde{B} = \Psi^{1/2}Q\beta\beta'\Psi^{1/2}$. Thus, (188) is the

sum of the following terms:

$$
\begin{aligned}
&\frac{1}{T}\operatorname{tr} E[Z_t\beta\beta' Z_t A Z_t\beta\beta' Z_t B] \\
&\sim \Big((\tilde\beta'\Psi\tilde\beta)\operatorname{tr}(\Psi^{1/2}Q\beta\beta'\Psi^{1/2}) + (\tilde\beta'\Psi^{1/2}Q\beta\beta'\Psi^{1/2}\tilde\beta)\operatorname{tr}(\Psi)\Big)\|\tilde\beta\|^2 \operatorname{tr}(\Sigma^2)(\operatorname{tr}(\Sigma))^2\frac{1}{T} \\
&+ E[\|\tilde\beta\|^4]((\operatorname{tr}\Psi)(\operatorname{tr}\Psi^{1/2}Q\beta\beta'\Psi^{1/2}) + 2\operatorname{tr}(\Psi\Psi^{1/2}Q\beta\beta'\Psi^{1/2}))(\operatorname{tr}(\Sigma^2))^2\frac{1}{T} \\
&\frac{1}{T}2\operatorname{tr}(E[Z_t\beta\beta' Z_t A Z_t B] + E[Z_t\beta\beta' Z_t B Z_t A]) \\
&\sim \frac{1}{T}4\|\tilde\beta\|^2\operatorname{tr}(\Psi\Psi^{1/2}Q\beta\beta'\Psi^{1/2})\operatorname{tr}(\Sigma)\operatorname{tr}(\Sigma^2) \\
&+ \frac{1}{T}4\|\tilde\beta\|^2(\operatorname{tr}(\Psi)\operatorname{tr}(\Psi^{1/2}Q\beta\beta'\Psi^{1/2}) - \operatorname{tr}(\Psi\Psi^{1/2}Q\beta\beta'\Psi^{1/2}))\operatorname{tr}(\Sigma^3) \\
&+ \frac{1}{T}4\Big(\tilde\beta'\Psi\tilde\beta\,(\operatorname{tr}\Psi^{1/2}Q\beta\beta'\Psi^{1/2}) + \tilde\beta'\Psi^{1/2}Q\beta\beta'\Psi^{1/2}\tilde\beta\,(\operatorname{tr}\Psi)\Big)\operatorname{tr}(\Sigma)(\operatorname{tr}(\Sigma^2)) \\
&\frac{1}{T}\operatorname{tr}(E[(\beta' Z_t A Z_t\beta)Z_t B] + E[(\beta' Z_t B Z_t\beta)Z_t A]) \\
&\sim \frac{1}{T}\Big(\tilde\beta'\Psi\tilde\beta\,(\operatorname{tr}\Psi^{1/2}Q\beta\beta'\Psi^{1/2}) + \tilde\beta'\Psi^{1/2}Q\beta\beta'\Psi^{1/2}\tilde\beta\,(\operatorname{tr}\Psi)\Big)(\operatorname{tr}\Sigma)^3 \\
&+ 2E[\|\tilde\beta\|^2]\frac{1}{T}(\operatorname{tr}\Psi)(\operatorname{tr}\Psi^{1/2}Q\beta\beta'\Psi^{1/2})\operatorname{tr}(\Sigma)\operatorname{tr}(\Sigma^2) \\
&\frac{1}{T}(2\operatorname{tr} E[Z_t A Z_t B] + E[\operatorname{tr}(Z_t A)\operatorname{tr}(Z_t B)]) \\
&\sim \Big((\operatorname{tr}\Psi)(\operatorname{tr}\Psi^{1/2}Q\beta\beta'\Psi^{1/2}) + 2\operatorname{tr}(\Psi\Psi^{1/2}Q\beta\beta'\Psi^{1/2})\Big)(\operatorname{tr}\Sigma)^2\frac{1}{T}
\end{aligned}
\tag{189}
$$

Now, $\operatorname{tr}(\beta\beta' D)$ is uniformly bounded almost surely for any bounded D. In addition, Assumption 2 implies that $\operatorname{tr}(\Sigma^2) = o(\operatorname{tr}(\Sigma)^2)$ and $\operatorname{tr}(\Sigma^3) = o(\operatorname{tr}(\Sigma)\operatorname{tr}(\Sigma^2))$. As a result,

many terms become negligible, and we get

$$
\begin{aligned}
&\frac{1}{T}\operatorname{tr} E[Z_t\beta\beta' Z_t A Z_t\beta\beta' Z_t B] \\
&\sim\ (\tilde\beta'\Psi^{1/2}Q\beta\beta'\Psi^{1/2}\tilde\beta)\operatorname{tr}(\Psi)\,\|\tilde\beta\|^2\ \operatorname{tr}(\Sigma^2)(\operatorname{tr}(\Sigma))^2\frac{1}{T} \\
&\frac{1}{T}2\operatorname{tr}(E[Z_t\beta\beta' Z_t A Z_t B]+E[Z_t\beta\beta' Z_t B Z_t A]) \\
&\sim\ \frac{1}{T}4\tilde\beta'\Psi^{1/2}Q\beta\beta'\Psi^{1/2}\tilde\beta\,(\operatorname{tr}\Psi)\operatorname{tr}(\Sigma)(\operatorname{tr}(\Sigma^2)) \\
&\frac{1}{T}\operatorname{tr}(E[(\beta' Z_t A Z_t\beta)Z_t B]+E[(\beta' Z_t B Z_t\beta)Z_t A]) \\
&\sim\ \frac{1}{T}\tilde\beta'\Psi^{1/2}Q\beta\beta'\Psi^{1/2}\tilde\beta\,(\operatorname{tr}\Psi)(\operatorname{tr}\Sigma)^3 \\
&\frac{1}{T}((\kappa_\varepsilon-1)\operatorname{tr} E[Z_t A Z_t B]\ +\ E[\operatorname{tr}(Z_t A)\operatorname{tr}(Z_t B)]) \\
&\sim\ (\operatorname{tr}\Psi)(\operatorname{tr}\Psi^{1/2}Q\beta\beta'\Psi^{1/2})(\operatorname{tr}\Sigma)^2\frac{1}{T}
\end{aligned}
\tag{190}
$$

Recall that $b_* = \operatorname{tr} E[\beta\beta'] = \operatorname{tr}((\Sigma_{F,t}\Psi) + \lambda'\nu_F))$. The first term is of the order $b_*^3 M\operatorname{tr}(\Sigma)\operatorname{tr}(\Sigma^2)$. The second term is of the order $b_*^2 M\operatorname{tr}(\Sigma)\operatorname{tr}(\Sigma^2)$. The third term is of the order of $b_*^2 M(\operatorname{tr}\Sigma)^3$ and hence it dominates the second term as well as the first term because $\operatorname{tr}(\Sigma^2) = o((\operatorname{tr}(\Sigma))^2)$. Thus, we are left with

$$
\begin{aligned}
&\frac{1}{T}\tilde\beta'\Psi^{1/2}Q\beta\beta'\Psi^{1/2}\tilde\beta\,(\operatorname{tr}\Psi)(\operatorname{tr}\Sigma)^3+(\operatorname{tr}\Psi)(\operatorname{tr}\Psi^{1/2}Q\beta\beta'\Psi^{1/2})(\operatorname{tr}\Sigma)^2\frac{1}{T} \\
&\sim\frac{1}{T}b_*\psi_{*,1}\operatorname{tr}(\Psi)\,(\operatorname{tr}(\Sigma))^3E[\beta'\Psi Q\beta]\ +\ (\operatorname{tr}\Psi)E[\beta'\Psi Q\beta](\operatorname{tr}\Sigma)^2\frac{1}{T}
\end{aligned}
\tag{191}
$$

where we have used that, by Lemma 8, $\beta'\Psi^{1/2}\tilde\beta \approx \operatorname{tr}((\Sigma_{F,t}\Psi) + \lambda'\nu_F))$ The proof of Lemma 18 is complete. □

I Expectation of $\hat R^M$

We will, for simplicity, assume $\sigma_* = 1$ and frequently use $\lambda = \nu_F$ notation. Indeed, $\lambda = E[FF']^{-1}E[F] \approx \Psi^{-1}\Psi\nu_F = \nu_F$.

Proposition 6 *We have*

$$E[R^F_{t+1}(z)] \;=\; \frac{\Gamma_{1,1}(z)}{1+\xi(z;c)}\,, \tag{192}$$

where

$$\Gamma_{1,1}(z) \;=\; \lim_{T,P\to\infty} \lambda' E[\Psi(zI+B_T)^{-1}\Psi]\lambda\,. \tag{193}$$

Proof of Proposition 6. We start by computing

$$E[F_{t+1}] \;=\; E[S_t' R_{t+1}] = E[S_t'(S_t\widetilde{F}_{t+1}+\varepsilon_{t+1})] \;=\; \mathrm{tr}(\Sigma)\nu_F \tag{194}$$

and therefore, by (115), we have

$$\begin{aligned} E[R^F_{t+1}(z)] \;&=\; E[\hat\beta(z)'F_{t+1}] \\ &=\; \mathrm{tr}(\Sigma)E[\frac{1}{T}\sum_t F_t'(zI+B_T)^{-1}]\nu_F \;\sim\; E[\frac{1}{T}\sum_t F_t'(zI+B_T)^{-1}]\nu_F\,, \end{aligned} \tag{195}$$

where we have used the normalization $\mathrm{tr}\,\Sigma = 1$. Now, by the interchangeability of F_t across t and the Sherman-Morrison formula, we have

$$\begin{aligned} &E[\frac{1}{T}\sum_t F_t'(zI+B_T)^{-1}]\nu_F \\ &=\; E[F_t'(zI+B_T)^{-1}\Psi]\lambda \;=\; E[F_t'(zI+B_{T,t})^{-1}\frac{1}{1+(T)^{-1}F_t'(zI+B_{T,t})^{-1}F_t}\Psi]\lambda\,, \end{aligned} \tag{196}$$

where

$$B_{T,t} \;=\; \frac{1}{T}\sum_{\tau\neq t} F_\tau F_\tau'.$$

By Lemma 10,

$$(T)^{-1}F_t'(zI + B_{T,t})^{-1}F_t \;\to\; \xi(z;c)$$

is probability and therefore

$$E[F_t'(zI + B_{T,t})^{-1}\frac{1}{1 + (T)^{-1}F_t'(zI + B_{T,t})^{-1}F_t}\Psi]\lambda \;\sim\; \frac{E[F_t'(zI + B_{T,t})^{-1}\nu_F]}{1 + \xi(z;c)}\,, \tag{197}$$

whereas $E[F_t] = \mathrm{tr}(\Sigma\Sigma_\varepsilon)\nu_F$ implies

$$E[F_t'(zI + B_{T,t})^{-1}\nu_F] \;=\; \mathrm{tr}(\Sigma\Sigma_\varepsilon)\lambda' E[\Psi(zI + B_{T,t})^{-1}\nu_F] \;\sim\; \Gamma_{1,1}(z)\,. \tag{198}$$

The proof of Proposition 6 is complete.

□

J Computing Expectations Involving Powers of Ψ

Lemma 19 *Let*

$$\psi_{*,k} \;=\; \lim P^{-1}\,\mathrm{tr}(\lambda'\Psi^k\lambda) \tag{199}$$

and

$$\Gamma_{k,l,T}(z) \;\equiv\; \lambda' E[\Psi^k(zI + B_T)^{-1}\Psi^\ell]\lambda\,. \tag{200}$$

We have

$$\psi_{*,k+\ell} \sim\; z\,\Gamma_{k,\ell,T}(z) \;+\; \left(\psi_{*,k+1}\Gamma_{1,\ell,T}(z) \;+\; \sigma_*\Gamma_{k+1,\ell,T}\right)(1 + \xi(z;c))^{-1} \tag{201}$$

Proof of Lemma 19. Using the Sherman-Morrison formula and Lemma 10, we get

$$F_t'(zI+B_T)^{-1} \;=\; F_t'(zI+B_{T,t})^{-1}(1+(T)^{-1}F_t'(zI+B_{T,t})^{-1}F_t)^{-1} \sim\; F_t'(zI+B_{T,t})^{-1}(1+\xi(z;c))^{-1}$$

We also have

$$\begin{aligned} E[F_tF_t'] \;&=\; ((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2/))\Psi\Sigma_F\Psi \\ &+\; \operatorname{tr}(\Sigma\circ\Sigma)(\kappa-3)\Psi^{1/2}\operatorname{diag}(\Psi^{1/2}\Sigma_F\Psi^{1/2})\Psi^{1/2}+\Psi\Big(\operatorname{tr}(\Sigma\Sigma_\varepsilon)+\operatorname{tr}(\Psi\Sigma_F)\operatorname{tr}(\Sigma^2)\Big) \quad (202) \\ &=\; \widehat{\Sigma}_F \;+\Psi\Sigma_F\Psi \;+\; \sigma_*\Psi\,, \end{aligned}$$

where $\|\widehat{\Sigma}_F\| = o(1)$, and

$$\Sigma_F \;=\; \lambda\lambda' \;+\; \Sigma_F^*\,. \quad (203)$$

We will need the following important observation:

Lemma 20 *For any sequence*

$$\lambda' A_P Q_P \lambda \;\to\; 0 \quad (204)$$

in probability, for any uniformly bounded Q_P (even if they correlate with λ) and any A_P with a uniformly bounded trace norm, such that A_P is independent of λ.

Proof of Lemma 20. We have

$$\begin{aligned}
\lambda' A_P Q_P \lambda &= \operatorname{tr}(\lambda\lambda' A_P Q_P) \\
&\le \|\lambda\lambda' A_P Q_P\|_1 \le \|Q_P\|_\infty \|\lambda\lambda' A_P\|_1 \\
&= \|Q_P\|_\infty \operatorname{tr}((\lambda\lambda' A_P A_P' \lambda\lambda')^{1/2}) = \|Q_P\|_\infty (\lambda' A_P A_P' \lambda)^{1/2} \operatorname{tr}((\lambda\lambda')^{1/2}) = (\lambda' A_P A_P' \lambda)^{1/2} \|\lambda\| \\
&= (\operatorname{tr}(A_P A_P' \lambda\lambda'))^{1/2} \|\lambda\| \to (P^{-1} \operatorname{tr}(\Sigma_\lambda))^{1/2} (P^{-1} \operatorname{tr}(A_P A_P' \Sigma_\lambda))^{1/2} \\
&\le (P^{-1} \operatorname{tr}(\Sigma_\lambda))^{1/2} \|\Sigma_\lambda\|^{1/2} (P^{-1} \operatorname{tr}(A_P A_P'))^{1/2} \to 0
\end{aligned} \tag{205}$$

The proof of Lemma 20 is complete. □

Thus, for any A_P with bounded trace norm, we get

$$\begin{aligned}
\psi_{*,k+\ell} &= P^{-1}\operatorname{tr}(\Psi^{k+\ell}\Sigma_\lambda) \approx \lambda'\Psi^{k+\ell}\lambda = \lambda' E[\Psi^k(zI+B_T)(zI+B_T)^{-1}\Psi^\ell]\lambda \\
&= z\Gamma_{k,\ell,T}(z) + \lambda' E[\Psi^k B_T(zI+B_T)^{-1}\Psi^\ell]\lambda \\
&\underbrace{=}_{\text{symmetry over } t} z\Gamma_{k,\ell,T}(z) + \lambda' E[\Psi^k F_t F_t'(zI+B_T)^{-1}\Psi^\ell]\lambda \\
&\underbrace{=}_{(77)} z\Gamma_{k,\ell,T}(z) + \lambda' E[\Psi^k F_t F_t'(zI+B_{T,t})^{-1}(1+(T)^{-1}F_t'(zI+B_{T,t})^{-1}F_t)^{-1}\Psi^\ell]\lambda \\
&\underbrace{\sim}_{\text{Lemma 10}} z\Gamma_{k,\ell,T}(z) + \lambda' E[\Psi^k F_t F_t'(zI+B_{T,t})^{-1}\Psi^\ell]\lambda(1+\xi(z;c))^{-1} \\
&\underbrace{\sim}_{(202)} z\Gamma_{k,\ell,T}(z) + \lambda' E[\Psi^k(\widehat{\Sigma}_F + \Psi\Sigma_F\Psi + \sigma_*\Psi)(zI+B_{T,t})^{-1}\Psi^\ell]\lambda(1+\xi(z;c))^{-1} \qquad (206) \\
&\sim z\Gamma_{k,\ell,T}(z) + \lambda' E[\Psi^k(\Psi(\Sigma_F+\lambda\lambda')\Psi + \sigma_*\Psi)(zI+B_T)^{-1}\Psi^\ell]\lambda(1+\xi(z;c))^{-1} \\
&\underbrace{\sim}_{(204)} z\Gamma_{k,\ell,T}(z) + \lambda' E[\Psi^k(\nu_F\nu_F' + \sigma_*\Psi)(zI+B_T)^{-1}\Psi^\ell]\lambda(1+\xi(z;c))^{-1} \\
&= z\Gamma_{k,\ell,T}(z) + \lambda'\Psi^{k+1}\lambda E[\nu_F'(zI+B_T)^{-1}\Psi^\ell]\lambda(1+\xi(z;c))^{-1} \\
&+ \nu_F'^{k+1}\sigma_*(zI+B_T)^{-1}\Psi^\ell\lambda(1+\xi(z;c))^{-1} \\
&\sim z\,\Gamma_{k,\ell,T}(z) + \left(\psi_{*,k+1}\Gamma_{1,\ell,T}(z) + \sigma_*\Gamma_{k+1,\ell,T}\right)(1+\xi(z;c))^{-1}
\end{aligned}$$

□

Lemma 21 *Let*

$$\delta(z) = -\sigma_* z^{-1}(1+\xi(z;c))^{-1}. \qquad (207)$$

Then,

$$\Gamma_{1,l}(z) = \frac{z^{-1}P^{-1}\operatorname{tr}(\Psi^{1+\ell}(I-\Psi\delta(z))^{-1}\Sigma_\lambda)}{1-\delta(z)P^{-1}\operatorname{tr}(\Psi^2(I-\Psi\delta(z))^{-1}\Sigma_\lambda)} \qquad (208)$$

and

$$\Gamma_{k,\ell} \;=\; z^{-1}P^{-1}\,\mathrm{tr}(\Psi^{k+\ell}(I-\Psi\delta(z))^{-1}\Sigma_\lambda)-z^{-1}P^{-1}\,\mathrm{tr}(\Psi^{k+1}(I-\Psi\delta(z))^{-1}\Sigma_\lambda)\Gamma_{1,\ell}(1+\xi(z;c))^{-1} \tag{209}$$

Proof. We have

$$\Gamma_{k,\ell} \;=\; a_{k+1}+\delta\,\Gamma_{k+1,\ell} \tag{210}$$

where

$$a_{k+1,\ell} \;=\; z^{-1}(\psi_{*,k+\ell}-\psi_{*,k+1}\Gamma_{1,\ell}(1+\xi(z;c))^{-1}),\;\delta(z) \;=\; -\sigma_* z^{-1}(1+\xi(z;c))^{-1}\,. \tag{211}$$

Let us pick $z>\max(1,\|\Psi\|)$ sufficiently large, so that $\sigma_* z^{-1}(1+\xi(z;c))^{-1}<1$ and[38]

$$|\delta^k\Gamma_{k,\ell}(z)| \;\le\; z^{-k+1}\|\lambda\|^2\|\Psi\|^{k+\ell} \;\to_{k\to\infty}\; 0\,. \tag{212}$$

Then, iterating forward, we get

$$\Gamma_{k,\ell} \;=\; \sum_{\tau=0}^{\infty} a_{k+\tau+1,\ell}\delta^\tau\,. \tag{213}$$

Now,

$$a_{k+\tau+1,\ell} \;=\; z^{-1}(\psi_{*,k+\tau+\ell}-\psi_{*,k+\tau+1}\Gamma_{1,\ell}(1+\xi(z;c))^{-1}),\;\delta(z) \;=\; -\sigma_* z^{-1}(1+\xi(z;c))^{-1}\,. \tag{214}$$

[38]This uniform exponential decay also implies that the infinite sum of the limits equals the limit of the infinite sum, as we pass to the $P\to\infty$ limit.

$$
\begin{aligned}
\Gamma_{1,\ell} &= \sum_{\tau=0}^{\infty} a_{\tau+2,\ell}\delta^{\tau} \\
&= \sum_{\tau=0}^{\infty} z^{-1}(\psi_{*,1+\tau+\ell} - \psi_{*,1+\tau+1}\Gamma_{1,\ell}(1+\xi(z;c))^{-1})\delta^{\tau} \\
&= \sum_{\tau=0}^{\infty} (z^{-1}(P^{-1}\operatorname{tr}(\Psi^{\tau+\ell+1}\Sigma_{\lambda}) - P^{-1}\operatorname{tr}(\Psi^{\tau+2}\Sigma_{\lambda})\Gamma_{1,\ell}(1+\xi(z;c))^{-1}))\delta^{\tau} \\
&= z^{-1}P^{-1}\operatorname{tr}(\Psi^{1+\ell}(I-\Psi\delta(z))^{-1}\Sigma_{\lambda}) - z^{-1}P^{-1}\operatorname{tr}(\Psi^{2}(I-\Psi\delta(z))^{-1}\Sigma_{\lambda})\Gamma_{1,\ell}(1+\xi(z;c))^{-1}\,,
\end{aligned}
\tag{215}
$$

implying that

$$
\Gamma_{1,l} = \frac{z^{-1}P^{-1}\operatorname{tr}(\Psi^{1+\ell}(I-\Psi\delta(z))^{-1}\Sigma_{\lambda})}{1-\delta(z)P^{-1}\operatorname{tr}(\Psi^{2}(I-\Psi\delta(z))^{-1}\Sigma_{\lambda})} \tag{216}
$$

Then, the same argument implies

$$
\Gamma_{k,\ell} = z^{-1}P^{-1}\operatorname{tr}(\Psi^{k+\ell}(I-\Psi\delta(z))^{-1}\Sigma_{\lambda}) - z^{-1}P^{-1}\operatorname{tr}(\Psi^{k+1}(I-\Psi\delta(z))^{-1}\Sigma_{\lambda})\Gamma_{1,\ell}(1+\xi(z;c))^{-1} \tag{217}
$$

Furthermore,

$$
\delta(z) = -\sigma_{*}z^{-1}(1+\xi(z;c))^{-1}, \tag{218}
$$

We have, with $\tilde{\lambda} = \nu_F$, that

$$
\Gamma_{1,1}(z) \approx \frac{z^{-1}\tilde{\lambda}'(I-\Psi\delta(z))^{-1}\tilde{\lambda}}{1-\delta(z)\tilde{\lambda}'(I-\Psi\delta(z))^{-1}\tilde{\lambda}} \tag{219}
$$

□

K Proof of Theorem 3, ii.: Second Moment of $\hat{R}^M$

We start with

Lemma 22 *We have*

$$G(z;c) \;=\; \frac{d}{dz}(z\xi(z;c)) \;\in (0, cz^{-2}] \tag{220}$$

satisfies

$$G(z;c) \;=\; \mathcal{M}(z; Z^*(z;c))\,, \tag{221}$$

where

$$\mathcal{M}(z;Z) \;=\; -1 \;+\; \frac{Z}{z + c\phi(Z)Z^2},\;\; \phi(z) \;=\; P^{-1}\,\mathrm{tr}(E[FF'](zI + E[FF'])^{-2})\,. \tag{222}$$

Proof of Lemma 22. By the master equation,

$$m(z;c) = \frac{1}{1 \;-\; c \;-\; c\,z\,m(z;c)}\, m_{\sigma_*\Psi}\left(\frac{z}{1 \;-\; c \;-\; c\,z\,m(z;c)}\right)\,. \tag{223}$$

whereas, by the definition of the $\xi(z;c)$ function,

$$\frac{c^{-1}\xi(z;c)}{1+\xi(z;c)} \;=\; 1 \;-\; m(-z;c)z\,. \tag{224}$$

and hence

$$\xi(z;c) \;=\; \frac{1 - zm(-z;c)}{c^{-1} - 1 + zm(-z;c)} \tag{225}$$

and hence

$$1+\xi(z;c) \;=\; \frac{c^{-1}}{c^{-1}-1+zm(-z;c)} \;=\; \frac{1}{1-c+czm(-z;c)} \tag{226}$$

Differentiating this identity, we get

$$\xi'(z;c) \;=\; -c(zm(-z;c))'(1+\xi(z;c))^2 \tag{227}$$

Furthermore, differentiating the identity

$$zm(-z;c) \;=\; Z^*(z;c)m(-Z^*(z;c)), \tag{228}$$

we get

$$(zm(-z;c))' \;=\; (zm(-z))'(Z^*)Z^{*\prime} \;=\; (zm(-z))'(Z^*)(1+\xi(z;c)+z\xi'(z;c)) \tag{229}$$

so that

$$\xi'(z;c) \;=\; -c(zm(-z))'(Z^*)(1+\xi(z;c)+z\xi'(z;c))(1+\xi(z;c))^2\,, \tag{230}$$

implying that

$$\xi'(z;c) \;=\; \frac{-c(zm(-z))'(Z^*)(1+\xi(z;c))^3}{1+c(zm(-z))'(Z^*)z(1+\xi(z;c))^2} \tag{231}$$

and hence

$$1+\xi(z;c)+z\xi'(z;c) \;=\; \frac{1+\xi(z;c)}{1+c(zm(-z))'(Z^*)z(1+\xi(z;c))^2} \;=\; \frac{Z^*(z;c)}{z+c(zm(-z))'(Z^*)Z^{*2}(z;c)} \tag{232}$$

□

Let

$$\overline{F_t} \;=\; \sum_t F_t\,.$$

Without loss of generality, we assume that $\kappa = 2$ because all kurtosis terms vanish asymptotically due to their vanishing trace norm. Using Lemma 6, we get[39]

$$\begin{aligned}
&E[(R^F_{t+1}(z))^2] \;=\; E[\frac{1}{T}\overline{F_t}'(zI+B_T)^{-1}F_{t+1}F_{t+1}'(zI+B_T)^{-1}\frac{1}{T}\overline{F_t}]\\
&=\; E[\frac{1}{T}\overline{F_t}'(zI+B_T)^{-1}E_{t-}[F_{t+1}F_{t+1}'](zI+B_T)^{-1}\frac{1}{T}\overline{F_t}]\\
&\underbrace{=}_{Lemma\ 6}\; E[\frac{1}{T}\overline{F_t}'(zI+B_T)^{-1}\bigg(((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))\Psi\Sigma_F\Psi+\Psi\Big(\operatorname{tr}(\Sigma\Sigma_\varepsilon)+\operatorname{tr}(\Psi\Sigma_F)\operatorname{tr}(\Sigma^2)\Big)\bigg)\\
&(zI+B_T)^{-1}\frac{1}{T}\overline{F_t}]\\
&\approx\; E[\frac{1}{T}\overline{F_t}'(zI+B_T)^{-1}\bigg((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)\bigg)\\
&(zI+B_T)^{-1}\frac{1}{T}\overline{F_t}]\\
&=\; \frac{1}{T^2}\sum_{t_1,t_2}E[F_{t_1}(zI+B_T)^{-1}\bigg((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_T)^{-1}F_{t_2}]\\
&\sim\; Term1\;+\;Term2
\end{aligned} \tag{233}$$

with

$$Term1 \;=\; \frac{1}{T}E[F_{t_1}'(zI+B_T)^{-1}\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_T)^{-1}F_{t_1}] \tag{234}$$

[39] E_{t-} denotes the expectation averaging over realizations of S_t and R_{t+1}.

and

$$Term2 \;=\; \frac{T(T-1)}{T^2} E[F_{t_1}'(zI+B_T)^{-1}\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_T)^{-1}F_{t_2}] \tag{235}$$

for any $t_1 \neq t_2$.

K.1 $Term1$ in (234)

We first deal with the first term. Using the Sherman-Morrison formula and Lemma 10, and Lemma 6, we get

$$\begin{aligned}
Term1 \;&=\; \frac{1}{T}\operatorname{tr} E[\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_T)^{-1}F_{t_1}F_{t_1}'(zI+B_T)^{-1}] \\
&\sim\; \frac{1}{T}\operatorname{tr} E[\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_{T,t_1})^{-1}F_{t_1}F_{t_1}'(zI+B_{T,t_1})^{-1}](1+\xi(z;c))^{-2} \\
&\sim \frac{1}{T}\operatorname{tr} E[\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_{T,t})^{-1} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_{T,t_1})^{-1}](1+\xi(z;c))^{-2}
\end{aligned} \tag{236}$$

We can now split this expression into several terms. We have

$$\begin{aligned}
&\frac{1}{T}\operatorname{tr} E[(\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi(zI+B_{T,t})^{-1}(\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi(zI+B_{T,t})^{-1}](1+\xi(z;c))^{-2} \\
&=\; \frac{1}{T}\operatorname{tr} E[\Psi\Sigma_F\Psi(zI+B_{T,t})^{-1}\Psi\Sigma_F\Psi(zI+B_{T,t})^{-1}](1+\xi(z;c))^{-2} \;\to 0
\end{aligned} \tag{237}$$

because

$$\operatorname{tr}(\Sigma_F) \;=\; \operatorname{tr}(\Sigma_{F,t}) \;+\; P^{-1}\|\lambda\|^2 \;=\; o(P) \;+\; O(1) = o(T)\,,$$

and all other matrices involved are uniformly bounded. The second term is

$$\frac{1}{T}\operatorname{tr} E[(\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi(zI+B_{T,t})^{-1}\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Psi(zI+B_{T,t})^{-1}]/(1+\xi(z;c))^2 \;=\; O(T^{-1}) \tag{238}$$

by the same argument. Finally, the last term is

$$(\operatorname{tr}(\Sigma\Sigma_\varepsilon))^2\frac{1}{T}\operatorname{tr} E[\Psi(zI+B_{T,t})^{-1}\Psi(zI+B_{T,t})^{-1}]/(1+\xi(z;c))^2 \tag{239}$$

and it needs to be evaluated directly.

Lemma 23 *We have*

$$\begin{aligned}
&\frac{1}{P}\operatorname{tr} E[F_{t_1}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_2}F_{t_2}'(zI+B_{T,t_1,t_2})^{-1}]\\
&\sim\; \sigma_*^2\frac{1}{P}\operatorname{tr} E[\Psi(zI+B_T)^{-1}\Psi(zI+B_T)^{-1}]\\
&\to\; \Gamma_3(z) \;=\; \left(1-(-z^2m'(-z;c)+2zm(-z;c)+c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2)\right)(1+\xi(z;c))^4\\
&=\; c^{-1}(\xi(z;c)\;+\;z\xi'(z;c))(1+\xi(z;c))^2
\end{aligned} \tag{240}$$

Proof. We have by the Sherman-Morrison formula that

$$\begin{aligned}
&\frac{1}{P}\frac{1}{T}\operatorname{tr} E[F_{t_1}F_{t_1}'(zI+B_T)^{-1}F_{t_1}F_{t_1}'(zI+B_T)^{-1}]\\
&\sim \frac{1}{c}\frac{1}{T^2}E[F_{t_1}'(zI+B_T)^{-1}F_{t_1}F_{t_1}'(zI+B_T)^{-1}F_{t_1}]\\
&=\; c^{-1}E\left[\left(\frac{\frac{1}{T}F_{t_1}'(zI+B_{T,t_1})^{-1}F_{t_1}}{1+\frac{1}{T}F_{t_1}'(zI+B_{T,t_1})^{-1}F_{t_1}}\right)^2\right]\\
&\sim\; c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2
\end{aligned} \tag{241}$$

by Lemma 10. Now,

$$m'(-z;c) \;=\; \lim P^{-1}\operatorname{tr} E[(zI+B_T)^{-2}] \tag{242}$$

and hence

$$
\begin{aligned}
1 &= \frac{1}{P}\operatorname{tr} E[(zI+B_T)(zI+B_T)^{-1}(zI+B_T)(zI+B_T)^{-1}] \\
&= \frac{1}{P}z^2 \operatorname{tr} E[(zI+B_T)^{-2}] \;+\; 2z\,\frac{1}{P}\operatorname{tr} E[(zI+B_T)^{-2}B_T] \\
&+ \frac{1}{P}\operatorname{tr} E[B_T(zI+B_T)^{-1}B_T(zI+B_T)^{-1}] \\
&\sim z^2 m'(-z;c) \;+\; 2z\,\frac{1}{P}\operatorname{tr} E[(zI+B_T)^{-2}(B_T+zI-zI)] \\
&+ \frac{1}{P}\frac{1}{T^2}\sum_{t_1,t_2}\operatorname{tr} E[F_{t_1}F_{t_1}'(zI+B_T)^{-1}F_{t_2}F_{t_2}'(zI+B_T)^{-1}] \\
&= -z^2 m'(-z;c) + 2zm(-z;c) + \frac{1}{P}\frac{1}{T}\operatorname{tr} E[F_{t_1}F_{t_1}'(zI+B_T)^{-1}F_{t_1}F_{t_1}'(zI+B_T)^{-1}] \\
&+ \frac{1}{P}\frac{T(T-1)}{T^2}\operatorname{tr} E[F_{t_1}F_{t_1}'(zI+B_T)^{-1}F_{t_2}F_{t_2}'(zI+B_T)^{-1}] \\
&\sim -z^2 m'(-z;c) + 2zm(-z;c) + c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2 \qquad (243) \\
&+ \frac{1}{P}\operatorname{tr} E[F_{t_1}F_{t_1}'(zI+B_T)^{-1}F_{t_2}F_{t_2}'(zI+B_T)^{-1}] \\
&\sim -z^2 m'(-z;c) + 2zm(-z;c) + c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2 \\
&+ \frac{1}{P}\operatorname{tr} E[F_{t_1}F_{t_1}'(zI+B_{T,t_1})^{-1}F_{t_2}F_{t_2}'(zI+B_{T,t_2})^{-1}]/(1+\xi(z;c))^2 \\
&\sim -z^2 m'(-z;c) + 2zm(-z;c) + c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2 \\
&+ \frac{1}{P}E[F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_2}F_{t_2}'(zI+B_{T,t_1,t_2})^{-1}F_{t_1}]/(1+\xi(z;c))^4 \\
&= -z^2 m'(-z;c) + 2zm(-z;c) + c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2 \\
&+ \frac{1}{P}\operatorname{tr} E[F_{t_1}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_2}F_{t_2}'(zI+B_{T,t_1,t_2})^{-1}]/(1+\xi(z;c))^4
\end{aligned}
$$

where we have defined

$$
B_{T,t_1,t_2} \;=\; \frac{1}{T}\sum_{\tau\notin\{t_1,t_2\}} F_\tau F_\tau'\,. \qquad (244)
$$

We also used that

$$F'_{t_1}(zI + B_T)^{-1} \;\sim\; F'_{t_1}(zI + B_{T,t_1})^{-1}/(1 + \xi(z;c))$$

by Lemma 10 and the Sherman-Morrison formula.

Now,

$$\begin{aligned}
&\frac{1}{P}\operatorname{tr} E[F_{t_1}F'_{t_1}(zI + B_{T,t_1,t_2})^{-1}F_{t_2}F'_{t_2}(zI + B_{T,t_1,t_2})^{-1}] \\
&= \frac{1}{P}\operatorname{tr} E[\Bigg(((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2))\Psi\Sigma_F\Psi \\
&+ \Psi\Big(\operatorname{tr}(\Sigma\Sigma_\varepsilon) + \operatorname{tr}(\Sigma_F\Psi)\operatorname{tr}(\Sigma^2)\Big)\Bigg)(zI + B_{T,t_1,t_2})^{-1}\Bigg(((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2))\Psi\Sigma_F\Psi \\
&+ \Psi\Big(\operatorname{tr}(\Sigma\Sigma_\varepsilon) + \operatorname{tr}(\Sigma_F\Psi)\operatorname{tr}(\Sigma^2)\Big)\Bigg)(zI + B_{T,t_1,t_2})^{-1}]
\end{aligned} \tag{245}$$

which coincides with the expression in (236). By the derivations in formulas (237) and (238), we get

$$\begin{aligned}
&\frac{1}{P}\operatorname{tr} E[F_{t_1}F'_{t_1}(zI + B_{T,t_1,t_2})^{-1}F_{t_2}F'_{t_2}(zI + B_{T,t_1,t_2})^{-1}] \\
&\sim \sigma_*^2\frac{1}{P}\operatorname{tr} E[\Psi(zI + B_T)^{-1}\Psi(zI + B_T)^{-1}]\,,
\end{aligned} \tag{246}$$

and hence

$$\begin{aligned}
1 \;&=\; -z^2m'(-z;c) + 2zm(-z;c) + c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2 \\
&+\; \sigma_*^2\frac{1}{P}\operatorname{tr} E[\Psi(zI + B_T)^{-1}\Psi(zI + B_T)^{-1}]/(1+\xi(z;c))^4
\end{aligned} \tag{247}$$

Finally,

$$\frac{\xi(z;c)}{1+\xi(z;c)} \;=\; c(1 - zm(-z;c)) \tag{248}$$

$$
\begin{aligned}
&(1+z^2m'(-z;c)-2zm(-z;c)-c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2)(1+\xi(z;c))^4\\
&= (\frac{d}{dz}(z(1-zm(-z;c)))-c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2)(1+\xi(z;c))^4\\
&= c^{-1}\Big(\frac{d}{dz}(\frac{z\xi(z;c)}{1+\xi(z;c)})(1+\xi(z;c))^2-(\xi(z;c))^2\Big)(1+\xi(z;c))^2\\
&= c^{-1}\Big(\frac{d}{dz}(z-\frac{z}{1+\xi(z;c)})(1+\xi(z;c))^2-(\xi(z;c))^2\Big)(1+\xi(z;c))^2\\
&= c^{-1}\Big(\Big(1-\frac{1}{1+\xi(z;c)}+\frac{z\xi'(z;c)}{(1+\xi(z;c))^2}\Big)(1+\xi(z;c))^2-(\xi(z;c))^2\Big)(1+\xi(z;c))^2\\
&= c^{-1}(\xi(z;c)\ +\ z\xi'(z;c))(1+\xi(z;c))^2
\end{aligned}
\tag{249}
$$

The proof of Lemma 23 is complete. □

We conclude that the first term from (233) characterized in (236) satisfies

$$
\begin{aligned}
Term1\ &=\ \frac{1}{T}E[F'_{t_1}(zI+B_T)^{-1}\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_T)^{-1}F_{t_1}]\\
&\sim\ (1+\xi(z;c))^{-2}c\Gamma_3(z)
\end{aligned}
\tag{250}
$$

because $1/T \sim c/P$.

K.2 $Term2$ in (235)

We now proceed with the second term (235). By the Sherman-Morrison formula and Lemma 10,

$$
\begin{aligned}
& E[F_{t_1}'(zI+B_T)^{-1}\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_T)^{-1}F_{t_2}] \\
& \sim\ E[F_{t_1}'(zI+B_{T,t_1})^{-1}\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_{T,t_2})^{-1}F_{t_2}]/(1+\xi(z;c))^2 \\
& \sim\ E[F_{t_1}'\Big((zI+B_{T,t_1,t_2})^{-1}-\frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_2}F_{t_2}'(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F_{t_2}'(zI+B_{T,t_1,t_2})^{-1}F_{t_2}}\Big) \\
& \Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)\Big((zI+B_{T,t_1,t_2})^{-1} \\
& -\frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}\Big)F_{t_2}]/(1+\xi(z;c))^2 \\
& =\ Term1\ +\ Term2\ +\ Term3
\end{aligned}
\tag{251}
$$

where

$$
\begin{aligned}
Term1\ &=\ E[F_{t_1}'(zI+B_{T,t_1,t_2})^{-1} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_{T,t_1,t_2})^{-1}F_{t_2}]/(1+\xi(z;c))^2 \\
Term2\ &=\ -2E[F_{t_1}'(zI+B_{T,t_1,t_2})^{-1} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big) \\
&\times\frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}F_{t_2}]/(1+\xi(z;c))^2 \\
Term3\ &=\ E[F_{t_1}'\frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_2}F_{t_2}'(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F_{t_2}'(zI+B_{T,t_1,t_2})^{-1}F_{t_2}} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)\frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}F_{t_2}]/(1+\xi(z;c))^2
\end{aligned}
\tag{252}
$$

We now analyze each term separately.

K.3 *Term*1 in (252)

We will need the following lemma.

Lemma 24 *We have*

$$F(A) \;=\; \lambda' E[(zI+B_T)^{-1}A(zI+B_T)^{-1}]\lambda \;\to\; 0 \tag{253}$$

for any A with uniformly bounded trace norm, with A independent of λ.

Proof of Lemma 24. We know from Lemma 20 that $\lambda' E[A(zI+B_T)^{-1}]\lambda \to 0$. Furthermore,

$$\begin{aligned}
&\lambda' E[A(zI+B_T)^{-1}]\lambda \;=\; \lambda' E[(zI+B_T)^{-1}(zI+B_T)A(zI+B_T)^{-1}]\lambda \\
&\underbrace{=}_{symmetry} \; z\lambda' E[(zI+B_T)^{-1}A(zI+B_T)^{-1}]\lambda \;+\; \frac{1}{T}\lambda' E[(zI+B_T)^{-1}F_tF_t'A(zI+B_T)^{-1}\lambda] \\
&= \; z\lambda' E[(zI+B_T)^{-1}A(zI+B_T)^{-1}]\lambda \\
&+ \; E[\Big((zI+B_{T,t})^{-1} - \frac{\frac{1}{T}(zI+B_{T,t})^{-1}F_tF_t'(zI+B_{T,t})^{-1}}{1+\frac{1}{T}F_t'(zI+B_{T,t})^{-1}F_t}\Big)F_tF_t'A(zI+B_T)^{-1}\lambda] \\
&\approx \; z\lambda' E[(zI+B_T)^{-1}A(zI+B_T)^{-1}]\lambda \;+\; (1+\xi(z;c))^{-1}\lambda' E[(zI+B_{T,t})^{-1}F_tF_t'A \\
&\times \Big((zI+B_{T,t})^{-1} - \frac{\frac{1}{T}(zI+B_{T,t})^{-1}F_tF_t'(zI+B_{T,t})^{-1}}{1+\xi(z;c)}\Big)\lambda] \\
&= \; z\lambda' E[(zI+B_T)^{-1}A(zI+B_T)^{-1}]\lambda \\
&+ \; (1+\xi(z;c))^{-1}\lambda' E[(zI+B_{T,t})^{-1}\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)A(zI+B_{T,t})^{-1}]\lambda \\
&- (1+\xi(z;c))^{-2}\lambda' E[(zI+B_{T,t})^{-1}F_tF_t'A\frac{1}{T}(zI+B_{T,t})^{-1}F_tF_t'(zI+B_{T,t})^{-1}]\lambda \\
&\approx \; z\lambda' E[(zI+B_T)^{-1}A(zI+B_T)^{-1}]\lambda \\
&+ \; (1+\xi(z;c))^{-1}\lambda' E[(zI+B_{T,t})^{-1}\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)A(zI+B_{T,t})^{-1}]\lambda \\
&- Q(z)(1+\xi(z;c))^{-2}\lambda' E[(zI+B_{T,t})^{-1}F_tF_t'(zI+B_{T,t})^{-1}]\lambda
\end{aligned} \tag{254}$$

where

$$Q(z) \;=\; F_t' A \frac{1}{T}(zI + B_{T,t})^{-1} F_t \;\to\; T^{-1} \operatorname{tr} E[\Psi A (zI + B_{T,t})^{-1}] \;\to 0 \tag{255}$$

because $\|A\|_1 = o(P)$ by assumption, and

$$\begin{aligned} &\lambda' E[(zI + B_{T,t})^{-1} F_t F_t' (zI + B_{T,t})^{-1}]\lambda \\ &= \lambda' E[(zI + B_{T,t})^{-1} \Big((\operatorname{tr}\Sigma)^2 \Psi \Sigma_F \Psi + \Psi \operatorname{tr}(\Sigma\Sigma_\varepsilon) \Big) (zI + B_{T,t})^{-1}]\lambda \;=\; O(1)\,. \end{aligned} \tag{256}$$

Thus, we get

$$o(1) \;\approx\; zF(A) \;+\; (1 + \xi(z;c))^{-1} F((\Psi\Sigma_F\Psi + \Psi)A) \tag{257}$$

where $o(1)$ is uniform, and the same iterative argument as in the proof of Lemma 21 give a power series representation for $F((\Psi\Sigma_F\Psi + \Psi)^k A)$ for all k, and the same uniform boundedness argument implies that $F(A) = 0$. The proof of Lemma 24 is complete. □

Now, $E[F_t] = \operatorname{tr}(\Sigma\Sigma_\varepsilon)\nu_F$ and therefore

$$\begin{aligned} &(1 + \xi(z;c))^2 Term1 \;=\; E[F_{t_1}' (zI + B_{T,t_1,t_2})^{-1} \\ &\Big((\operatorname{tr}\Sigma)^2 \Psi \Sigma_F \Psi + \Psi \operatorname{tr}(\Sigma\Sigma_\varepsilon) \Big) (zI + B_{T,t_1,t_2})^{-1} F_{t_2}] \\ &\sim\; \frac{1}{N^3} (\operatorname{tr}(\Sigma))^2 \nu_F'\, E[(zI + B_{T,t_1,t_2})^{-1} \\ &\Big((\operatorname{tr}\Sigma)^2 \Psi (\Sigma_{F,t} + \lambda\lambda') \Psi + \Psi \operatorname{tr}(\Sigma\Sigma_\varepsilon) \Big) (zI + B_{T,t_1,t_2})^{-1}] \nu_F \\ &=\; \frac{1}{N^4} (\operatorname{tr}(\Sigma))^2 \nu_F'\, E[(zI + B_{T,t_1,t_2})^{-1} (\operatorname{tr}\Sigma)^2 \Psi \Sigma_{F,t} \Psi (zI + B_{T,t_1,t_2})^{-1}] \nu_F \\ &+ \frac{1}{N^4} (\operatorname{tr}(\Sigma))^2 \nu_F'\, E[(zI + B_{T,t_1,t_2})^{-1} (\operatorname{tr}\Sigma)^2 \Psi \lambda \nu_F' (zI + B_{T,t_1,t_2})^{-1}] \nu_F \\ &+ \frac{1}{N^3} (\operatorname{tr}(\Sigma))^2 \nu_F'\, E[(zI + B_{T,t_1,t_2})^{-1} (\operatorname{tr}\Sigma\Sigma_\varepsilon) \Psi (zI + B_{T,t_1,t_2})^{-1}] \nu_F \\ &\sim\; \Gamma_{1,1}(z)^2 \;+\; \Gamma_{4,T}(z)\,, \end{aligned} \tag{258}$$

where Γ_4 is defined in the following lemma.

Lemma 25 *We have*

$$\begin{aligned} &\sigma_* \nu_F' E[(zI + B_{T,t_1,t_2})^{-1} \Psi (zI + B_{T,t_1,t_2})^{-1}] \nu_F \;=\; \Gamma_{4,T}(z) \\ &\to\; \Gamma_4(z) \;=\; \frac{\Gamma_{1,1}(z) + z\Gamma_{1,1}'(z) - (\Gamma_{1,1}(z))^2 (1 + \xi(z;c))^{-2}}{(1 + \xi(z;c))^{-2}} \end{aligned} \tag{259}$$

Proof. We have by the symmetry across t and the Sherman-Morrison formula and Lemma 10 that

$$\begin{aligned} &\Gamma_{1,1}(z) \;\sim\; \lambda' E[\Psi (zI + B_T)^{-1} \Psi] \lambda \;=\; \lambda' E[\Psi (zI + B_T)^{-1} (zI + B_T)(zI + B_T)^{-1} \Psi] \lambda \\ &=\; z\, \lambda' E[\Psi (zI + B_T)^{-1} (zI + B_T)^{-1} \Psi] \lambda \;+\; \lambda' E[\Psi (zI + B_T)^{-1} B_T (zI + B_T)^{-1} \Psi] \lambda \\ &=\; -z\, \Gamma_{1,1,T}'(z) \;+\; \lambda' E[\Psi (zI + B_T)^{-1} \frac{1}{T} \sum_t F_t F_t' (zI + B_T)^{-1} \Psi] \lambda \\ &=\; -z\, \Gamma_{1,1,T}'(z) \;+\; \lambda' E[\Psi (zI + B_T)^{-1} F_t F_t' (zI + B_T)^{-1} \Psi] \lambda \\ &\sim\; -z\, \Gamma_{1,1,T}'(z) \;+\; \lambda' E[\Psi (zI + B_{T,t})^{-1} F_t F_t' (zI + B_{T,t})^{-1} \Psi] \lambda (1 + \xi(z;c))^{-2} \\ &=\; -z\, \Gamma_{1,1,T}'(z) \\ &+\; \lambda' E[\Psi (zI + B_{T,t})^{-1} \Big(((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2)) \Psi \Sigma_F \Psi \\ &+\; \Psi \Big(\operatorname{tr}(\Sigma \Sigma_\varepsilon) + \operatorname{tr}(\Sigma_F \Psi) \operatorname{tr}(\Sigma^2) \Big) \Big) (zI + B_{T,t})^{-1} \Psi] \lambda (1 + \xi(z;c))^{-2} \\ &\sim\; -z\, \Gamma_{1,1,T}'(z) \;+\; (\Gamma_{1,1}(z))^2 (1 + \xi(z;c))^{-2} \\ &+\; \Gamma_{4,T}(z) (1 + \xi(z;c))^{-2} \end{aligned} \tag{260}$$

The claim follows now because $\Gamma_{1,1,T}'(z) \to \Gamma_{1,1}'(z)$ by standard properties of analytic functions. The proof of Lemma 25 is complete. □

K.4 $Term2$ in (252)

The next term in (252) is (note the factor of 2 as it appears two times):

$$
\begin{aligned}
Term2 \;&=\; -2E[F_{t_1}'(zI+B_{T,t_1,t_2})^{-1} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big) \\
&\times \frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}F_{t_2}]/(1+\xi(z;c))^2 \\
&=\; -2E[F_{t_1}'(zI+B_{T,t_1,t_2})^{-1} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big) \\
&\times \frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}\nu_F]\operatorname{tr}(\Sigma)/(1+\xi(z;c))^2 \\
&\sim\; -2E[F_{t_1}'(zI+B_{T,t_1,t_2})^{-1} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big) \\
&\times \frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}\nu_F]/(1+\xi(z;c))^2 \\
&= -2(1+\xi(z;c))^{-2}E[X_TY_T],
\end{aligned}
\tag{261}
$$

where we have used that

$$
E[F_{t_2}] \;=\; \nu_F\,, \tag{262}
$$

and where

$$
\begin{aligned}
Y_T \;&=\; F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}\lambda \\
X_T \;&=\; F_{t_1}'(zI+B_{T,t_1,t_2})^{-1} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big) \\
&\times \frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}{1+\frac{1}{T}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}
\end{aligned}
\tag{263}
$$

We will need the following technical lemma whose proof follows directly from the Cauchy-Schwarz inequality.

Lemma 26 *If $X_T \to X$ in probability and is uniformly bounded and $E[Y_T^2]$ is uniformly bounded. Then,*

$$E[(X_T - X)Y_T] \ \to \ 0$$

Then, we will need

Lemma 27 *We have*

$$E[(Y_T)^2]$$

is uniformly bounded whereas

$$E[Y_T] = E[F'_{t_1}(zI + B_{T,t_1,t_2})^{-1}\lambda] \ \to \ \Gamma_{1,1}(z)\,. \tag{264}$$

Proof. Recall that

$$\lambda'\Psi^k(zI + B_T)^{-1}\Psi^\ell\lambda \ \to \ \Gamma_{k,l}(z) \tag{265}$$

by Lemma 21.

We have

$$
\begin{aligned}
& E[\big(F_{t_1}'(zI + B_{T,t_1,t_2})^{-1}\lambda\big)^2] \\
&= E[F_{t_1}'(zI + B_{T,t_1,t_2})^{-1}\lambda\lambda'(zI + B_{T,t_1,t_2})^{-1}F_{t_1}] \\
&= \operatorname{tr} E[(zI + B_{T,t_1,t_2})^{-1}\lambda\lambda'(zI + B_{T,t_1,t_2})^{-1}F_{t_1}F_{t_1}'] \\
&\sim \operatorname{tr} E[(zI + B_{T,t_1,t_2})^{-1}\lambda\lambda'(zI + B_{T,t_1,t_2})^{-1} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)] \;\leq\; Kz^{-2}
\end{aligned}
\tag{266}
$$

for some $K > 0$. The proof of Lemma 27 is complete. □

Recall that

$$
Y_T \;=\; F_{t_1}'(zI + B_{T,t_1,t_2})^{-1}\lambda
$$

and

$$
\begin{aligned}
X_T \;&=\; F_{t_1}'(zI + B_{T,t_1,t_2})^{-1} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big) \\
&\times \frac{\frac{1}{T}(zI + B_{T,t_1,t_2})^{-1}F_{t_1}}{1 + \frac{1}{T}F_{t_1}'(zI + B_{T,t_1,t_2})^{-1}F_{t_1}}
\end{aligned}
\tag{267}
$$

Now, we know from the proof of Lemma 14 that

$$
\frac{1}{T}F_t'AF_t - \frac{1}{T}\operatorname{tr}(AE[F_tF_t']) \;\to\; 0
$$

in L_2 and

$$
\begin{aligned}
& F'_{t_1}(zI + B_{T,t_1,t_2})^{-1} \\
& \Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)\frac{1}{T}(zI + B_{T,t_1,t_2})^{-1}F_{t_1} \\
& \sim \ \frac{1}{T}\operatorname{tr} E[(zI + B_{T,t_1,t_2})^{-1}\Big(\Psi(\Sigma_{F,t} + \lambda\lambda')\Psi + \sigma_*\Psi\Big) \\
& \times (zI + B_{T,t_1,t_2})^{-1}\Big(\Psi(\Sigma_{F,t} + \lambda\lambda')\Psi + \sigma_*\Psi\Big)] \\
& \overbrace{\qquad\qquad}^{\sim} \\
& \text{(204) } \textit{and Lemma 24} \\
& \frac{1}{T}\operatorname{tr} E[(zI + B_{T,t_1,t_2})^{-1}\Big(\Psi\lambda\nu'_F + \sigma_*\Psi\Big) \\
& \times (zI + B_{T,t_1,t_2})^{-1}\Big(\Psi\lambda\lambda'\Psi + \sigma_*\Psi\Big)] \\
& \sim \ \frac{1}{T}\operatorname{tr} E[(zI + B_{T,t_1,t_2})^{-1}\Psi\lambda\lambda'\Psi(zI + B_{T,t_1,t_2})^{-1}\Psi\lambda\nu'_F] \\
& + \ 2\frac{1}{T}\operatorname{tr} E[(zI + B_{T,t_1,t_2})^{-1}\Psi\lambda\lambda'\Psi(zI + B_{T,t_1,t_2})^{-1}\Psi\sigma_*] \\
& + \ \sigma_*^2\frac{1}{T}\operatorname{tr} E[(zI + B_{T,t_1,t_2})^{-1}\Psi(zI + B_{T,t_1,t_2})^{-1}\Psi] \\
& \sim \ c\Gamma_3(z)
\end{aligned}
\tag{268}
$$

by Lemma (23) because the λ-terms are $O(T^{-1})$. Furthermore, X_T is uniformly bounded by the Cauchy-Schwarz inequality. Thus,

$$X_T \ \to \ \frac{c\Gamma_3(z)}{1 + \xi(z;c)}$$

and

$$E[Y_T] \to \Gamma_{1,1}(z)$$

by Lemma 27, and Lemma 26 and formula (261) imply that

$$Term2 \;\sim -2\frac{c\Gamma_3(z)\Gamma_{1,1}(z)}{(1+\xi(z;c))^3}\,. \tag{269}$$

K.5 $Term3$ in (252)

Finally, we now deal with $Term3$ in (252).

Lemma 28 *$Term3$ in (252) converges to zero.*

Proof of Lemma 28. We have

$$\begin{aligned} Term3 \;&=\; E[F'_{t_1}\frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_2}F'_{t_2}(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F'_{t_2}(zI+B_{T,t_1,t_2})^{-1}F_{t_2}} \\ &\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)\frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}F'_{t_1}(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F'_{t_1}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}F_{t_2}]/(1+\xi(z;c))^2 \\ &=\; E[X_TY_T]\,/(1+\xi(z;c))^2\,, \end{aligned} \tag{270}$$

where we have defined

$$X_T \;=\; \frac{\left(\frac{1}{T}F'_{t_1}(zI+B_{T,t_1,t_2})^{-1}F_{t_2}\right)^2}{(1+\frac{1}{T}F'_{t_1}(zI+B_{T,t_1,t_2})^{-1}F_{t_1})(1+\frac{1}{T}F'_{t_2}(zI+B_{T,t_1,t_2})^{-1}F_{t_2})}$$

and

$$Y_T \;=\; F'_{t_2}(zI+B_{T,t_1,t_2})^{-1}\Big(\Psi\Sigma_F\Psi \;+\; \sigma_*\Psi\Big)(zI+B_{T,t_1,t_2})^{-1}F_{t_1}\,.$$

The first observation is that X_T is uniformly bounded by the Cauchy-Schwarz inequality and has a $O(1/T)$ L_2-norm by Lemma 29. Since the first component of Y_T,

$$F'_{t_2}(zI+B_{T,t_1,t_2})^{-1}\Psi\Sigma_F\Psi(zI+B_{T,t_1,t_2})^{-1}F_{t_1}\,.$$

has a $o(T)$ L_2-norm, we get that this part is negligible by Lemma 26.

Lemma 29 *We have that*

$$E[(F_{t_1}' A F_{t_2})^2] \;=\; O(\|A\|_1 \, \|A\|_\infty) \, .$$

for any A. Thus,

$$\left(\frac{1}{T} F_{t_1}' (zI + B_{T,t_1,t_2})^{-1} F_{t_2} \right)^2$$

converges to zero in L_1, while

$$F_{t_2}' (zI + B_{T,t_1,t_2})^{-1} \Psi \Sigma_F \Psi (zI + B_{T,t_1,t_2})^{-1} F_{t_1}$$

has a uniformly bounded L_2-norm because $\mathrm{tr}(\Sigma_F) = o(T)$.

Proof. We have

$$\begin{aligned} & E[(F_{t_1}' A F_{t_2})^2] \;=\; N^{-2} E[F_{t_1}' A F_{t_2} F_{t_2}' A F_{t_1}] \\ & =\; N^{-2} \, \mathrm{tr} \, E[A F_{t_2} F_{t_2}' A F_{t_1} F_{t_1}'] \\ & \sim \; \mathrm{tr} \, E[A \left(\Psi \Sigma_F \Psi \;+\; \sigma_* \Psi \right) \\ & \times A \left(\Psi \Sigma_F \Psi \;+\; \sigma_* \Psi \right))] \end{aligned} \tag{271}$$

The proof of Lemma 29 is complete. □

Lemma 30 *We have*

$$E[(F_{t_1}' A F_{t_2})^4] \;=\; O(P^2)$$

for any uniformly bounded A.

Indeed, Lemma 30 implies that

$$E[X_T^2] \;\le\; T^{-4}E[(F'_{t_1}(zI + B_{T,t_1,t_2})^{-1}F_{t_2})^4] = O(P^2/T^4)$$

while Lemma 29 implies that

$$E[Y_T^2] \;=\; O(P)\,.$$

Thus,

$$|E[X_T Y_T]|^2 \;\le\; E[X_T^2]E[Y_T^2] \;=\; O(P^2/T^4)\,O(P) \;\to\; 0$$

and the claim follows.

Proof of Lemma 30. Without loss of generality, we may assume that A is symmetric. Recall that

$$R_t \;=\; S_{t-1}\beta_t \;+\; \varepsilon_t, \tag{272}$$

and

$$F_t \;=\; S'_{t-1}R_t \;=\; S'_{t-1}S_{t-1}\beta_t + S'_{t-1}\varepsilon_t \;=\; Z_t\beta \;+\; S'_{t-1}\varepsilon_t \tag{273}$$

and therefore

$$F_tF'_t \;=\; Z_t\beta\beta' Z_t + S'_{t-1}\varepsilon_t\beta' Z_t + Z_t\beta\varepsilon'_t S_{t-1} + S'_{t-1}\varepsilon_t\varepsilon'_t S_{t-1}\,. \tag{274}$$

and formula (149) applied to $t = t_1$ implies

$$\begin{aligned}
E[(F'_{t_1}AF_{t_2})^4] &= E[F'_{t_1}AF_{t_2}F'_{t_2}AF_{t_1}F'_{t_1}AF_{t_2}F'_{t_2}AF_{t_1}] \\
&= \operatorname{tr} E[F_{t_1}F'_{t_1}AF_{t_2}F'_{t_2}AF_{t_1}F'_{t_1}AF_{t_2}F'_{t_2}A] \\
&= \operatorname{tr} E[Z_{t_1}\beta\beta'Z_{t_1}AF_{t_2}F'_{t_2}AZ_{t_1}\beta\beta'Z_{t_1}AF_{t_2}F'_{t_2}A] \\
&+ \operatorname{tr} E[Z_{t_1}\beta\beta'Z_{t_1}AF_{t_2}F'_{t_2}AZ_{t_1}AF_{t_2}F'_{t_2}A] \\
&+ 2\operatorname{tr} E[(\beta'Z_{t_1}AF_{t_2}F'_{t_2}AZ_{t_1}\beta)Z_{t_1}AF_{t_2}F'_{t_2}A] \\
&+ ((\kappa_\varepsilon - 1)\operatorname{tr} E[Z_{t_1}AF_{t_2}F'_{t_2}AZ_{t_1}AF_{t_2}F'_{t_2}A] \\
&+ E[\operatorname{tr}(Z_{t_1}AF_{t_2}F'_{t_2}A)^2]
\end{aligned} \tag{275}$$

We then again apply (149) to $t = t_2$. It is then straightforward to show that the leading contribution will be

$$\begin{aligned}
E[\operatorname{tr}(Z_{t_1}AZ_{t_2}A)^2] &= E[\left(\sum X_{i_1,k_1,t_1}\lambda_{i_1}(\Sigma)X_{i_1,k_2,t_1}\lambda_{k_2}(\tilde{A})X_{i_2,k_2,t_2}\lambda_{i_2}(\Sigma)X_{i_2,k_1,t_2}\lambda_{k_1}(\tilde{A})\right)^2] \\
&= E[\sum X_{i_1,k_1,t_1}\lambda_{i_1}(\Sigma)X_{i_1,k_2,t_1}\lambda_{k_2}(\tilde{A})X_{i_2,k_2,t_2}\lambda_{i_2}(\Sigma)X_{i_2,k_1,t_2}\lambda_{k_1}(\tilde{A}) \\
&\times X_{\tilde{i}_1,\tilde{k}_1,t_1}\lambda_{\tilde{i}_1}(\Sigma)X_{\tilde{i}_1,\tilde{k}_2,t_1}\lambda_{\tilde{k}_2}(\tilde{A})X_{\tilde{i}_2,\tilde{k}_2,t_2}\lambda_{i_2}(\Sigma)X_{\tilde{i}_2,\tilde{k}_1,t_2}\lambda_{\tilde{k}_1}(\tilde{A})]
\end{aligned} \tag{276}$$

Non-zero terms must have that $(i_1,k_1),(i_1,k_2),(\tilde{i}_1,\tilde{k}_1),(\tilde{i}_2,\tilde{k}_2)$ is coming in at least two identical pairs. For example, $k_1 = k_2$, $\tilde{k}_1 = \tilde{k}_2$ will give $\operatorname{tr}(\Sigma)^4(P)^2$. All other terms will be even smaller because more indices should be equal. For example, if $k_1 = \tilde{k}_1$ we ought to have $i_1 = \tilde{i}_1$. The proof of Lemma 30 is complete. □

Thus, (270) converges to zero.

The proof of Lemma 28 is complete. □

Summarizing, we get from (261) and (258), (269), that

$$Term2 \;=\; (1+\xi(z;c))^{-2}(\Gamma_{1,1}(z)^2+\Gamma_4(z)) \;-2\frac{c\Gamma_3(z)\Gamma_{1,1}(z)}{(1+\xi(z;c))^3} \tag{277}$$

and (233) implies

$$\begin{aligned} &E[(R^F_{t+1}(z))^2] \underbrace{\sim}_{(233)} Term1 \;+\; Term2 \\ &\underbrace{\sim}_{(250)} (1+\xi(z;c))^{-2}c\Gamma_3(z) \;+\; Term2 \\ &\underbrace{\sim}_{(277)} (1+\xi(z;c))^{-2}c\Gamma_3(z) \;+\; (1+\xi(z;c))^{-2}(\Gamma_{1,1}(z)^2+\Gamma_4(z)) \;-2\frac{c\Gamma_3(z)\Gamma_{1,1}(z)}{(1+\xi(z;c))^3} \end{aligned} \tag{278}$$

and the final expression follows from Lemma 25:

$$\Gamma_{1,1}(z)^2+\Gamma_4(z) \;=\; \Gamma_{1,1}(z)^2+\frac{\Gamma_{1,1}(z)+z\Gamma'_{1,1}(z)-(\Gamma_{1,1}(z))^2(1+\xi(z;c))^{-2}}{(1+\xi(z;c))^{-2}} \tag{279}$$

L Proof of Theorem 3, iv.: Pricing Errors

Our analysis relies on the quantities

$$\bar{F}_{OS} \;=\; E_{OS}[F] \;\in\; \mathbb{R}^P,\; B_{OS} \;=\; E_{OS}[FF'] \;\in\; \mathbb{R}^{P\times P} \tag{280}$$

where $E_{OS}[X] \;=\; \frac{1}{T_{OS}}\sum_{t\in(T,T+T_{OS}]} X_t$ denotes an out-of-sample time series average. The pricing error properties of the SDF are particularly tractable to derive when the test assets are the P factors F_t that underly the SDF. The out-of-sample pricing error vector is

$$\mathcal{E}_{OS}(z;P;T) \;=\; \frac{1}{T_{OS}}\sum_{t\in(T,T+T_{OS}]} F_t\hat{M}_t(z;P;T) \;\in\; \mathbb{R}^P\,. \tag{281}$$

Finally, following Hansen and Jagannathan (1997), we define the out-of-sample HJD as[40]

$$\mathcal{D}_{OS}^{HJ}(z;P;T) \;=\; \mathcal{E}_{OS}(z;P;T)'\, B_{OS}^{+}\, \mathcal{E}_{OS}(z;P;T)\,, \tag{282}$$

where B_{OS}^{+} is the Moore-Penrose quasi-inverse of the potentially degenerate matrix B_{OS}.

The following is true.

Proposition 7 *We have*

$$\mathcal{D}_{OS}^{HJ}(z;P;T) \;-\; \bar{F}_{OS}' B_{OS}^{+} \bar{F}_{OS} \;=\; -2E_{OS}[\hat{R}_t^M(z;P;T)] \;+\; E_{OS}[(\hat{R}_t^M(z;P;T))^2] \tag{283}$$

When $P > T_{OS}$ and both are sufficiently large, we have

$$\bar{F}_{OS}' B_{OS}^{+} \bar{F}_{OS} \;\approx\; 1 \tag{284}$$

and hence

$$\mathcal{D}_{OS}^{HJ}(z;P;T) \;\approx\; E_{OS}[(1-\hat{M}_t(z;P;T))^2]\,. \tag{285}$$

In expectation, we have

$$\lim_{P,T,T_{OS}\to\infty,\; P/T\to c,\; P>T_{OS}} E[\mathcal{D}_{OS}^{HJ}(z;P;T)] \;=\; (1+G(z;c))\mathcal{R}(Z^*(z;c)), \tag{286}$$

Proposition 7 shows a surprising identity for expected out-of-sample pricing errors. The high complexity error $\mathcal{D}_{OS}^{HJ}(z;c)$ is proportional to the infeasible error $\mathcal{R}(Z^*(z;c))$, subject to implicit regularization (i.e., z is replaced by $Z^*(z;c)$). The proportionality factor equals one plus the complexity risk.

[40] At first glance, it may not be obvious whether we should define the HJD weighting matrix as the in-sample or out-of-sample second moment of factors. Upon further inspection, we find that the out-of-sample second moment is preferable because it allows us to establish a direct correspondence between $\mathcal{D}_{OS}^{HJ}(z;P;T)$ and the out-of-sample SDF Sharpe ratio.

Perhaps surprisingly, the out-of-sample pricing error (286) does not always converge to zero even when $c = z = 0$. Instead,

$$\lim_{P,T,T_{OS}\to\infty,\ P/T\to 0} E[\mathcal{D}_{OS}^{HJ}(0;P;T)] \ \to\ \lim E[\bar{F}_{OS}' B_{OS}^{+} \bar{F}_{OS}] \ -\ E[F]E[FF']^{-1}E[F]\,. \quad (287)$$

Note that $E[F]E[FF']^{-1}E[F] = \mathcal{E}(0)$ is the expected return on the efficient portfolio, whereas $E[\bar{F}_{OS}' B_{OS}^{+} \bar{F}_{OS}]$ can be computed based on the following result.

Lemma 31 *We have*

$$\lim E[\bar{F}_{OS}'(zI + B_{OS})^{-1}\bar{F}_{OS}] \ =\ \frac{\mathcal{E}(Z^*(z;c_{OS}))\ +\ \xi(z;c)}{1+\xi(z;c)}\,. \quad (288)$$

In the ridgeless limit,

$$\lim E[\bar{F}_{OS}' B_{OS}^{+} \bar{F}_{OS}] \ =\ \begin{cases} \mathcal{E}(0)(1-c_{OS})\ +\ c_{OS}, & c_{OS} < 1 \\ 1, & c_{OS} \geq 1\,, \end{cases} \quad (289)$$

and, hence,

$$\lim_{P,T,T_{OS}\to\infty,\ P/T\to 0} E[\mathcal{D}_{OS}^{HJ}(0;P;T)] \ \to\ \begin{cases} c_{OS}(1-\mathcal{E}(0)), & c_{OS} < 1 \\ 1-\mathcal{E}(0), & c_{OS} \geq 1\,. \end{cases} \quad (290)$$

The pricing error (290) remains strictly positive as long as $c_{OS} > 0$. Only when $c = 0$ do we recover the true SDF in the large T limit, so it must price all assets without error. In this case, because the test assets (F_t) are the same factors that underly the SDF, the factors are essentially trying to "price themselves." However, when $c_{OS} > 0$, the out-of-sample factor moments ($\bar{F}_{OS},\ B_{OS}$) are so severely misestimated that $\mathcal{D}_{OS}^{HJ}(0;P;T)$ does not converge to zero *even if we have learned the true SDF in training.*

Finally, we can relate the out-of-sample HJD to the out-of-sample Sharpe ratio. Consider a scale parameter α such that

$$\hat{M}_t(z;P;T) \;=\; 1 - \alpha\, \hat{R}^M_t(z;P;T). \tag{291}$$

Then (285) implies that the optimal α is $\alpha = E_{OS}[\hat{R}^M_t(z;P;T)]/E_{OS}[(\hat{R}^M_t(z;P;T))^2]$, and we get

$$\mathcal{D}^{HJ}_{OS}(z;P;T) = \bar{F}'_{OS} B^+_{OS} \bar{F}_{OS} - SR^2_{OS}(\hat{R}^M(z;P;T)). \tag{292}$$

Thus, the larger the out-of-sample Sharpe ratio, the lower the out-of-sample pricing error. Pricing errors are minimized when the complex feasible ridge SDF achieves the same out-of-sample Sharpe ratio as the *ex-post* out-of-sample tangency portfolio of factors. This is, in essence, an out-of-sample counterpart to the Gibbons et al. (1989) statistic.

Proof of Proposition 7. We have

$$\begin{aligned}
PricingError(z;q;c) \;&=\; E[F'(1-\lambda(z;q)'F(q))]\, E[FF']^{-1} E[(1-\lambda(z;q)'F)F] \\
&=\; (E[F] - E[FF(q)']\lambda(z;q))' E[FF']^{-1} (E[F] - E[FF(q)']\lambda(z;q)) \\
&=\; E[F]' E[FF']^{-1} E[F] - 2 \underbrace{E[R^F(z;q)F'] E[FF']^{-1} E[F]}_{directional} \\
&+ \underbrace{E[R^F(z;q)F'] E[FF']^{-1} E[R^F(z;q)F]}_{risk} \\
&=\; E[F]' E[FF']^{-1} E[F] - 2E[R^F(z;q)] \;+\; E[(R^F(z;q))^2]
\end{aligned} \tag{293}$$

We have

$$E\left[\hat{\lambda}(z;q)' \left(\frac{1}{\hat{T}} \sum_{\tau} (F_\tau(q)) F'_\tau\right) ((0+)I + \hat{B}_{\hat{T}})^{-1} \left(\frac{1}{\hat{T}} \sum_{\tau} F_\tau\right)\right] \tag{294}$$

Now, all matrices here have a block structure:

$$\left(\frac{1}{\hat{T}}\sum_{\tau}(F_{\tau}(q))F_{\tau}'\right) \;=\; [\hat{B}_{\hat{T}}(q)+(0+)I\,,\hat{\Psi}_{1,2}] \tag{295}$$

where $\hat{\Psi}_{1,2}\in\mathbb{R}^{P\times(P-P)}$ and, assuming for simplicity that

$$\left(\frac{1}{\hat{T}}\sum_{\tau}(F_{\tau}(q))F_{\tau}'\right)((0+)I+\hat{B}_{\hat{T}})^{-1} \;=\; [I_{P\times P},0_{P\times(P-P)}] \tag{296}$$

by the definition of the inverse matrix. Namely,

$$(A,B)\begin{pmatrix}A & B\\ C & D\end{pmatrix}^{-1} \;=\; (I,0) \tag{297}$$

Thus,

$$E[R^{F}(z;q)F']E[FF']^{-1} \;=\; \hat{\lambda}(z;q)'(I,0) \tag{298}$$

and hence

$$\begin{aligned} &E[R^{F}(z;q)F']E[FF']^{-1}E[R^{F}(z;q)F]\\ &=\; E[R^{F}(z;q)F']E[FF']^{-1}E[FF']E[FF']^{-1}E[R^{F}(z;q)F]\\ &=\; \hat{\lambda}(z;q)'E[F(q)F(q)']\hat{\lambda}(z;q)\,. \end{aligned} \tag{299}$$

Finally, the last identity follows from

$$\mathcal{D} \;=\; 1 \;-\; 2E[\hat{R}^{M}] \;+\; E[(\hat{R}^{M})^{2}] \;=\; 1-2\mathcal{E}(Z^{*}) \;+\; \mathcal{V}(Z^{*})+G(z;c)\mathcal{R}(Z^{*}) \;=\; \mathcal{R}(Z^{*})+G(z;c)\mathcal{R}(Z^{*}) \tag{300}$$

□