NBER WORKING PAPER SERIES

COMPLEXITY IN FACTOR PRICING MODELS

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

Working Paper 31689
http://www.nber.org/papers/w31689

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2023

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 Fabio Trojani, Neng Wang, and seminar participants at the Wharton School of Management, Temple University, the University of California San Diego, the Imperial College London, and the National University of Singapore. We are especially grateful to Mohammad Pourmohammadi for his numerous constructive comments and suggestions. 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 or the National Bureau of Economic Research.

Complexity in Factor Pricing Models
Antoine Didisheim, Shikun (Barry) Ke, Bryan T. Kelly, and Semyon Malamud
NBER Working Paper No. 31689
September 2023
JEL No. C1,C4,C58,G1,G10,G12,G14,G17

ABSTRACT

We theoretically characterize the behavior of machine learning asset pricing models. We prove that expected out-of-sample model performance—in terms of SDF Sharpe ratio and test asset pricing errors—is improving in model parameterization (or "complexity"). Our empirical findings verify the theoretically predicted "virtue of complexity" in the cross-section of stock returns. Models with an extremely large number of factors (more than the number of training observations or base assets) outperform simpler alternatives by a large margin.

Antoine Didisheim
Extranef 251
University of Lausanne
Lausanne, Swit 1015
Switzerland
antoine.didisheim@unil.ch

Bryan T. Kelly
Yale School of Management
165 Whitney Ave.
New Haven, CT 06511
and NBER
bryan.kelly@yale.edu

Shikun (Barry) Ke
Yale School of Management
barry.ke@yale.edu

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

1 Introduction

In this paper, we develop a statistical theory of heavily parameterized or "complex" asset pricing models. We build our analysis around the stochastic discount factor (SDF). A true SDF, if one exists, is representable as a tradable portfolio (Hansen and Richard, 1987):

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

R_{t+1} is the vector of excess returns on the N risky assets in the economy. Vector $w(X_t)$ contains the SDF's conditional portfolio weights, with X_t representing conditioning variables that span the time t information set.

The literature has primarily investigated (1) with tightly constrained models of the function w. A leading example is the Fama and French (1993) model, which restricts M_{t+1} to a three-parameter combination of pre-defined portfolios.[1] These small SDF parameterizations are motivated in part by the econometric principle of parsimony (e.g. Tukey, 1961; Box and Jenkins, 1970).[2]

1.1 Virtue of Complexity

In contrast to earlier work, an emergent literature documents that extremely large SDF parameterizations achieve smaller out-of-sample pricing errors than their parsimonious

[1] The Fama and French (1993) three-factor SDF may be written as $w(X_t) = c_1 w_{MKT} + c_1 w_{SMB} + c_1 w_{HML}$. The weights w_{MKT}, w_{SMB}, and w_{HML} use information such as assets' market values and book-to-market ratios to construct factor weights in an entirely researcher-dictated manner with no estimated parameters. The three corresponding factors must then be aggregated into an SDF based on three estimated parameters: c_1, c_2, and c_3.

[2] 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 function 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.

predecessors.[3] An understanding of this surprising empirical phenomenon is only beginning to take shape. Kelly et al. (2021) (KMZ henceforth) provide a first step by theoretically characterizing the behavior of high-dimensional machine learning models in return prediction applications. They prove under general conditions that the performance of time series forecasting models—both in terms of forecast accuracy and market timing strategy returns—is increasing in model complexity.

The first contribution of our paper is to move beyond the pure prediction setting of KMZ and theoretically characterize the "virtue of complexity" in highly parameterized asset pricing models. To do so, we build on KMZ in two ways. First, we reorient the statistical objective from time series forecasting to SDF estimation—that is, minimizing pricing errors among test assets and maximizing the SDF Sharpe ratio. Second, we move from the single asset time series setting to a panel setting with an arbitrary number of risky assets. Like KMZ, we study a class of high-dimensional ridge estimators that provides the necessary analytical link to random matrix theory that underpins our theoretical analysis. We explicitly derive key properties of an SDF—its expected out-of-sample Sharpe ratio and pricing errors—when the number of SDF parameters becomes large.[4]

We discuss two interesting interpretations of the virtue of complexity in SDF models. The first is that a complex SDF is a factor pricing model with an extremely large number of factors. A large SDF approximating model of this sort may be written as $w(X_t) \approx \sum_{p=1}^{P} \lambda_p S_p(X_t)$, where each $S_p(X_t)$ is some nonlinear basis function of X_t. Thus, the SDF

[3]Examples include Gu et al. (2020a), Kozak et al. (2020), Kelly et al. (2020), Chen et al. (2023), and Preite et al. (2022), among others. See Kelly and Xiu (2023) for a survey of financial machine learning, including machine learning models of the SDF.

[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. Most importantly, we remove the equal ex-ante predictive power assumption of KMZ and allow for a generic distribution of risk premia across factors.

approximation is

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

and each "factor" $F_{p,t+1}$ is a characteristic-managed portfolio of base assets that uses the nonlinear asset "characteristics" $S_p(X_t)$ as portfolio weights. From this representation, our main result shows that the out-of-sample pricing errors from a factor pricing model are decreasing in the number of factors. This interpretation of the virtue of complexity is a challenge for the traditional APT perspective that a small number of risk factors should capture the risk-return tradeoff among assets. We establish the surprising result 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 adding these factors to the pricing model continually improves its out-of-sample performance.

The second interpretation is based on the theoretical equivalence of an SDF and the mean-variance optimal portfolio. We prove that the out-of-sample efficient portfolio of risky assets improves with the number of managed portfolios $F_{p,t+1}$ that it incorporates. That is, the best mean-variance trading strategy available to an investor combines an exorbitant number of factors, with each of these factors themselves a trading strategy based on a single nonlinear basis function of the conditioning variables. The virtue of SDF complexity means that we can *always* find a new, nonlinear characteristic factor that boosts the out-of-sample Sharpe ratio of the efficient portfolio, even when the "true" model ensures no arbitrage. This interpretation helps rationalize the prominence of "anomaly" portfolios in empirical asset pricing. The abundance of anomalies (or so-called "factor zoo") is not a puzzle to be solved or evidence of a corrupt research process.[5] Instead, it is the theoretically expected

[5] Jensen et al. (Forthcoming) reach a similar conclusion based on the rationale that the risk-return tradeoff 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

outcome in a complex asset pricing environment. 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 Sharpe ratio of the SDF and reduces out-of-sample pricing errors.

As a corollary, we theoretically prove a long-standing conjecture[6] in the literature that managed portfolios are sufficient for approximating the conditionally efficient portfolio. This is true, however, only in the high complexity limit. The true SDF is the conditional Markowitz portfolio of base assets R_t. It thus requires the true conditional covariance matrix of R_t, though this is unobservable and impossible to recover in small samples. We show that with a complex model, the best portfolio available to an investor need not estimate the covariance of R_t at all. Instead, the investor needs only find the *unconditionally optimal* portfolio of factors.

1.2 Limits to Learning

Low complexity settings—with many more observations than parameters to estimate—are the purview of traditional econometrics. In these conditions, the law of large numbers kicks in, and appropriate estimators tend to recover the true model.[7] In high-complexity settings, the number of parameters is large relative to the number of observations—this is the machine learning case. Here, the law of large numbers breaks down and even correctly specified estimators fail to converge on the true model because there is not enough data to go around. 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

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."

[6]This conjecture is clearly articulated by Kozak and Nagel (2023).

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

this as "limits to learning." Perhaps surprisingly, our theory can explicitly quantify limits to learning based on the properties of the training sample.

Limits to learning are intimately connected to the phenomenon of in-sample overfit. When the number of parameters is large relative to the number of observations, in-sample model performance is exaggerated. Overfit is defined as the difference between the in-sample performance of the trained model versus the performance of the true but infeasible model. Note that the term overfit is sometimes a misnomer. In high complexity models, overfit is driven primarily by the dearth of training observations rather than by excessive parameterization.[8]

Together, limits to learning and overfit make up the "complexity wedge," which is the difference between in-sample model performance and expected out-of-sample performance, which decomposes into overfit plus limits to learning:

$$\begin{aligned}
\text{Complexity Wedge} &= \text{In-sample Performance} - \text{Out-of-sample Performance} \\
&= (\text{In-sample Performance} - \text{True Predictability}) \\
&\quad - (\text{Out-of-sample Performance} - \text{True Predictability}) \\
&= \text{Overfit} + \text{Limits to Learning},
\end{aligned}$$

Complexity wedges can be partially mitigated by shrinkage, which reduces the extent of overfit and improves the limits to learning. But as long as complexity is greater than zero, the complexity wedge and its components are positive regardless of the amount of shrinkage. When complexity is low, the law of large numbers eliminates these wedges because expected in-sample, out-of-sample, and true performance are equalized.

[8] This is most easily understood in the case of a correctly specified model which, by definition, has no "excess" parameters. Yet, with insufficient data, it produces in-sample fits that exceed the fit of the true model. For example, when $P \geq T$, one can achieve perfect in-sample fits because there are enough parameters to fit each training observation, but out-of-sample performance will suffer because the model was underfitted (it did not see enough training data to learn the true model).

1.3 Evaluating Machine Learning Asset Pricing Models

Our third main contribution proposes an approach to asset pricing model evaluation for the machine learning context. An immediate implication of the complexity wedge is that traditional approaches to model evaluation—which generally involve some form of significance test for in-sample pricing errors—lose their meaning for heavily parameterized models.

Machine learning asset pricing models must be evaluated based on out-of-sample performance. As a replacement for in-sample comparisons, we recommend comparing highly parameterized models based on the out-of-sample Hansen-Jagannathan distance (HJD). The HJD (Hansen and Jagannathan, 1997) has a number of attractive model comparison properties.[9] First, it averages pricing errors among test assets using a common weighting matrix for all models. This is important because it puts all models on equal footing for comparison, unlike other alpha or GMM-based comparisons. Second, the weighting matrix is economically motivated and can be interpreted as the pricing error of the portfolio of test assets that is *most mispriced* by each model. Third, 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. Fourth, 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.

These properties of the out-of-sample HJD make it a valuable tool for model evaluation in the age of machine learning. Each research team that proposes a new machine learning asset pricing model can publicly post a data set with out-of-sample SDF returns from their model. As new models are developed, they may be compared to previously proposed models based via HJD.[10] A particularly interesting aspect of this protocol is that it forces the researcher

[9] In addition to (Hansen and Jagannathan, 1997), the literature has further highlighted advantages of the HJD, including Kan and Robotti (2009), Chen and Ludvigson (2009), and Kelly and Xiu (2023).

[10] The question of test assets is at the discretion of researchers, and the set of extant models may be re-scrutinized via the HJD for new and perhaps more demanding test assets.

to settle on a univariate SDF representation of their model rather than a multivariate factor representation. This means that the researcher must take a stand on the SDF's factor weights in real time and let the out-of-sample SDF returns fall where they may. For example, evaluating the Fama-French model in this way requires the researcher to decide on MKT, SMB, and HML weights based on training data and report the univariate out-of-sample SDF returns for HJD evaluation. Furthermore, this puts both parsimonious asset pricing models (like Fama-French) and heavily parameterized machine learning models on equal footing for out-of-sample comparison. These comparisons teach us both about the merit of candidate machine learning architectures and the merit of (implicit) priors imposed by researchers who advocate for simpler models.

1.4 Empirical Findings

Our final contribution is an empirical investigation of our theoretical predictions. We design data experiments that mirror our theoretical environment in order to evaluate the role of complexity in the performance of empirical asset pricing models. We study the sample of monthly US stocks and fix the conditioning set—X_t in equation (1)—to be a large collection of 130 stock-level predictors from Jensen et al. (Forthcoming). To vary our empirical models from low to high complexity, we adapt the machine learning method of random features regression (as used in KMZ) to the SDF estimation problem. This converts the fixed set of raw stock characteristics into any desired number P of "random features." The random features constitute an arbitrarily rich set of nonlinear transformations of the raw variables, equivalent to the features engineered in the hidden layer of a wide two-layer neural network.[11] A key attraction of this formulation is the ability to evaluate the effects of empirical SDF

[11] In the first layer of the network, fixed weights (randomly drawn, as opposed to estimated) aggregate the raw inputs X_t which are then fed through a nonlinear activation function to produce the "random features" S_t. In the second layer, the random features are combined with estimated weights to optimize the SDF performance objective (with ridge shrinkage).

complexity simply by varying the number of random features derived from the conditioning set X_t (while holding X_t itself fixed).

Our first empirical result documents a virtue of complexity in pricing the cross-section of returns. We find that the realized out-of-sample performance of the empirical SDF generally improves with model complexity. Increasing the number of model parameters (i.e., the number of factors) consistently raises the out-of-sample SDF Sharpe ratio and reduces its out-of-sample pricing errors in a manner that closely tracks our theoretical predictions. Our empirical "VoC curves," which plot model performance as a function of model complexity, support the theoretical prediction that the gains in approximation accuracy from incorporating more model parameters dominate the statistical costs of estimating those additional parameters. Our high-complexity models also outperform standard low-dimensional benchmark models (such as the Fama-French model) by a large margin.

The virtue of complexity in our empirical asset pricing models is robust. It is not driven by any particular subset of the stock universe. We find nearly identical patterns in complex model behavior when the SDF is estimated only from a subset of the broader sample (e.g., among stocks broken in various market capitalization groups). Recent literature notes that some results in the financial machine learning literature result in infeasible trading strategies that are heavily dependent on predictability induced by limits-to-arbitrage (e.g. Jensen et al., 2022; Avramov et al., 2023). We show that our results are robust to excluding high-turnover signals and when restricting the sample to the largest and most liquid US stocks.

To further elaborate the virtue of complexity in empirical asset pricing models, we replace the large set of 130 stock characteristics in X_t with the smaller set of five characteristics underlying the Fama-French model. We show that even if one were to prefer this narrow set of characteristics, out-of-sample SDF performance is improved by increasing the number of model parameters. Compared to the original Fama-French model, out-of-sample pricing

errors are cut by more than half by including thousands of factors formed from nonlinear transformations of the five underlying Fama-French characteristics.

Recent work by (Kozak et al., 2020) suggests that a successful SDF does not require many factors because the asset pricing properties of those factors are adequately summarized by a small number of their principal components.[12] Their "sparse PC-based SDF" cleverly avoids model complexity through a dimension reduction of the factors. This begs the question: Can complex models be reduced to achieve similar performance with potentially many fewer parameters? We show that this is not possible. For each complex model that we study, we consider replacing its large number of factors with a smaller number of their principal components. We show that dimension reduction significantly impairs performance relative to the full complex model.

1.5 Literature

This paper is broadly related to the financial machine-learning literature.[13] Within this literature it relates particularly closely to machine learning methods that directly estimate the SDF from characteristic-based factors and focus on the link between the SDF and conditionally efficient portfolios. Empirical work in this vein includes Brandt et al. (2009), Kozak et al. (2020), DeMiguel et al. (2020), Chen et al. (2023), Bryzgalova et al. (2020), and Liu et al. (2020). Our paper establishes the theoretical foundations for using heavily parameterized models for SDF estimation.

Our theoretical analysis is also related to the literature on conditional SDF estimation and the conditional Hansen-Jagannathan distance (e.g. Nagel and Singleton, 2011; Antoine et al., 2020a,b; Gagliardini and Ronchetti, 2020, and references therein). As Antoine et al. (2020a) argue, an SDF constructed from factors is "pseudo-true" because it is mis-specified and, even

[12] Relatedly, papers such as Lettau and Pelger (2020), Kelly et al. (2020), and Gu et al. (2020a) demonstrate the success of dimension reduction methods when estimating asset pricing models with a large number of candidate factors.

[13] See Kelly and Xiu (2023) for a recent survey.

in the large sample (zero complexity) limit, only converges to the best approximate SDF in the space of SDFs spanned by factors. Our results imply that mis-specification vanishes in the limit of a very large number of factors and the pseudo-true SDF converges to the truth. However, the cost of this improved approximation is increased complexity, the associated breakdown of the law of large numbers, and estimation errors that survive even in large samples. We derive explicit expressions for these errors (dubbed "limits to learning" in our paper) and show that, despite these errors, SDF approximation improvements afforded by large factor models are well worth the added statistical complexity.

Our paper also relates to machine learning methods for analyzing factor pricing models, including Connor et al. (2012), Fan et al. (2016), Kelly et al. (2020), Lettau and Pelger (2020), Giglio and Xiu (2021), Gu et al. (2020a), and Giglio et al. (2022). These papers provide evidence that introducing conditioning information into latent factor betas improves model performance. Many papers in this line of work argue that retaining a few leading principal components is sufficient to explain the cross-section of returns. This typically results in a low-complexity model environment in which the true conditional SDF can be consistently estimated. In contrast, we work in a theoretical setting where complexity precludes consistent recovery of the SDF. In this sense, our paper is part of an emergent literature analyzing "limits to learning," or the fact that realistically complex asset pricing models cannot be accurately recovered from the limited size of financial data sets (see Martin and Nagel, 2021; Da et al., 2022). Our theory and empirics show that, despite limits to learning in complex environments, high-complexity models deliver more powerful out-of-sample performance than low-complexity models.

Through its coupling with the literature on high-dimensional factor pricing models, our work also relates to the empirical literature surrounding the "factor zoo" and factor replicability, including Cochrane (2011), Harvey et al. (2016), McLean and Pontiff (2016), Hou et al. (2020), Feng et al. (2020), Jensen et al. (Forthcoming), and Chen and Zimmermann

(2021). Our theoretical analysis helps rationalize the continued discovery of unspanned factors that capture nonlinear impacts of conditioning variables on the SDF that are missed by simpler precedent models.

Many papers in the financial machine learning literature focus on predicting asset returns using complex machine learning models, including 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). While this literature is largely agnostic about the link between expected returns and the risk-return tradeoff, its demonstrated success in predicting the cross-section of returns with heavily parameterized models is a manifestation of the virtue of complexity in the panel setting developed in this paper.

In the remainder of the paper, Section 2 outlines the foundational assumptions of our theory. Sections 3 and 4 provide our core theoretical analyses of complex SDF models in the correctly specified and mis-specified settings, respectively. Section 5 documents the empirical virtue of complexity in a canonical data set of monthly US stock returns and stock-level predictors, and Section 6 concludes.

2 Environment

In this section, we present assumptions that form the foundation of our theoretical results in Sections 3 and 4.

2.1 Assets and Conditioning Information

Our first assumption concerns the conditional properties of risky assets in the economy.

Assumption 1 (Returns Have a Conditional Factor Structure) *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*

$$R_{t+1} = S_t \tilde{F}_{t+1} + \varepsilon_{t+1}, \tag{3}$$

where $E_t[\varepsilon_{t+1}] = 0$ *and* $E_t[\varepsilon_{t+1}\varepsilon_{t+1}'] = \Sigma_{\varepsilon,t}$. *The latent factors satisfy* $E_t[\tilde{F}_{t+1}] = \lambda_F$[14] *and* $\Sigma_{F,t} = E_t[\tilde{F}_{t+1}\tilde{F}_{t+1}']$ *satisfies* $\text{tr}(\Sigma_{F,t}) = O(1)$ *as* $P \to \infty$.

The SDF summarizes the risk-return tradeoff among N risky assets in the economy. Thus, to make progress on characterizing SDF behavior, we require assumptions on assets' dependence structure and expected returns. We assume that asset riskiness is describable with a latent factor structure. The factor structure in (3) is generic. It allows for an arbitrary number of factors P, with the only restriction being that the trace of the factor covariance matrix remains bounded.[15]

We are especially interested in understanding the behavior of *conditional* asset pricing models. Assumption 1 defines the relevant conditioning information in this economy. As natural in a factor pricing model, conditioning information is summarized by the conditional factor loadings, denoted by S_t (an $N \times P$ matrix). Throughout, we also refer to S_t as "characteristics" or "signals" in connection with the empirical asset pricing literature.[16] The P-vector of factor risk prices, denoted λ_F, together with the conditional loadings determine asset expected returns. We make the following assumption about the covariance structure of the signals S_t.

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*

[14] The assumption of constant conditional expected returns is without loss of generality and can be achieved by expanding the set of factors.

[15] 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.

[16] 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.

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* $\lim_{N\to\infty} \text{tr}(\Sigma^2)/(\text{tr}(\Sigma))^2 = \lim_{N\to\infty} \text{tr}((\Sigma_{\varepsilon,t}^{-1}\Sigma)^2)/(\text{tr}(\Sigma_{\varepsilon,t}^{-1}\Sigma))^2 = 0$.

By Assumption 2,

$$E[S_t'S_t] \;=\; \text{tr}(\Sigma)\Psi \;\in\; \mathbb{R}^{P\times P} \quad \text{and} \quad E[S_tS_t'] \;=\; \text{tr}(\Psi)\Sigma \in\; \mathbb{R}^{N\times N}. \tag{4}$$

While the matrix Ψ captures the covariance structure of signals across factors, Σ captures the covariance structure of signals across assets. The assumption of bounded $\text{tr}(\Sigma)$ is a no-arbitrage condition, ensuring that the predictable variation in returns stays bounded. The last two limits in Assumption 2 ensure that the Herfindahl indices of the eigenvalues of Σ and $\Sigma_{\varepsilon,t}^{-1}\Sigma$ converge to zero.[17] These assumptions guarantee that characteristics-based portfolios offer a sufficient amount of diversification across stocks.[18]

2.2 Neural Network Interpretation

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,[19]

$$R_{i,t+1} = \beta(X_{i,t})'G_{t+1} + u_{i,t+1}, \tag{6}$$

[17] For a matrix A with eigenvalues $\lambda_1, \cdots, \lambda_N$, we have $\text{tr}(A^2)/\,\text{tr}(A)^2 \;=\; \frac{\sum_i \lambda_i^2}{(\sum_i \lambda_i)^2}$.

[18] For example, suppose that $\text{rank}\Sigma = 1$, so that $\Sigma^{1/2} = \pi\pi'$ for some $\pi \in \mathbb{R}^N$. Then, $S_t = \pi\pi' X_t \Psi^{1/2}$ and therefore all factors are given by

$$F_{t+1} \;=\; \Psi^{1/2}X_t'\pi(\pi' R_{t+1})\,, \tag{5}$$

implying that all factor returns are proportional to returns on a single portfolio, $\pi' R_{t+1}$. Thus, there are no diversification benefits from constructing a portfolio of factors. The same happens when Σ has only a few large eigenvalues. Assumption 2 ensures that this pathological situation cannot occur.

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

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}. Absent 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

$$\beta(X_{i,t}) \approx \sum_{p=1}^{P} \xi_p S_{i,t,p} = \underbrace{\Xi}_{K\times P} \underbrace{S_{i,t}}_{P\times 1}, \tag{7}$$

where

$$S_{i,t} = A(\Omega X_{i,t}) = (A(\omega_p' X_{i,t}))_{p=1}^{P}. \tag{8}$$

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 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 (7) 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 (7) can accurately approximate the true conditional beta function under regularity conditions.[20]

[20]The approximating structure in (7) is analyzed by Gu et al. (2020a) and is a semi-nonparametric extension of the IPCA model in Kelly et al. (2020).

To tie this back to Assumption 1, we may stack assets' beta coefficients into an $N \times K$ matrix and substitute (7) into (6) to deliver

$$R_{t+1} \approx S_t F_{t+1} + u_{t+1}, \tag{9}$$

where $\tilde{F}_{t+1} = \Xi' G_{t+1}$ (and likewise $\lambda_F = \Xi' E_t[G_{t+1}]$). 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.

2.3 Efficient Portfolios and Characteristic-managed Portfolios

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

Proposition 1 *A conditional stochastic discount factor is*

$$\tilde{M}_{t+1} \;=\; 1 \;-\; \tilde{w}(S_t)' R_{t+1}\,, \tag{10}$$

where

$$\tilde{w}(S_t) \;=\; (S_t \Sigma_{F,t} S_t' + \Sigma_{\varepsilon,t})^{-1} S_t \lambda_F \tag{11}$$

is the conditional mean-variance efficient portfolio and

$$E_t[R_{i,t+1}\tilde{M}_{t+1}] \;=\; 0,\; i = 1, \cdots, N\,. \tag{12}$$

The SDF in (10) is stated as a portfolio of the basic risky assets, R. Estimating the conditional mean-variance portfolio of basic assets is extremely 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,[21]

$$F_{t+1} \;=\; S_t' R_{t+1}. \tag{13}$$

The hope is that by studying the *unconditional* properties of factors, we can learn about the conditional properties of asset markets. The conjecture underlying this approach is that, by interacting basic asset returns with conditioning characteristics, managed portfolios succinctly capture the conditional properties of asset returns. For example, Kozak et al. (2020) approximate the conditional SDF using the unconditional mean-variance efficient portfolio of managed portfolios.[22] Yet it is easy to see that the mean-variance portfolio of F_t,

$$\lambda = E[F_{t+1}F_{t+1}']^{-1}E[F_{t+1}] \tag{14}$$

is generally different from the conditionally efficient portfolio of basic assets that determines the true SDF in (10). This is particularly true for standard, low-dimensional factor models (e.g., when F_t is the vector of Fama-French factors).

We prove the surprising result that, in the high complexity (large P) setting, the unconditional optimal portfolio of factors and the true conditional SDF indeed coincide.

Proposition 2 (Unconditionally Optimal Portfolios of Factors Are Conditionally Optimal) *Suppose that in the limit, as $P \to \infty$, the vector of latent risk premia λ_F is uniformly bounded*

[21] 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.

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

and satisfies

$$\lambda_F' A \lambda_F \;\to\; 0 \tag{15}$$

in probability, for any symmetric, positive definite A with uniformly bounded trace.[23] *Suppose also that $\Sigma_{\varepsilon,t} = \sigma_t I$ for some $\sigma_t > 0$,*[24] *and let*

$$M_{t+1} = 1 - \lambda' F_{t+1} \;=\; 1 \;-\; w(S_t)' R_{t+1}, \;\; \textit{with } w(S_t) \;=\; \lambda' S_t \,, \tag{16}$$

be the factor approximation for the SDF with λ given by (14). Then, M_{t+1} converges to $\tilde{M}_{t+1}$ 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 $P \to \infty$.

This result is striking. It states that as long as the number of factors is large, factors are indeed sufficient to recover the conditionally efficient portfolio and, thus, the SDF. The condition $\lambda_F' A \lambda_F \;\to\; 0$ requires factor risk premia to be non-trivially distributed across factors. Note that, in the neural network interpretation above, this is essentially guaranteed by the fact that the true underlying factor premia are distributed across characteristic-managed portfolios through the Ξ matrix in (7). As a result of (15), the conditional covariance matrix $S_t \Sigma_{F,t} S_t' + \Sigma_\varepsilon$ of basic assets appearing in (11) drops out from the large P limit. What is left to be estimated is the covariance matrix of factors, which is an unconditional object and, thus, more tractable to estimate.

In summary, the key theoretical implication of Proposition 2 is that the complex setting allows us to characterize the SDF by describing the unconditional portfolio of factors and avoids the daunting problem of finding the conditionally optimal portfolio of basic assets. Just as continuous time limits conveniently simplify a range of asset pricing derivations, Proposition 2 shows the convenience of complexity for simplifying asset pricing

[23] For example, this is the case when $\lambda_F \in N(0, \Sigma_\lambda / P)$ for some bounded matrix Σ_λ. In this case, $E[\lambda_F' A \lambda_F] = \text{tr}(E[A \lambda_F \lambda_F']) = \text{tr}(A E[\lambda_F \lambda_F']) = \text{tr}(A \Sigma_\lambda)/P \leq \text{tr}(A) \|\Sigma_\lambda\| / P \to 0$.

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

model derivations by reducing conditional SDF estimation to an unconditional problem. Equivalently, managed portfolios efficiently incorporate all conditional information in the large P limit.[25] Therefore, in the remaining theoretical development, we leave behind the basic assets R_t and work directly with managed portfolios, F_t. Going forward, we refer to $M_{t+1} = 1 - \lambda' F_{t+1}$ in (16) as the true SDF and analyze estimators of λ when P is large.

3 Properties of Machine Learning SDF Models

This section derives the theoretical behavior of complex asset pricing models in the environment of Section 2. Throughout this section, we assume that the estimator is correctly specified in the sense that the factors used by the econometrician, F_t, represent the true and complete set of factors that enter linearly into the SDF, as in (16). On the one hand, the case of a correctly specified SDF estimator is unrealistic because an econometrician cannot know the true inputs to the SDF. However, the correctly specified setting is a useful starting point for understanding the basic properties of high-dimensional SDF estimators. This provides a foundation for our analysis of the more realistic and more interesting mis-specified setting in Section 4.

3.1 Ridge Estimation for a Complex SDF

Our analysis centers on the *ridge SDF estimator*, defined as

$$\hat{\lambda}(z) \;=\; \hat{\lambda}(z;P;T) \;=\; \left(zI + \hat{E}[F_t F_t']\right)^{-1} \hat{E}[F_t] = \arg\min_{\lambda} \left\{ \sum_{t=1}^{T} (1 - \lambda' F_t)^2 + z\|\lambda\|^2 \right\}, \quad (17)$$

[25] See, Appendix B for technical details.

where $\hat{E}[F_t] = \frac{1}{T}\sum_t F_t$ and $\hat{E}[F_t F_t'] = \frac{1}{T}\sum_t F_t F_t'$ are the sample mean and covariance of factors. The corresponding ridge portfolio return and SDF are

$$\hat{R}^M_{T+1}(z;P;T) = \hat{\lambda}(z)' F_{T+1}, \quad \hat{M}_{T+1}(z;P;T) = 1 - \hat{R}^M_{T+1}(z;P;T). \tag{18}$$

Britten-Jones (1999) points out that the population tangency portfolio of factors in (14) can be viewed as the coefficient in a time series regression: $\min_\lambda E[(1-\lambda' F_t)^2]$. Intuitively, this regression finds the combination of the risky assets F_t that behaves as closely as possible to a positive constant (in the ℓ_2 sense), which is tantamount to finding the portfolio with the highest Sharpe ratio. Kelly and Xiu (2023) dub this "maximum Sharpe ratio regression" (or MSRR) and discuss its attractiveness for incorporating machine learning methods into SDF estimation problems, such as the ridge estimator used here.

When $z = 0$ and $P \leq T$, expression (17) is the OLS estimator of the SDF and is the exact sample counterpart of (14):

$$\hat{\lambda}(0) = \hat{E}[F_t F_t']^{-1} \hat{E}[F_t]. \tag{19}$$

To equate the tangency portfolio of factors with the SDF per Proposition 2, we require a high complexity model in which $P \to \infty$. Yet when P is greater than the number of training observations T, $\hat{E}[F_t F_t']$ is ill-defined, and the OLS regression has an infinite number of solutions, all of which exactly fit the training data (delivering an excess return that is equal to one in all periods).

We overcome the deficiency of $\hat{E}[F_t F_t']$ by introducing a ridge penalty into the regression problem, shown in the second equality of (17). This augments the mean-variance efficient portfolio of factors by shrinking the sample covariance to the identity matrix in proportion

to the ridge parameter z, which has the effect of constraining the magnitude of λ.[26] When $z > 0$, the SDF solution $\hat{\lambda}(z)$ is unique and has a finite in-sample Sharpe ratio.

We may also interpret $\hat{\lambda}(z)$ in terms of "pricing errors" in the standard investor Euler equation. Since the factors are tradable assets and F_t is in the space of excess returns, a true SDF prices these with zero error due to the marginal investor's first-order optimality condition:

$$E[M_t F_t] = 0. \tag{20}$$

Or, by substituting (16), we see the standard notion of pricing errors as the divergence between expected factor returns and their riskiness:

$$E[F_t] - E[F_t F_t']\lambda = 0, \tag{21}$$

which is exactly zero for the true SDF parameters λ. In other words, the population regression solution (14) exactly prices all factors F_t. The deficiency of ordinary least squares regression when $P > T$ means an infinite number of SDF solutions price all factors with exactly zero error in-sample. The ridge regression in (17) coincides with a "regularized Euler equation" with the in-sample pricing errors proportional to factor risk prices:[27]

$$\hat{E}[M_t F_t] = z\hat{\lambda}(z). \tag{22}$$

Amid high complexity ($P > T$), ridge regularization provides a unique SDF solution whose in-sample pricing errors are non-zero.

A particularly interesting case of the ridge SDF that we reference throughout is the

[26] Such ridge shrinkage belongs to the family of spectral shrinkage estimators of the covariance matrix that only shrinks the empirical eigenvalues thereof. See, e.g., Kozak et al. (2020) and Ledoit and Wolf (2017) for other spectral shrinkage estimators.

[27] Indeed, $\hat{E}[M_t F_t] = (I - (zI + \hat{E}[FF'])^{-1}\hat{E}[FF'])\hat{E}[F_t] = z(zI + \hat{E}[FF'])^{-1})\hat{E}[F_t] = z\hat{\lambda}(z)$.

ridgeless SDF estimator, defined as

$$\hat{\lambda}(0^+) = \lim_{z\to 0^+} \left(zI + \hat{E}[F_t F_t']\right)^{-1} \hat{E}[F_t] = \hat{E}[F_t F_t']^+ \hat{E}[F_t], \tag{23}$$

where X^+ denotes the Moore-Penrose pseudo-inverse of X. The infinitesimal penalty $z \to 0$ means that the ridgeless SDF exactly fits the training data. Thus, it has an infinite in-sample Sharpe ratio and zero pricing errors in-sample. But among all least squares solutions that exactly fit the training data, $\hat{\lambda}(0^+)$ is the solution with the smallest magnitude (the smallest ℓ_2 norm). This property is important in understanding complex SDF behavior, as discussed below.

We also introduce the *infeasible ridge SDF estimator*

$$\lambda(z) = (zI + E[FF'])^{-1} E[F] \tag{24}$$

and its return and SDF,

$$R^M_{T+1}(z) \;=\; \lambda(z)' F_{T+1}, \quad M_{T+1}(z) = 1 - R^M_{T+1}(z). \tag{25}$$

The formula (24) is a useful population counterpart to the ridge SDF estimator $\hat{\lambda}(z)$. While $\hat{\lambda}(z)$ is a feasible portfolio that uses sample moments of the factors, the infeasible $\lambda(z)$ relies on the true means and covariances of factors. The special case $\lambda = \lambda(0)$ corresponds to the true SDF in (16). Naturally, as z increases from zero, the Sharpe ratio of $R^M_{T+1}(z)$ declines. The portfolio $\lambda(z)$ is an intermediate object between the true SDF $\lambda(0)$ and the feasible estimator $\hat{\lambda}(z)$ that will be useful for characterizing the properties of $\hat{\lambda}(z)$. Finally, we define

$$\mathcal{E}(z) \;\equiv\; E[R^M_{T+1}(z)] \;=\; E[F]'(zI + E[FF'])^{-1} E[F] \;\in\; (0,1)\,, \tag{26}$$

and note that, by a direct calculation,

$$\mathcal{V}(z) \equiv E[(R^M_{T+1}(z))^2] = \frac{d}{dz}(z\mathcal{E}(z)). \tag{27}$$

We will also need

$$\mathcal{R}(z) \equiv E[(1 - R^M_{T+1}(z))^2], \tag{28}$$

the mean squared deviation of $R^M_{T+1}(z)$ from one.

3.2 The Ridge SDF and Random Matrix Theory

To characterize large P behavior of $\hat{\lambda}(z)$, we use asymptotic analysis that allows the number of parameters P to grow with the number of observations T at a fixed rate ($P/T \to c > 0$). We refer to the ratio c as SDF *complexity*. In machine learning models, the number of parameters is typically large (often much larger than the number of observations), so $c \gg 0$. In this case, traditional large T asymptotic results such as the law of large number do not hold and therefore

$$\hat{\lambda}(z) = \left(zI + \hat{E}[F_t F_t']\right)^{-1} \hat{E}[F_t] \not\approx (zI + E[FF'])^{-1} E[F] = \lambda(z). \tag{29}$$

The central challenge to understanding the feasible SDF is the $P \times P$ matrix $\hat{E}[F_t F_t']$, whose dimension grows with the number of SDF parameters. Such analysis requires the apparatus of random matrix theory (RMT), on which we draw heavily to derive our results. This approximates the SDF estimator's behavior as we gradually increase the number of parameters holding the amount of data fixed.

The eigenvalue distribution of the factor population covariance matrix, $E[FF']$, determines the large P behavior of $\hat{E}[F_t F_t']$ via the Marčenko and Pastur (1967) theorem, a

cornerstone of RMT. A key technical insight in our analysis is that, by incorporating a ridge penalty in the SDF estimation problem per (17), we can connect $\hat{\lambda}(z)$ and the Marčenko and Pastur (1967) theorem to characterize the SDF estimator when the number of parameters is large. The eigenvalue distribution of $E[FF']$ in the large P limit is summarized by the Stieltjes transform

$$m(-z) \;=\; \lim_{P\to\infty} \frac{1}{P} \operatorname{tr}\left((E[F_t F_t'] + zI)^{-1}\right). \tag{30}$$

Because of its high dimensionality, $\hat{E}[F_t F_t']$ converges not to $E[FF']$ but to a distortion of it. The Marčenko and Pastur (1967) theorem describes the nature of this distortion by relating the limiting eigenvalue distribution of $\hat{E}[F_t F_t']$ to that of $E[FF']$. Theorem 10 in Appendix C shows that the limit

$$m(-z;c) \;=\; \lim_{P\to\infty,\ P/T\to c} \frac{1}{P} \operatorname{tr}\left(\left(zI + \hat{E}[F_t F_t']\right)^{-1}\right) \tag{31}$$

exists, is non-random, with $m(z;c)$ being the unique positive solution to the nonlinear master equation

$$m(z;c) = \frac{1}{1 \;-\; c \;-\; c\,z\,m(z;c)}\, m\left(\frac{z}{1 \;-\; c \;-\; c\,z\,m(z;c)}\right). \tag{32}$$

Equation (32) links $m(-z;c)$ in (31) to $m(-z)$ in (30). When complexity is small, $c \approx 0$, (32) implies $m(z;c) \approx m(z)$, as predicted by the standard law of large numbers. However, for $c > 0$, the link between the empirical Stieltjes transform and the "true" Stieltjes transform becomes very subtle. Formally, if one knows $m(z)$, one can compute $m(z;c)$ by solving the implicit algebraic equation (32). However, in reality, $m(z)$ is not observable, and we can only perform our inference based on (31). It is possible to show that, when complexity is high enough (e.g., $c > 1$), large parts of information about $m(z)$ are lost in finite samples,

and, hence $m(z)$ cannot be recovered from $m(z;c)$. A remarkable property of the theoretical expressions derived in this paper is that they only depend on the empirically observable $m(z;c)$. Thus, the knowledge of the true Stieltjes transform is not needed for understanding the properties of the SDF in the high complexity limit.

3.3 Expected Return of the Complex SDF

Our first result describes the expected return of the ridge SDF estimator in the high-complexity regime.

Theorem 3 *In the limit as* $P,\ T \to \infty,\ P/T \to c$, *the expected out-of-sample return of the ridge SDF satisfies*

$$\lim E\left[\hat{R}^M_{T+1}(z;P;T)\right] \ = \ \mathcal{E}(Z^*(z;c)), \tag{33}$$

where $\mathcal{A}_1$ *is defined in* (26) *and the function* $Z^*(z;c)$ *is the "equivalent infeasible shrinkage" given by*

$$Z^*(z;c) \ = \ z\,(1+\xi(z;c)) \ \in (z,\ z+c)\,, \tag{34}$$

and where

$$\xi(z;c) \ = \ \frac{c(1 \ - \ m(-z;c)z)}{1-c(1 \ - \ m(-z;c)z)}\,. \tag{35}$$

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

$$Z^*(z;c) \ \to \ \begin{cases} 0, & c<1 \\ 1/\tilde{m}(c), & c>1 \end{cases} \tag{36}$$

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

$$c - 1 \;=\; \frac{\int \frac{dH(x)}{\tilde{m}(1+\tilde{m}\,x)}}{\int \frac{x dH(x)}{1+\tilde{m}\,x}}, \tag{37}$$

and H *is the limiting eigenvalue distribution of* $E[F_t F_t']$.

Theorem 3 shows how complexity inhibits the performance of the SDF estimator. Intuitively, the large number of parameters 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. The fascinating aspect of Theorem 3 is that we can explicitly characterize the severity of limits to learning based on the eigenvalue distribution of the factor covariance matrix.

For a given choice of ridge parameter z, the cost of complexity can be described in terms of the infeasible ridge portfolio's return. At a complexity of zero, $Z^*(z;0) = z$, 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 $c > 0$ raises $Z^*(z,c)$ and drives down the expected return of the SDF estimator. By how much? By the same amount that the expected return drops when the infeasible portfolio's shrinkage rises from z to $Z^*(z,c)$. 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, we can characterize $Z^*(z;c)$ in closed form thanks to the expression (35) for $\xi(z;c)$ from RMT.

The monotonicity of $Z^*(z;c)$ 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 $Z^*(z;c)$ is uniformly bounded away from zero in the high complexity regime (when $c > 1$).

3.3.1 Variance of the Complex SDF

Next, we analyze the role of complexity in determining the variance of the ridge SDF estimator.

Theorem 4 *In the limit as* $P,\ T \to \infty,\ P/T \to c$, *the expected out-of-sample second moment of the return of the ridge SDF satisfies*

$$\lim E[(\hat{R}^M_{T+1}(z;P;T))^2] \ = \ \underbrace{\mathcal{V}(Z^*(z;c))}_{implicit\ \ shrinkage} \ + \ \underbrace{G(z;c)\,\mathcal{R}(Z^*(z;c))}_{complexity\ \ risk}, \tag{38}$$

where $\mathcal{V}(z),\ \mathcal{R}(z)$ *are defined in* (27), (28), *and*

$$G(z;c) \ = \ \frac{d}{dz}(z\xi(z;c)) \ \in (0, cz^{-2}] \tag{39}$$

is monotone decreasing in z *and increasing in* c. *Furthermore,*

$$G(z;c) \ = \ \mathcal{M}(z;Z_*(z;c))\,, \tag{40}$$

where

$$\mathcal{M}(z;Z) \ = \ -1 \ + \ \frac{Z}{z + c\phi(Z)Z^2},\ \phi(z) \ = \ P^{-1}\,\mathrm{tr}(E[FF'](zI + E[FF'])^{-2})\,. \tag{41}$$

The variance of the complex ridge SDF is characterized by two terms on the right side of (38). The first term mirrors the behavior of the mean in Theorem 3. In the large P limit, the SDF based on ridge parameter z has the same volatility as the infeasible portfolio with a larger ridge parameter $Z^*(z;c) > z$. Rising complexity (holding z fixed) raises the infeasible ridge portfolio's effective complexity $Z^*(z;c)$. In the high complexity regime ($c > 1$), SDF volatility is therefore decreasing in complexity. In other words, higher complexity imposes additional *implicit shrinkage* of the ridge SDF, above and beyond the explicit shrinkage z.

Through this regularization, complexity reduces SDF variance and may improve the risk-return tradeoff. 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 λ vector further.

While the rationale for the first term in (38) is qualitatively similar to Theorem 3, 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 governed by the function $G(z;c)$ that is independent of expected factor returns and only depends on the eigenvalue distribution of $E[FF']$. When $c = 0$, there are infinitely more observations than parameter parameters, so the SDF estimator $\hat{\lambda}(z)$ converges to a non-random limit. As a result, $G(z;0) = 0$, and there is no complexity risk. But 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 complexity-induced limits to learning.

3.3.2 Sharpe Ratio

Combining the preceding results, we can characterize the complex SDF's limiting Sharpe ratio. As usual, we use $\text{Var}[X] = E[X^2] - E[X]^2$ to denote the variance of a random variable.

Theorem 5 *Let*

$$SR(z;c) = \lim_{P,T\to\infty, P/T\to c} \frac{E[\hat{R}^M_{T+1}(z;P;T)]}{\text{Var}[(\hat{R}^M_{T+1}(z;P;T))^2]^{1/2}}. \tag{42}$$

Then,

$$\frac{1}{SR^2(z;c)} \;=\; (1+G(z;c))\frac{1}{SR^2(Z_*(z;c);0)} + \; G(z;c)\left(\frac{1-\mathcal{E}(Z_*(z;c))}{\mathcal{E}(Z_*(z;c))}\right)^2 \tag{43}$$

Theorem 5 implies that complexity always creates a gap between the feasible and infeasible Sharpe ratios due to limits to learning. Since the infeasible Sharpe ratio, $SR(z;0)$, is monotone decreasing in z, Theorem 5 shows how complexity affects the SDF Sharpe ratio by both reducing out-of-sample SDF expected returns (through implicit regularization) and raising out-of-sample SDF volatility (by introducing complexity risk).

The formula (43) implies a lower bound for limits to learning that depends only on the eigenvalue distribution of the covariance matrix $E[FF']$:

$$SR^2(z;c) \;\leq\; \frac{SR^2(Z_*(z;c);0)}{1+G(z;c)} \;\leq\; \frac{SR^2(z;0)}{1+G(z;c)}\,. \tag{44}$$

Thus, no matter how big the expected returns on the factors are, the Sharpe ratio will drop by a factor of at least $\frac{1}{1+G(z;c)}$.

3.4 Complex SDF Pricing Errors

The fourth and last performance metric that we study quantifies the magnitude of pricing errors for the SDF. In particular, we study the HJD (Hansen and Jagannathan, 1997) that aggregates squared pricing errors of test assets weighted by the test assets' inverse covariance matrix.

The distinction between in-sample and out-of-sample performance is an essential ingredient in the analysis of machine learning models (see, e.g. Martin and Nagel, 2021, for a related discussion). Likewise, 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 $t \in (T+1, T+T_{OOS}]$, where T_{OOS} is the number of out-of-sample periods.

Our analysis relies on the quantities

$$\bar{F}_{OOS} \;=\; E_{OOS}[F] \;\in\; \mathbb{R}^P,\; B_{OOS} \;=\; E_{OOS}[FF'] \;\in\; \mathbb{R}^{P\times P} \tag{45}$$

where $E_{OOS}[X] \;=\; \frac{1}{T_{OOS}} \sum_{t\in(T,T+T_{OOS}]} X_t$ 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

$$\mathcal{E}_{OOS}(z;P;T) \;=\; \frac{1}{T_{OOS}} \sum_{t\in(T,T+T_{OOS}]} F_t \hat{M}_t(z;P;T) \;\in\; \mathbb{R}^P\,. \tag{46}$$

Finally, following Hansen and Jagannathan (1997), we define the out-of-sample HJD as[28]

$$\mathcal{D}^{HJ}_{OOS}(z;P;T) \;=\; \mathcal{E}_{OOS}(z;P;T)'\, B^{+}_{OOS}\, \mathcal{E}_{OOS}(z;P;T)\,, \tag{47}$$

where B^{+}_{OOS} is the Moore-Penrose quasi-inverse of the potentially degenerate matrix B_{OOS}.

The following is true.

Proposition 6 *We have*

$$\mathcal{D}^{HJ}_{OOS}(z;P;T) \;-\; \bar{F}'_{OOS} B^{+}_{OOS} \bar{F}_{OOS} \;=\; -2E_{OOS}[\hat{R}^M_t(z;P;T)] \;+\; E_{OOS}[(\hat{R}^M_t(z;P;T))^2] \tag{48}$$

[28] 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 $\mathcal{D}^{HJ}_{OOS}(z;P;T)$ and the out-of-sample SDF Sharpe ratio.

When $P > T_{OOS}$ and both are sufficiently large, we have

$$\bar{F}'_{OOS} B^{+}_{OOS} \bar{F}_{OOS} \;\approx\; 1 \tag{49}$$

and hence

$$\mathcal{D}^{HJ}_{OOS}(z;P;T) \;\approx\; E_{OOS}[(1-\hat{M}_t(z;P;T))^2]\,. \tag{50}$$

In expectation, we have

$$\lim_{P,T,T_{OOS}\to\infty,\ P/T\to c,\ P>T_{OOS}} E[\mathcal{D}^{HJ}_{OOS}(z;P;T)] \;=\; (1+G(z;c))\mathcal{R}(Z^*(z;c)), \tag{51}$$

Proposition 6 shows a surprising identity for expected out-of-sample pricing errors. The high complexity error $\mathcal{D}^{HJ}_{OOS}(z;c)$ 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.

Perhaps surprisingly, the out-of-sample pricing error (51) does not always converge to zero even when $c = z = 0$. Instead,

$$\lim_{P,T,T_{OOS}\to\infty,\ P/T\to 0} E[\mathcal{D}^{HJ}_{OOS}(0;P;T)] \;\to\; \lim E[\bar{F}'_{OOS} B^{+}_{OOS} \bar{F}_{OOS}] - E[F]E[FF']^{-1}E[F]\,. \tag{52}$$

Note that $E[F]E[FF']^{-1}E[F] = \mathcal{E}(0)$ is the expected return on the efficient portfolio, whereas $E[\bar{F}'_{OOS} B^{+}_{OOS} \bar{F}_{OOS}]$ can be computed based on the following result.

Lemma 1 *We have*

$$\lim E[\bar{F}'_{OOS}(zI + B_{OOS})^{-1}\bar{F}_{OOS}] \;=\; \frac{\mathcal{E}(Z^*(z;c_{OOS})) \;+\; \xi(z;c)}{1+\xi(z;c)}\,. \tag{53}$$

In the ridgeless limit,

$$\lim E[\bar{F}'_{OOS} B^{+}_{OOS} \bar{F}_{OOS}] \;=\; \begin{cases} \mathcal{E}(0)(1-c_{OOS}) \;+\; c_{OOS}, & c_{OOS} < 1 \\ 1, & c_{OOS} \geq 1\,, \end{cases} \tag{54}$$

and, hence,

$$\lim_{P,T,T_{OOS}\to\infty,\; P/T\to 0} E[\mathcal{D}^{HJ}_{OOS}(0;P;T)] \;\to\; \begin{cases} c_{OOS}(1-\mathcal{E}(0)), & c_{OOS} < 1 \\ 1-\mathcal{E}(0), & c_{OOS} \geq 1\,. \end{cases} \tag{55}$$

The pricing error (55) remains strictly positive as long as $c_{OOS} > 0$. 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 $c_{OOS} > 0$, the out-of-sample factor moments ($\bar{F}_{OOS},\ B_{OOS}$) are so severely misestimated that $\mathcal{D}^{HJ}_{OOS}(0;P;T)$ 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

$$\hat{M}_t(z;P;T) \;=\; 1 - \alpha\, \hat{R}^M_t(z;P;T). \tag{56}$$

Then (50) implies that the optimal α is $\alpha = E_{OOS}[\hat{R}^M_t(z;P;T)]/E_{OOS}[(\hat{R}^M_t(z;P;T))^2]$, and we get

$$\mathcal{D}^{HJ}_{OOS}(z;P;T) = \bar{F}'_{OOS} B^{+}_{OOS} \bar{F}_{OOS} - SR^2_{OOS}(\hat{R}^M(z;P;T)). \tag{57}$$

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.

4 Mis-specified SDF Models and the Virtue of Complexity

So far, we have implicitly assumed a correctly specified model for the SDF. KMZ point out that model complexity comparative statics are of limited use in the theoretical setting of correctly specified models because, as model complexity varies, the complexity of both the empirical model *and the true model* are changing. As a result, comparative statics involving model complexity cannot be taken to the data.

Instead, they argue that the more interesting theoretical case to consider is one that varies the complexity of a mis-specified empirical model while holding the true DGP fixed. The practical motivation for this approach is that, as the empirical machine learning model becomes more complex, its ability to approximate the truth improves, and the degree of mis-specification lessens.

In this section, we develop the theory for this more realistic mis-specified environment. In particular, we assume only a fraction $q = \frac{P_1}{P} < 1$ of factors are observable. The subset of factors, $F_{t+1}(q) = (F_{i,t+1})_{i=1}^{P_1}$, has a covariance matrix $E[F_t(q)F_t(q)'] \in \mathbb{R}^{P_1 \times P_1}$. As in KMZ, we consider a case where the true number of factors P is large, and the ordering of factors is irrelevant. We are interested in characterizing SDF behavior as P_1 varies and approaches P from below (when $P_1 = P$, the model is correctly specified).

Our analysis of the estimator and its theoretical properties mirrors that of the correctly specified estimator in Section 3. In particular, the complex SDF parameter estimates are

$$\hat{\lambda}(z; P_1; T) \;=\; (zI + \hat{E}[F_t(q)F_t(q)'])^{-1}\hat{E}[F_t(q)] \tag{58}$$

with corresponding portfolio return and SDF of

$$\hat{R}^M_{T+1}(z;P_1;T) \;=\; \hat{\lambda}(z;P_1;T)'F_{T+1}(q), \quad \hat{M}_{T+1}(z;P_1;T) = 1 - \hat{R}^M_{T+1}(z;P_1;T).$$

We also define the analog of the infeasible ridge SDF estimator (24) for the mis-specified case:

$$\lambda(z;q) = (zI + E[F(q)F(q)'])^{-1}E[F(q)] \tag{59}$$

and its return and SDF,

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

From the infeasible estimator, we define the analogs of functions (26) through (28):

$$\begin{aligned}
\mathcal{E}(z;q) \;&=\; E[R^M_{T+1}(z;q)] \;=\; E[F(q)]'(zI + E[F(q)F(q)'])^{-1}E[F(q)] \\
\mathcal{V}(z;q) \;&=\; E[(R^M_{T+1}(z;q))^2] \;=\; \frac{d}{dz}(z\mathcal{E}(z;q)) \\
\mathcal{R}(z;q) \;&=\; E[(1 - R^M_{T+1}(z;q))^2]\,.
\end{aligned} \tag{60}$$

Next, we require analogues of the m, Z^*, ξ, and G functions that depend on the degree of mis-specification, q:

$$\begin{aligned}
m(-z;cq;q) \;&=\; \lim_{P_1\to\infty,\ P_1/T\to cq} \frac{1}{P_1}\operatorname{tr}\left(\left(zI + \hat{E}[F_t(q)F_t(q)']\right)^{-1}\right) \\
\xi(z;cq) \;&=\; \frac{cq(1 \;-\; m(-z;cq;q)z)}{1 - cq(1 \;-\; m(-z;cq;q)z)} \\
Z^*(z;cq;q) \;&=\; z\,(1 + \xi(z;cq;q) \in (z, z + cq) \\
G(z;cq;q) \;&=\; (z\xi(z;cq;q))' \;\in (0, cqz^{-2}].
\end{aligned} \tag{61}$$

The complexity of the mis-specified model is $cq = P_1/T$. As q increases, so does cq, and, hence, the impact of complexity on portfolio performance is also determined by $cq \in (0, c)$.

Finally, we define the out-of-sample pricing errors and HJD under mis-specification as

$$\begin{aligned}
\mathcal{E}_{OOS}(z;q;P;T) &= \frac{1}{T_{OOS}} \sum_{t\in(T,T+T_{OOS}]} F_t \hat{M}_t(z;q) \in \mathbb{R}^P \\
\mathcal{D}^{HJ}_{OOS}(z;q;P;T) &= \mathcal{E}_{OOS}(z;q;P;T)' \, B^+_{OOS} \, \mathcal{E}_{OOS}(z;q;P;T) \, .
\end{aligned}$$

Note that $\mathcal{E}_{OOS}(z;q;P;T) \in \mathbb{R}^P$ and $B^+_{OOS} \in \mathbb{R}^P$ while $F_t(q) \in \mathbb{R}^{P_1}$. In other words, the mis-specified SDF attempts to price all P factors using only a subset of P_1 factors.

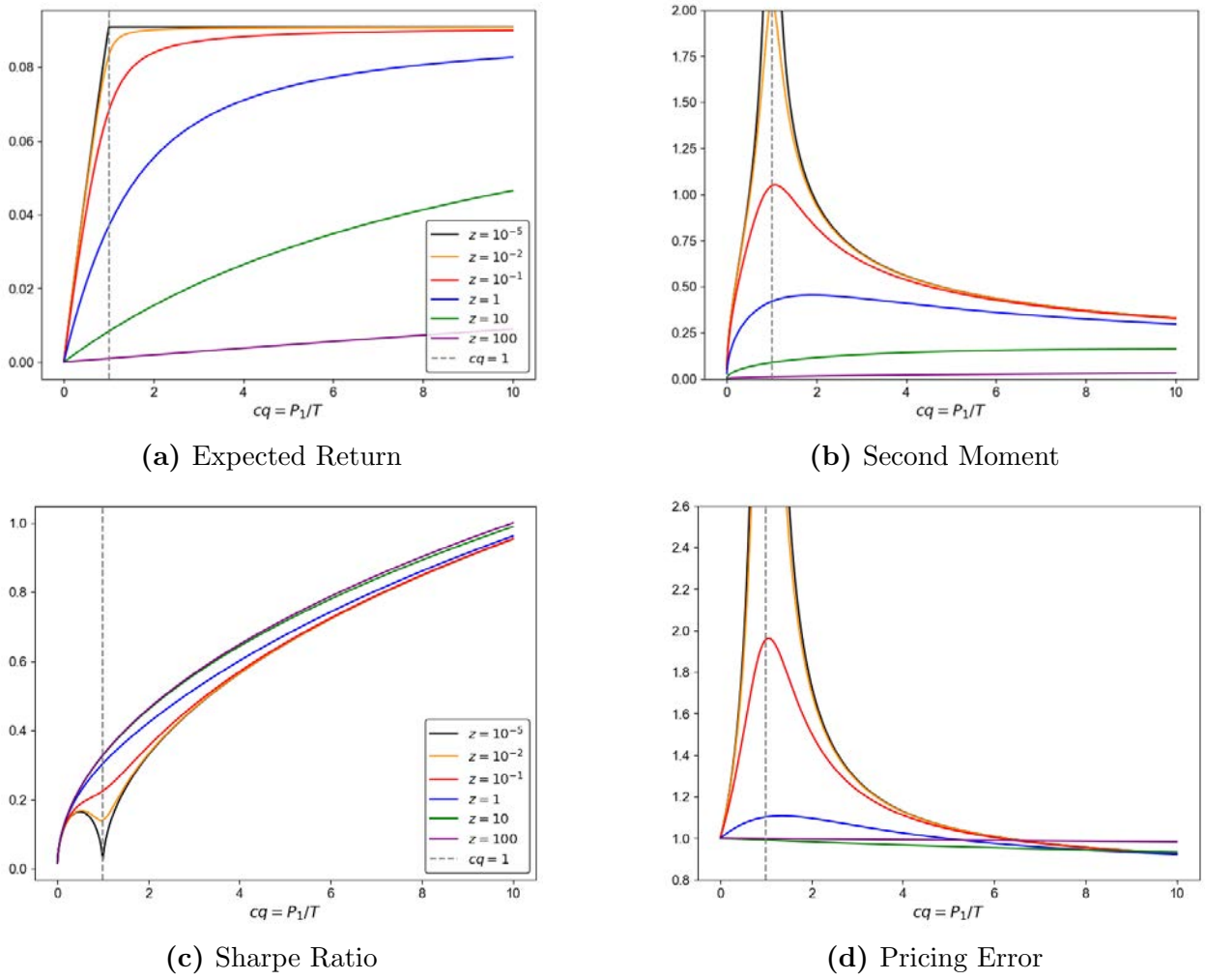
We can now state our main theoretical results for the mis-specified case.

Theorem 7 *In the limit as $P_1,\ T \to \infty,\ P_1/T \to cq$, the expected out-of-sample moments of the ridge SDF portfolio satisfy*

$$\begin{aligned}
&i.\ \lim E[\hat{R}^M_{T+1}(z;P_1;T)] = \mathcal{E}(Z^*(z;cq;q);q) \\
&ii.\ \lim E[(\hat{R}^M_{T+1}(z;P_1;T))^2] = \mathcal{V}(Z^*(z;cq;q);q) + G(z;cq;q)\,\mathcal{R}(Z^*(z;cq;q);q) \\
&iii.\ \frac{1}{SR^2(z;cq;q)} = \frac{1}{\mathcal{S}(z;Z_*(z;cq;q);q)},\ \textit{where} \\
&\quad \frac{1}{\mathcal{S}(z;Z;q)} = (1+\mathcal{M}(z;Z,q))\frac{1}{SR^2(Z;0;q)} + \mathcal{M}(z;Z,q)\left(\frac{1-\mathcal{E}(Z;q)}{\mathcal{E}(Z;q)}\right)^2 \\
&iv.\ \lim E[\mathcal{D}^{HJ}_{OOS}(z;q;P;T)] = -(1-\mathcal{E}(0;1))\max(1-c_{OOS},0) + \mathcal{D}(z;Z_*(z;cq;q);q),\ \textit{where} \\
&\quad \mathcal{D}(z;Z;q) = (1+\mathcal{M}(z;Z,q))\mathcal{R}(Z;q).
\end{aligned} \tag{62}$$

4.1 The Virtue of Complexity

Theorem 7 is the foundation for understanding the virtue of complexity in SDF models. In this section, we attempt to draw out the intuition behind the theorem to understand the behavior of complex SDFs. To aid our discussion, Figure 1 plots the "VoC curves"



(a) Expected Return

(b) Second Moment

(c) Sharpe Ratio

(d) Pricing Error

Figure 1: Expected Out-of-sample Performance of Mis-specified Complex SDF Models

Note. Limiting out-of-sample mean, variance, Sharpe ratio, and pricing error (HJD) of the SDF as a function of c, q and z from Theorem 7 assuming Ψ is the identity matrix and $\lambda \sim N(0, I)$.

to illustrate the effect of model complexity on SDF behavior. Each curve corresponds to a different choice of ridge penalty z, and each point on a curve corresponds to a different number of model parameters P_1. VoC curves show how the complexity of the empirical model (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. In Figure 1, these moments are calculated from the random matrix theoretical limits in Theorem 7 with calibration

choices of $\Psi = I$ and $\lambda \sim N(0, I)$. We set the true complexity $c = P/T$ equals 10, and we gradually increase the fraction of observable factors $q = P_1/P$.

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. This is true for all ridge penalty levels and throughout the full range of complexity. Expected return curves are flatter with higher 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), SDF volatility is highly sensitive to model complexity, a point critical to understanding the broader behavior of complex SDF models. When the complexity of the empirical model (cq) approaches unity, the variance of the ridgeless SDF spikes. The logic for this behavior follows from arguments in KMZ. For $cq \to 1$, the unregularized sample covariance matrix of factors becomes unstable, and because $\hat{\lambda}(0)$ relies on the unregularized inverse of this covariance, the estimator's variance explodes. Intuitively, when $cq = 1$, the number of model parameters equals the number of time series observations, so there is a unique estimator that fits every data point with zero error. Without regularization, this estimator is badly overfitting and produces disastrous out-of-sample behavior.

When $cq \gg 1$ the ridgeless estimator $\lim_{z\to 0} \hat{\lambda}(z)$ is the unique estimator that exactly fits the training data while maintaining the smallest ℓ_2 norm. Thus, the ridgeless estimator implicitly regularizes the SDF, leading to low SDF volatility in the high-complexity regime. 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.[29] At low complexity ($cq \ll 1$), the Sharpe ratio rises with complexity as larger models show improved approximation power benefits. But near $cq = 1$, the Sharpe ratio collapses to zero due to the explosion in SDF variance. Finally, at high complexity ($cq \gg 1$), 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.[30]

Finally, Panel (d) illustrates the behavior of out-of-sample SDF pricing errors (HJD) as a function of complexity. As shown by the inverse association between the HJD and Sharpe ratio in (57), the pricing errors in Panel (d) mirror the patterns for the Sharpe ratio in Panel (c). The more complex the SDF, the better its ability to price all factors F_t 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 that 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 (we cannot learn parameters accurately due to the dearth of data), and as a result, out-of-sample pricing errors are bounded away from zero (this is true even for high complexity models that are *correctly* specified).

A fascinating consequence of the pricing error result in part (iv) of Theorem 7 is the fact that *out-of-sample pricing errors are independent of the set of test assets* when $c_{OOS} > 1$. As an illustration, consider an SDF model that only involves P_1 factors $F_1, \cdots, F_{P_1}$,

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

[30] Another way to understand how complexity benefits the SDF Sharpe ratio is by examining the estimation problem directly:

$$\min_{\lambda \in \mathbb{R}^{P_1}} \{E[(1 - \lambda' F_t(q))^2] + z\|\lambda\|^2\} \;=\; 1 \;-\; \mathcal{E}(z;q)\,.$$

As $P_1 = qP$ increases, the objective is optimized over a larger subset of factors. Hence, the expected SDF return, $\mathcal{E}(z;q)$, is monotone increasing in q. When z is small, the effect of penalization is negligible, and $SR(z;0;q) = E[R^M_{T+1}(z;q)]/(E[(R^M_{T+1}(z;q))^2]^{1/2}$ is also monotone increasing, reflecting the larger diversification possibilities when more factors are used.

and let $F_1, \cdots, F_{P_1}, F_{P_1+1}, \cdots, F_P$ be the set of test assets, containing $F_1, \cdots, F_{P_1}$.[31] As is emphasized in the literature (e.g. Gagliardini and Ronchetti, 2020), computing the conditional HJD requires using all possible managed portfolios in the set of test assets. Theorem 7 implies that complexity imposes an inherent limit on what may be inferred with a limited amount of out-of-sample test data (aka *"limits-to-testing"*). Effectively, increasing the number of priced assets beyond the number of test periods has no impact on the pricing errors. When $c_{OOS} > 1$, it does not even matter which assets we try to price. The pricing errors only depend on the properties of the SDF itself (i.e., properties of $F_1, \cdots, F_{P_1}$) and not the properties of the test asset set!

To best illustrate the virtue of SDF complexity, the plots in Figure 1 are calculated from our random matrix theory derivations in a specific calibration. However, the virtue of SDF complexity holds more generally, as stated in the next result.

Theorem 8 (The Virtue of Complexity) *Suppose that $\frac{\partial}{\partial q}\mathcal{S}(z;Z;q)$ is sufficiently large relative to $\frac{\partial}{\partial Z}\mathcal{S}(z;Z;q)$. Then, the OOS Sharpe ratio is monotone increasing in q. Next, suppose that $\frac{\partial}{\partial q}\mathcal{D}(z;Z;q)$ is sufficiently large relative to $\frac{\partial}{\partial Z}\mathcal{D}(z;Z;q)$. Then, HJD is monotone decreasing in q.*

To understand the virtue of complexity, consider the formula

$$\lim E[\mathcal{D}^{HJ}_{OOS}(z;q;P;T)] \;=\; \mathcal{D}(z;Z^*(z;cq);q), \tag{63}$$

where, by Theorem 7,

$$\mathcal{D}(z;Z;q) \;=\; (1+\mathcal{M}(z;Z;q))\mathcal{R}(Z;q)\,. \tag{64}$$

Complexity defines a tradeoff between the ability of the model to approximate the truth and the estimation risk due to complexity. A larger model leads to better-diversified factor

[31] Formally, we just need to make sure that all $F_1, \cdots, F_{P_1}$ belong to the span of the test assets.

portfolios, and, as a result, $\mathcal{R}(Z;q)$ is monotone, decreasing in q. At the same time, larger q leads to stronger implicit regularization, pushing Z_* up and, hence, higher $\mathcal{R}(Z_*;q)$. A similar mechanism pushes $\mathcal{M}(z;Z;q)$ up because a larger q increases complexity risk. The tradeoff is then determined by

$$\frac{d}{dq}\mathcal{D}(z;Z_*(z;cq;q);q) \;=\; \frac{\partial}{\partial Z}\mathcal{D}(z;Z_*(z;cq;q);q)\frac{d}{dq}Z_*(z;cq;q) + \frac{\partial}{\partial q}\mathcal{D}(z;Z_*(z;cq;q);q) \quad (65)$$

When the marginal diversification benefit $\frac{\partial}{\partial q}\mathcal{D} > 0$ is large enough relative to the loss due to implicit regularization, $\frac{\partial}{\partial Z}\mathcal{D}$, we obtain the virtue of complexity.[32]

5 Empirics

In this section, we empirically investigate the effect of model complexity on out-of-sample SDF behavior. We develop direct empirical analogs to the theoretical comparative statics for mis-specified models in Section 4.

5.1 Data

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 thus focus our analysis on a standard empirical problem in asset pricing: estimating the SDF from US stock returns

[32] As an illustration, consider the simple case of uncorrelated factors so that $\mathrm{Cov}(F) = I$ and suppose that risk premia are uniformly distributed across factors. Then,

$$\begin{aligned}
\mathcal{E}(z;q) \;&=\; \frac{q}{1+z+q}\\
\mathcal{V}(z;q) \;&=\; \frac{q+q^2}{(1+z+q)^2}\\
\mathcal{R}(z;q) \;&=\; \frac{1+z^2+q}{(1+z+q)^2}\\
1+\mathcal{M}(z;Z;q) \;&=\; \frac{Z}{z+cq\frac{Z^2}{(1+Z)^2}},
\end{aligned} \quad (66)$$

while Z_* can be computed using RMT.

at monthly frequency (see Lettau and Pelger (2020); Kozak et al. (2020)). In addition to stock returns, we require data on the conditioning variables, X_t, that determine conditional stock-level weights in the SDF portfolio, as in equation (1). For this, we use the data set constructed in Jensen et al. (Forthcoming) (JKP henceforth), which is a comprehensive and standardized collection of stock-level return predictors from the finance literature.[33] It includes monthly observations of 153 characteristics for each stock from 1963 to 2019. The JKP universe includes NYSE/AMEX/NASDAQ securities with CRSP share code 10, 11, or 12, excluding "nano" stock as classified by JKP (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 130 characteristics with the fewest missing values.[34]. 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 a panel of characteristics $X_t = (X_{i,k,t})_{i,k} \in \mathbb{R}^{N_t \times d}$.

5.2 Empirical Design

Following our theoretical development, our empirical analysis pursues an SDF of the form

$$M_{t+1} = 1 - \lambda F_{t+1} = 1 - \lambda' S_t' R_{t+1} \approx 1 - w(X_t)' R_{t+1}$$

[33]The JKP stock-level data are accessible at https://wrds-www.wharton.upenn.edu/pages/get-data/contributed-data-forms/global-factor-data/.

[34]To mitigate concerns that our empirical findings are dependent on high-turnover characteristics, we rerun our main experiments while further restricting the predictor set to $d = 110$ by removing the 20 characteristics with the highest turnover. This helps alleviate concerns that the complexity effects that we document are artifacts of limits-to-arbitrage arising from trading costs (see the related critiques of Jensen et al., 2022; Avramov et al., 2023) We present the results in Appendix M.

as in equations (10) and (16). The factors $F_{t+1} = S_t' R_{t+1}$ are a set of P characteristic-managed portfolios, one for each of the P characteristics in the $N \times P$ matrix S_t.

The genesis of S_t is critical for linking the empirical analysis to our theory. We define S_t as nonlinear basis functions of the raw predictors X_t such that $S_t\lambda$ is a generic nonparametric approximator of $w(X_t)$. To evaluate the complexity comparative statics implied by our theory, we require a framework that allows us to smoothly transition from low complexity ($P \ll T$) to high complexity ($P \gg T$) models holding the underlying information set fixed. Following KMZ, we accomplish this using the machine learning method of random Fourier features, or RFF (Rahimi and Recht, 2007). RFF converts the 130 original signals X_t into a pair of new signals

$$S_{i,t} \in \mathbb{R}^{N_t \times 2} = [\sin(\gamma X_t \omega_i), \cos(\gamma X_t \omega_i)]', \quad \omega_i \sim \text{i.i.d. } N(0, I). \tag{67}$$

$S_{i,t}$ is a random linear combination (ω_i) of the raw characteristics X_t fed through the trigonometric activation functions. RFF is an ideal tool for our analysis because it uses a fixed set of input data, X_t, to create a set of features with any desired dimension P. For a low-dimensional model of, say, $P = 2$, one generates a single pair of RFFs. For a very high-dimensional model of, say $P = 10{,}000$, one can instead draw many random weight vectors ω_i, $i = 1, ..., 5{,}000$. The larger the number of random features, the richer the approximation $S_t\lambda$ provides to the true SDF weight function $w(X_t)$. The RFF approach is a wide two-layer neural network with fixed weights in the first layer (in the form of ω_i) and optimized weights in the second layer (in the form of least squares estimates of λ).[35]

We construct empirical VoC curves by varying the dimension of S_t from $P = 1, ..., 1{,}000{,}000$ and considering a grid of ridge penalties $z \in \{10^n \mid n \in \{-12, \dots, 3\}\}$. For

[35]The parameter γ controls the Gaussian kernel bandwidth in generating random Fourier features. Following Kelly et al. (2022), we randomly choose 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. For each nonlinear feature that we generate, we again cross-sectionally rank-standardize it to lie in the [-0.5,0.5] interval. Finally, because the size of the cross-section varies over time, our empirical analysis uses $F_{t+1} = \frac{1}{N_t^{1/2}} R_{t+1}' S_t \in R^P$.

each model of size P with ridge penalty z, we conduct a rolling out-of-sample SDF model performance analysis. Starting in January 1993, on each date t we use the most recent 360 months of data to estimate the ridge SDF in (17).[36] 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 then calculate the realized out-of-sample SDF expected return, variance, Sharpe ratio, and HJD.[37]

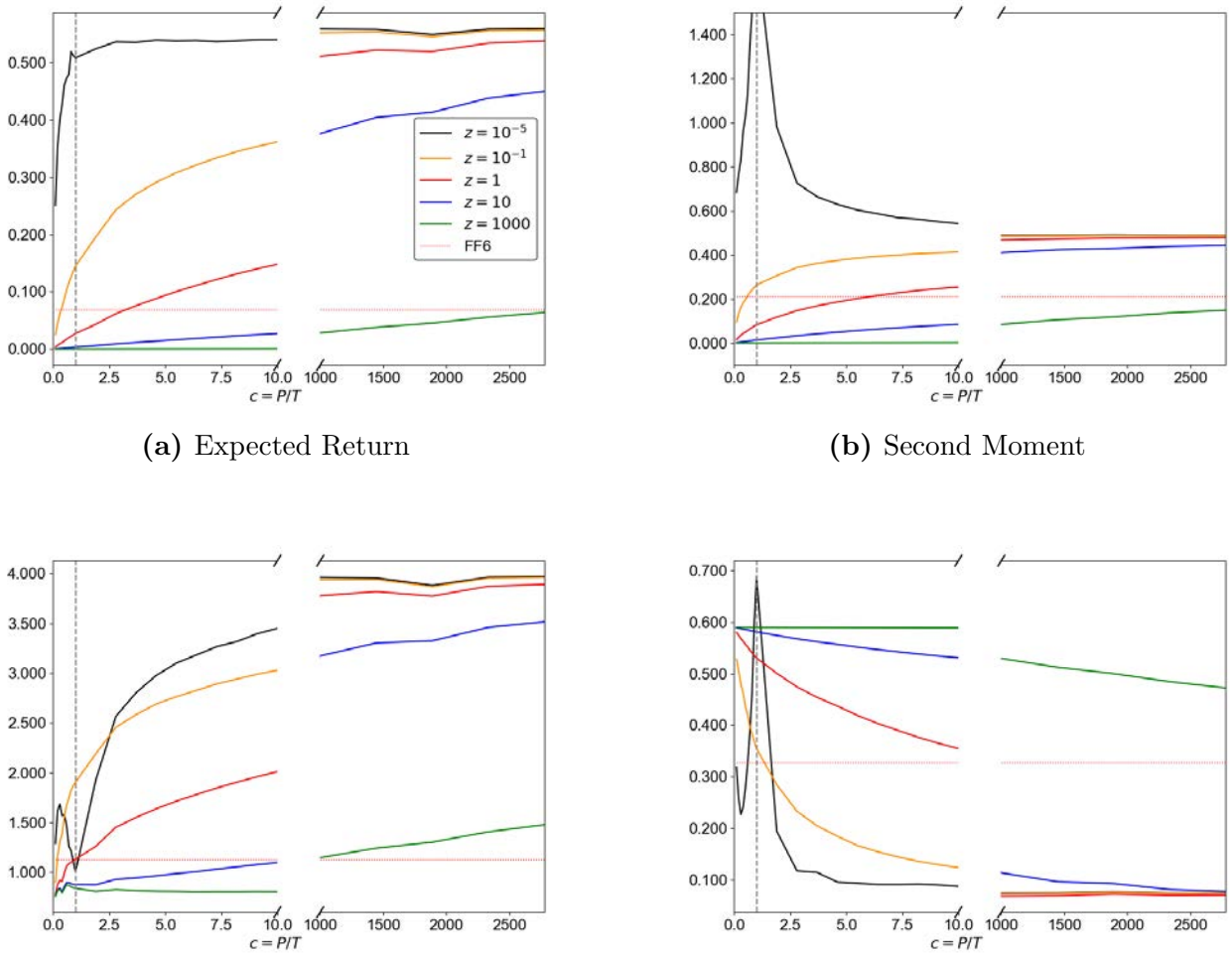
5.3 Main Results

Our main analysis uses the full JKP data set described above. We plot the empirical VoC curves in Figure 2. The first and central conclusion of our analysis is that the data demonstrates the virtue of complexity predicted by out theory. As we increase the number of factors in our empirical SDF model we find that i) the average SDF return rises, ii) SDF volatility spikes as P approaches T and decreases monotonically thereafter, iii) the SDF Sharpe ratio exhibits double ascent for low ridge penalties and permanent ascent for higher penalties, and iv) pricing error patterns are the inverse of those for Sharpe ratio and are generally decreasing with complexity. In short, increasing the number of factors enhances the out-of-sample performance of factor SDF models.

The second conclusion from Figure 2 is that complexity benefits are large in magnitude. Low complexity SDF models deliver Sharpe ratios on the order of 0.5 to 1.5, while the highest complexity SDFs that we consider achieve Sharpe ratios near 4.0. Likewise, complex SDFs are much better situated to price our demanding set of test assets (one million

[36]The stochastic nature of RFF means that there is inherent variability in the estimated SDF model, particularly when P is small. To mitigate this variability, we repeat the RFF-based estimation 20 times and ensemble the SDF parameter estimates in an equally weighted average. See KMZ for further discussion of this point.

[37]When calculating the HJD for an SDF model of dimension P, we use the entire set of 1,000,000 factors as test assets. So, when $P = 1{,}000{,}000$, the set of factors underlying the SDF exactly coincides with the set of test assets.



(a) Expected Return

(b) Second Moment

(c) Sharpe Ratio

(d) Pricing Error

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

Note. Realized out-of-sample SDF portfolio average return, second moment, Sharpe ratio, and pricing error (HJD). The horizontal axis shows model complexity $c = P/T$, with P ranging from 2 to 1,000,000 and $T = 360$ months. Factors underlying the SDF are characteristic-managed portfolios constructed with random Fourier features derived from JKP stock characteristics.

nonlinear characteristic-managed portfolios). From the out-of-sample HJD, we see that high complexity SDFs reduce pricing errors by a factor of six relative to low complexity models.

To further benchmark magnitudes, the figures report out-of-sample performance of the Fama-French six-factor SDF (including momentum and denoted "FF6"). Specifically, we use the MKT, HML, SMB, RMW, CMA, and MOM factors from Ken French's website and construct the out-of-sample FF6 SDF as the rolling 360-month six-factor tangency portfolio.

From this, we calculate the same out-of-sample SDF performance metrics used for complex portfolios. The out-of-sample FF6 Sharpe ratio is 1.1. While this is in line with the other low complexity models, it is nonetheless an impressive feat. The FF6 SDF achieves its performance with a much smaller information set than the 130 characteristics flowing into the low complexity RFF models. Nonetheless, this falls far short of the Sharpe ratios delivered by high complexity models.[38]

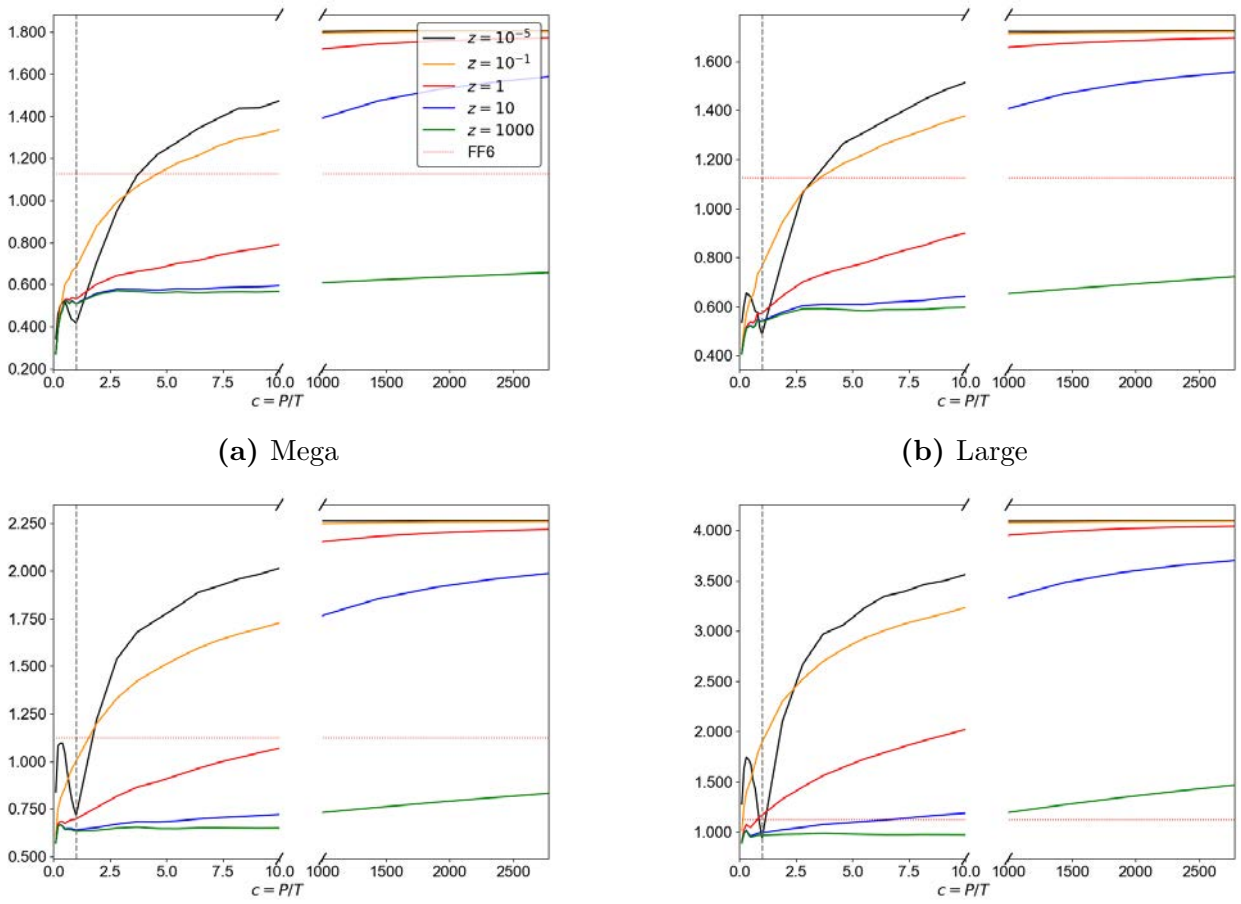
5.4 Results By Market Capitalization

A Sharpe ratio hovering near 4.0 for high complexity SDF models is an indication that the model is likely picking up inefficiencies associated with illiquidity. To understand the role of complexity in factor pricing models while abstracting from the question of asset liquidity, we perform our empirical experiments separately for different market capitalization groups. We study four stock 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 3 plots out-of-sample Sharpe ratio VoC curves for SDFs estimated separately within each size group, and Figure 4 plots pricing errors. The main conclusion from the figure is that the virtue of complexity conforms to our theory predictions in all stock size groups. This means that the patterns in Figure 2 are not driven by illiquidity and limits-to-arbitrage among the underlying assets. Instead, the virtue of complexity reflects that models with a large number of factors are better suited to price assets in the cross section.

In terms of magnitudes, Figure 3 reveals that the high SDF Sharpe ratios in Figure 2 are indeed driven by micro capitalization stocks. While complex SDFs built from micro stocks achieve a Sharpe ratio near 4.0, the SDF Sharpe ratios based on mega or large stocks are

[38] When building the FF6 tangency portfolio we do not use ridge shrinkage because the number of assets is far smaller than the number of time series observations. Even with ex post optimal ridge shrinkage in the FF6 tangency calculation, the resulting out-of-sample SDF portfolio Sharpe increases only negligibly from 1.06 to 1.13.



(a) Mega **(b)** Large

(c) Small **(d)** Micro

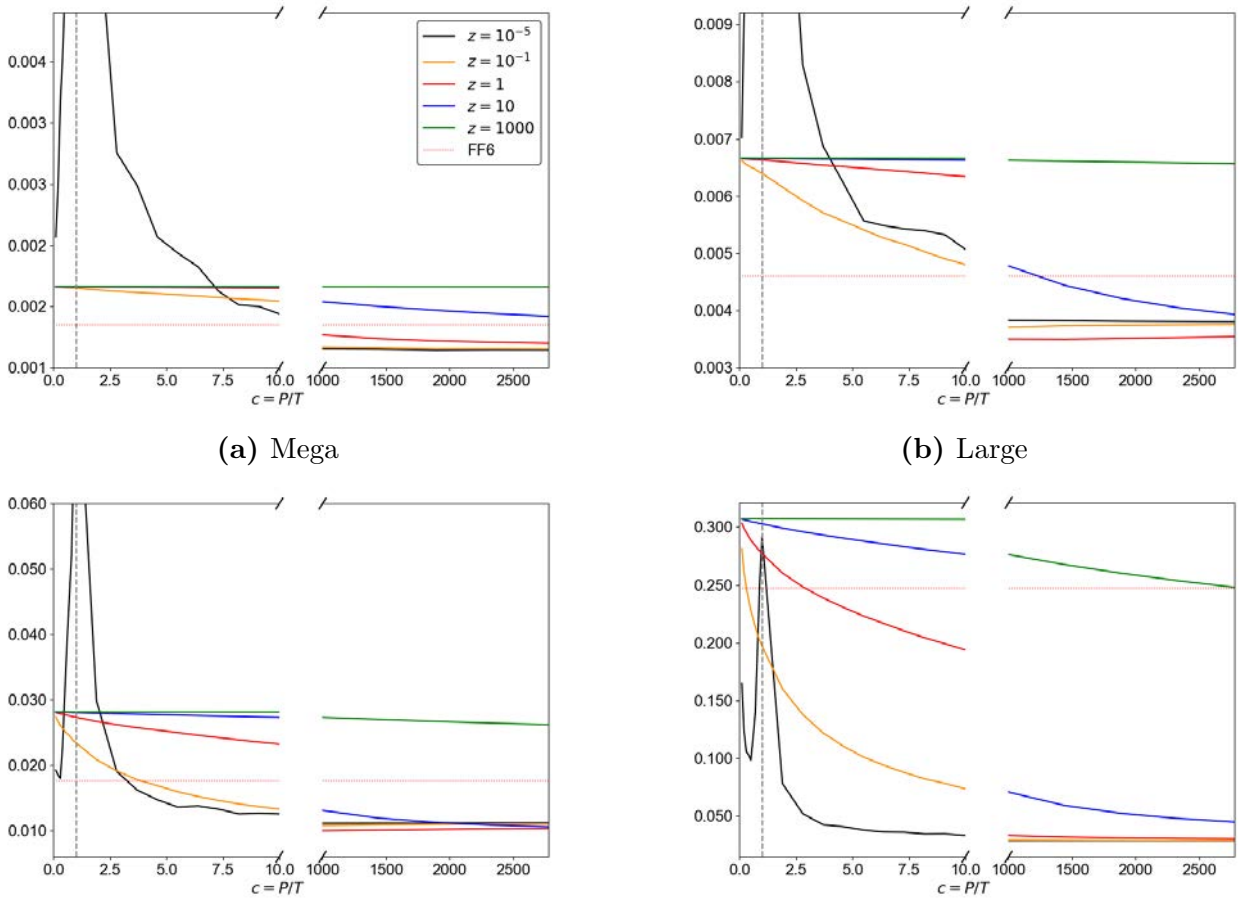
Figure 3: Out-of-sample Sharpe Ratios of Complex SDF Models By Size Group

Note. Realized out-of-sample SDF Sharpe ratio in subsamples based on market capitalization. The horizontal axis shows model complexity $c = P/T$, with P ranging from 2 to 1,000,000 and $T = 360$ months.

on the order of 1.7. Nonetheless, it is impressive that 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 1.1 (which uses the full cross section of stocks).

5.5 The Nonlinear Fama-French Model

In the preceding analysis, we compare high complexity models derived from 130 signals to the FF6 model, which is a low complexity model derived from a small set of signals. Naturally,



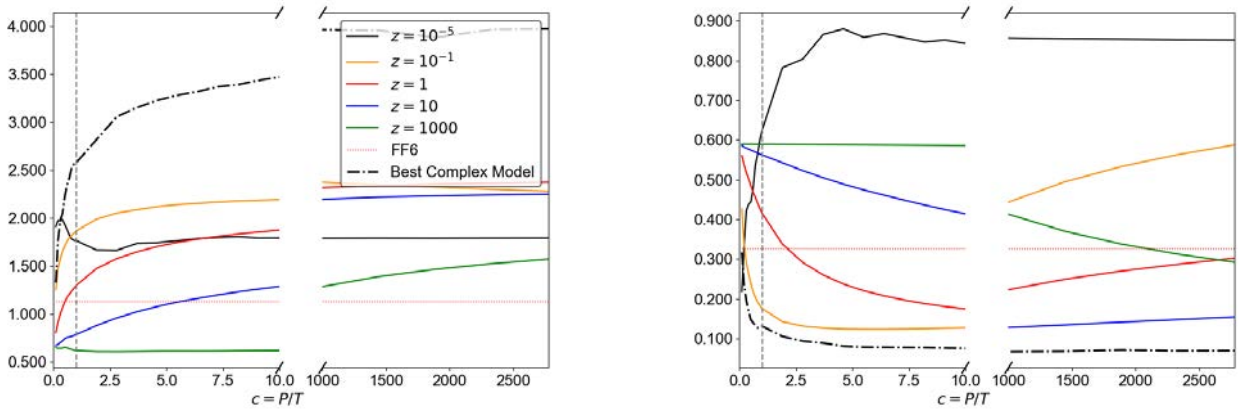
(a) Mega **(b)** Large

(c) Small **(d)** Micro

Figure 4: Out-of-sample Pricing Errors of Complex SDF Models By Size Group

Note. Realized out-of-sample SDF pricing error (HJD) in subsamples based on market capitalization. The horizontal axis shows model complexity $c = P/T$, with P ranging from 2 to 1,000,000 and $T = 360$ months. In each subsample, the test assets are a set of 1,000,000 factor portfolios managed on the basis of nonlinear random Fourier features. Like the SDF, the test assets are constructed from stocks within a given subsample, which is why the FF6 pricing error varies across the four panels.

one may wonder what the benefits of complexity are if we restrict the raw data inputs to those used by Fama-French. To investigate this, we construct complex SDFs using random features derived from *only* the five characteristics that underly the FF6 model. Each of these complex models is a reformulation of Fama-French that employs a large number of factors based on nonlinear transformations of size, value, investment, profitability, and momentum characteristics.



(a) Sharpe Ratio **(b)** Pricing Error

Figure 5: Nonlinear Fama-French Model Performance

Note. Realized out-of-sample SDF Sharpe ratio and pricing error (HJD). The horizontal axis shows model complexity $c = P/T$, with P ranging from 2 to 1,000,000 and $T = 360$ months. In each subsample, the test assets are a set of 1,000,000 factor portfolios managed on the basis of nonlinear random Fourier features. Factors underlying the SDF are characteristic-managed portfolios constructed with random Fourier features derived from the five characteristics underlying the FF6 model: size, value, investment, profitability, and momentum. For ease of reference, "Best Complex Model" shows the best performing complex model from Figure 2 that uses all JKP stock characteristics.

Figure 5 reports the performance on nonlinear Fama-French models with varying degrees of complexity and shrinkage, and compares it to the baseline FF6 model. By allowing for nonlinearity in the SDF, high complexity Fama-French models achieve Sharpe ratios over 2.2, doubling that of the baseline FF6 model, and Fama-French pricing errors drop by more than half thanks to complexity.

The key conclusion from this analysis is that the benefits of SDF complexity accrue even when starting from a small conditioning information set. Complexity uses *any* set of conditioning variables in a flexible manner to more fully express their nonlinear impact.

5.6 SDF Complexity or SDF Sparsity?

A recent spate of financial machine learning research suggests that it is possible to estimate a successful SDF through the imposition of sparsity.[39] 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). They study a collection of difficult-to-price anomaly portfolios which serve as their test assets. They then show that a simple linear SDF—comprised of just 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 SDF complexity that we document above. While our results thus far demonstrate that a complex SDF can identify some effective nonlinear pricing factors, is it possible that we also introduce unnecessary redundancy by using many thousands of factors? We investigate this possibility now.

In our main Figure 2, each point on each curve is a model with a particular number of factors, P, and a particular shrinkage parameter, z. To understand the potential benefits of SDF sparsity, for each SDF model in Figure 2 we fits P nonlinear factors to a small number K of their principal components. We then estimate the ridge SDF from these K components and track the resulting out-of-sample SDF performance.

Figure 6 reports the results. In the left column, we consider a $K = 5$ component dimension reduction of each complex factor model, while the right column shows a reduction to $K = 25$ components. The top row reports out-of-sample SDF Sharpe ratios, and the bottom row reports pricing errors. As a frame of reference, the dotted lines in each plot show the performance of the highest complexity SDF in Figure 2 without dimension reduction. The main conclusion from Figure 6 is that imposing sparsity on the SDF via a principal components dimension reduction inhibits SDF performance relative to the unreduced, high-complexity counterpart. When $K = 5$, the Sharpe ratio of the dimension-reduced SDF is

[39] 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.

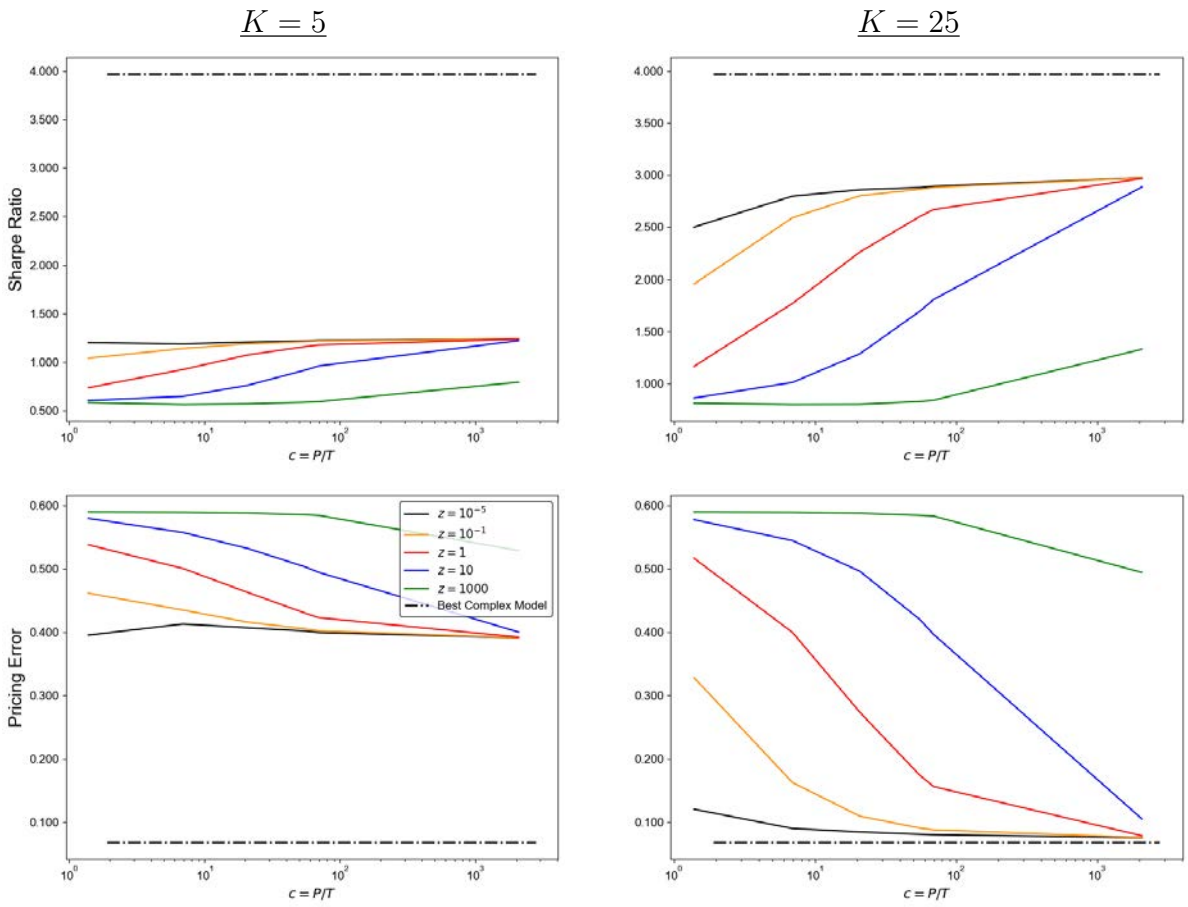


Figure 6: The Effect of Sparsity on Out-of-sample SDF Performance

Note. Realized out-of-sample SDF Sharpe ratio (top row) and pricing error (HJD, bottom row) for complex models with dimension reduction to $K = 5$ (left column) and $K = 25$ (right column) principal components. The horizontal axis shows model complexity $c = P/T$ in log scale, with P ranging from 2 to 1,000,000 and $T = 360$ months. For ease of reference, "Best Complex Model" shows the best performing complex model from Figure 2 that uses all factors without dimension reduction.

roughly 1.2, compared to 4.0 for the full complexity model. Likewise, pricing errors are 0.39 for the $K = 5$ SDF versus 0.07 for the full complexity model.

Two important 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 impossible to detect with enough complexity.[40] Second and more surprisingly, even low-variance components have a significant Sharpe ratio;

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

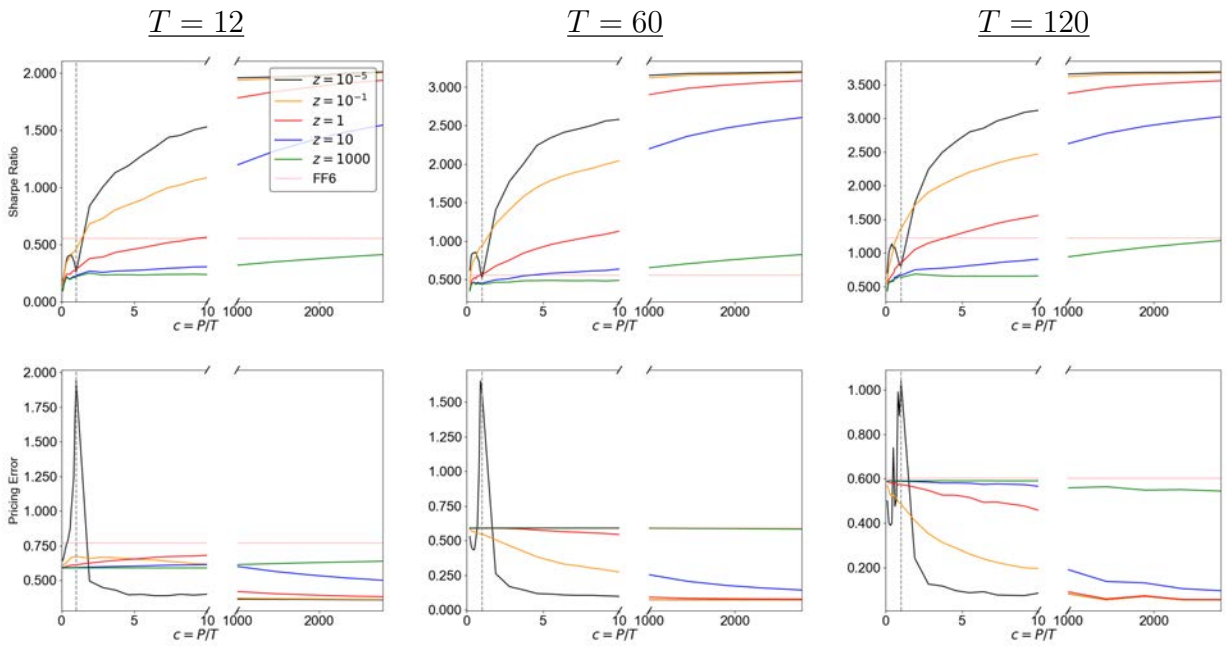


Figure 7: The Effect of Training Sample Size on Out-of-sample SDF Performance

Note. Realized out-of-sample SDF pricing error (HJD) for different training window sizes of $T = 12, 60,$ and 120 months. The horizontal axis shows model complexity $c = P/T$, with P ranging from 2 to 1,000,000.

hence, dropping them leads to a drop in out-of-sample SDF performance. Thus, we should include all components in the SDF portfolio. While this seems counter-intuitive from the point of view of arbitrage pricing theory, the high Sharpe ratios of low-variance components are impossible to realize on a standalone basis because these components are so difficult to estimate.

5.7 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 7 we plot VoC curves for training sample samples of $T = 12, 60$ and 120 months. We find identical patterns in SDF behavior as a function of complexity in training windows as short as a single year. However, for shorter

training windows, SDF model performance weakens. The highest out-of-sample Sharpe ratio for a complex SDF is roughly 2.0, 3.1, 3.6, and 4.0 for $T = 12, 60, 120$, and 360 months, respectively. Likewise, pricing errors reduce from roughly 0.38 for $T = 12$ months to 0.07 for $T = 360$.

6 Conclusion

In this paper we 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 SDF complexity. In essence, out-of-sample performance of factor pricing models (both in terms of SDF Sharpe ratio and test asset pricing errors) generally improves with the number of factors. We also characterize the limits to learning that arise from 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.

Using monthly US stock data, we document an empirical virtue of complexity that bears a strikingly close resemblance to the predictions of our theory. The most successful factor models that we study are those with the very largest number of factors—as many as one million factors in our implementation, with each factor constructed as a managed portfolio whose weights are nonlinear functions of conditioning characteristics. Indeed, we find non-trivial benefits of introducing additional factors even after controlling for several thousand other factors. Relatedly, we show that principal components of our high-dimensional factor sets are incapable of delivering the same caliber of out-of-sample SDF performance that we observe for the complex SDF built from the full set of factors.

References

Antoine, Bertille, Kevin Proulx, and Eric Renault, "Pseudo-true SDFs in conditional asset pricing models," *Journal of Financial Econometrics*, 2020, *18* (4), 656–714.

___ , ___ , and ___ , "Rejoinder on: Pseudo-True SDFs in Conditional Asset Pricing Models," *Journal of Financial Econometrics*, 2020, *18* (4), 776–790.

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.

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.

Brandt, Michael W, Pedro Santa-Clara, and Rossen Valkanov, "Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns," *The Review of Financial Studies*, 2009, *22* (9), 3411–3447.

Britten-Jones, Mark, "The sampling error in estimates of mean-variance efficient portfolio weights," *The Journal of Finance*, 1999, *54* (2), 655–671.

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.

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

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.

DeMiguel, Victor, Alberto Martin-Utrera, Francisco J Nogales, and Raman Uppal, "A transaction-cost perspective on the multitude of firm characteristics," *The Review of Financial Studies*, 2020, *33* (5), 2180–2222.

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

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 and Diego Ronchetti, "Comparing asset pricing models by the conditional Hansen-Jagannathan distance," *Journal of Financial Econometrics*, 2020, *18* (2), 333–394.

___, 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.

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

Jensen, Theis Ingerslev, Bryan T Kelly, and Lasse Heje Pedersen, "Is there a replication crisis in finance?," Technical Report, Journal of Finance Forthcoming.

___ , ___ , **Semyon Malamud, and Lasse Heje Pedersen**, "Machine Learning and the Implementable Efficient Frontier," *Available at SSRN 4187217*, 2022.

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," *Working Paper*, 2023.

___ , **Semyon Malamud, and Kangying Zhou**, "The Virtue of Complexity in Return Prediction," *Swiss Finance Institute Research Paper*, 2021, (21-90).

___ , **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.

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

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

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

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

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.

Liu, Yang, Guofu Zhou, and Yingzi Zhu, "Maximizing the Sharpe ratio: A genetic programming approach," *Available at SSRN 3726609*, 2020.

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.

Nagel, Stefan and Kenneth J Singleton, "Estimation and evaluation of conditional asset pricing models," *The Journal of Finance*, 2011, *66* (3), 873–909.

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.

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.

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.

A Properties of the Infeasible Portfolio

By a direct calculation,[41]

$$\lambda \;=\; E[FF']^{-1}E[F] \;=\; \frac{1}{1+MaxSR^2}\,\mathrm{Var}[F]^{-1}E[F]\,, \tag{68}$$

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

$$MaxSR^2 \;=\; E[F]'\mathrm{Var}[F]^{-1}E[F] \tag{69}$$

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

$$E[\lambda' F_{t+1}] \;=\; E[(\lambda' F_{t+1})^2] \;=\; E[F]'E[FF']^{-1}E[F] \;=\; \frac{MaxSR^2}{1+MaxSR^2}\,. \tag{70}$$

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 $\mathrm{Var}[F] \;=\; U\,\mathrm{diag}(\mu)U'$, we can define PC_i to be the i-th column of $U'F$. In the sequel, we will use

$$\theta \;=\; U'E[F] \tag{71}$$

[41] See the Sherman-Morrison formula (80).

to denote the vector of mean returns of the PCs. Then, we can rewrite the maximal Sharpe ratio (69) as

$$MaxSR^2 \;=\; \sum_i \frac{\theta_i^2}{\mu_i} \;=\; \sum_i (SR(PC_i))^2\,. \tag{72}$$

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

$$R_{t+1}^{infeas}(z) \;=\; E[F]'(zI + \mathrm{Var}[F])^{-1}F_{t+1}\,. \tag{73}$$

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, $R_{t+1}^{infeas}(0) = \lambda' F_{t+1}$ achieves the $MaxSR$, and

$$\mathcal{E}(z) \;=\; E[R^{infeas}(z)] \;=\; E[F]'(zI + E[FF'])^{-1}E[F] \;=\; \frac{A(z)}{1+A(z)}\,, \tag{74}$$

where we have defined

$$\begin{aligned} A(z) \;&=\; E[F]'(zI + \mathrm{Var}[F])^{-1}E[F] \\ &=\; \sum_i (SR(PC_i))^2 \frac{\mu_i}{\mu_i + z} \\ &=\; \sum_i (SR(PC_i))^2 \frac{1}{1 + z/\mu_i} \;\approx\; \sum_{i:\mu_i > z} (SR(PC_i))^2 \end{aligned} \tag{75}$$

and

$$A'(z) \;=\; -\sum_i \theta_i^2 \frac{1}{(\mu_i + z)^2}\,. \tag{76}$$

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

$$(zA(z))' \;=\; \sum_i (SR(PC_i))^2 \left(\frac{\mu_i}{\mu_i + z}\right)^2 , \tag{77}$$

a somewhat tedious calculation implies that

$$\mathrm{Var}[R^{infeas}(z)] = \frac{(zA(z))'}{(1 + A(z))^2} . \tag{78}$$

and

$$\begin{aligned}
\mathcal{V}(z) \;&=\; E[(R^{infeas}(z))^2] \\
&= \frac{1}{(1 + A(z))^2} E[(E[F]'(zI + \Psi)^{-1} F_t)^2] \;=\; \frac{1}{(1 + A(z))^2} E[E[F]'(zI + \Psi)^{-1} F_t F_t'(zI + \Psi)^{-1} E[F]] \\
&= \frac{1}{(1 + A(z))^2} E[E[F]'(zI + \Psi)^{-1} F_t F_t'(zI + \Psi)^{-1} E[F]] \\
&= E[F]'(zI + \Psi)^{-1} \Psi (zI + \Psi)^{-1} E[F] \;+\; \mathcal{R}_1(z)^2 \\
&= \frac{1}{(1 + A(z))^2} \sum_i \theta_i^2 (z + \mu_i)^{-2} \mu_i \;+\; \left(\frac{A(z)}{1 + A(z)}\right)^2 \\
&= \frac{A(z) + zA'(z) + A^2(z)}{(1 + A(z))^2} \\
&= \frac{(A(z) + zA'(z))(1 + A(z)) - zA(z)A'(z)}{(1 + A(z))^2} \\
&= \frac{d}{dz}\left(\frac{zA(z)}{1 + A(z)}\right) .
\end{aligned} \tag{79}$$

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

B Proof of Proposition 2

We will frequently be using the Sherman-Morrison formula

$$(A+xx')^{-1} \;=\; A^{-1} - A^{-1}xx'A^{-1}/(1+x'Ax),\; (A+xx')^{-1}x \;=\; A^{-1}x/(1+x'Ax) \quad (80)$$

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

Lemma 3 *We have*

$$(A+B)^{-1} \;=\; B^{-1} - (A+B)^{-1}AB^{-1}\,, \quad (81)$$

and

$$(A+B)^{-1}AB^{-1} \;\leq\; A \quad (82)$$

in the sense of positive semi-definite order.

Proof of Lemma 3. We have

$$(A+B)^{-1}AB^{-1} \;=\; B^{-1/2}(\hat{A}+I)^{-1}\hat{A}B^{-1/2} \;\leq\; B^{-1/2}\hat{A}B^{-1/2} = B^{-1}AB^{-1} \quad (83)$$

□

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

$$\tilde{w}(S_t) \;=\; (S_t\Sigma_{F,t}S_t' + \Sigma_\varepsilon)^{-1}S_t\lambda_F \quad (84)$$

is the conditionally efficient portfolio with the return

$$R'_{t+1}\tilde{w}(S_t) \;=\; F'_{t+1}(S_t\Sigma_{F,t}S'_t \;+\; \Sigma_\varepsilon)^{-1}S_t\lambda_F\,. \tag{85}$$

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

$$Q_t \;=\; (S_t\Sigma_F^*S'_t \;+\; \Sigma_\varepsilon)^{-1} \;=\; \Sigma_\varepsilon^{-1} \;-\; (S_t\Sigma_F^*S'_t \;+\; \Sigma_\varepsilon)^{-1}S_t\Sigma_F^*S'_t\Sigma_\varepsilon^{-1}\,, \tag{86}$$

we get

$$\begin{aligned}
&E[R'_{t+1}\tilde{w}(S_t)] \\
&=\; E[(S_t\widetilde{F}_{t+1} \;+\; \varepsilon_{t+1})'(S_t(\Sigma_F)S'_t \;+\; \Sigma_\varepsilon)^{-1}S_t\lambda_F] \\
&=\; E[\lambda'_FS'_t(S_t(\Sigma_F)S'_t \;+\; \Sigma_\varepsilon)^{-1}S_t\lambda_F] \\
&=\; E[\lambda'_FS'_t(S_t(\lambda_F\lambda'_F + \Sigma_F^*)S'_t \;+\; \Sigma_\varepsilon)^{-1}S_t\lambda_F] \\
&=\; E[\lambda'_FS'_t((S_t\lambda_F)(S_t\lambda_F)' + (S_t\Sigma_F^*S'_t \;+\; \Sigma_\varepsilon))^{-1}S_t\lambda_F] \\
&\underbrace{=}_{(80)}\; E[\lambda'_FS'_t(Q_t - Q_tS_t\lambda_F\lambda'_FS'_tQ_t(1 + \lambda'_FS'_tQ_tS_t\lambda_F)^{-1})S_t\lambda_F] \\
&=\; E[Z_t - Z_t^2(1+Z_t)^{-1}] \;=\; E[Z_t/(1+Z_t)]\,,
\end{aligned} \tag{87}$$

where we have defined

$$Z_t \;=\; \lambda'_FS'_tQ_tS_t\lambda_F \;=\; \lambda'_FS'_t\Sigma_\varepsilon^{-1}S_t\lambda_F - q\,, \tag{88}$$

with

$$q \;=\; \lambda_F' S_t' \Sigma_\varepsilon^{-1} S_t \lambda_F - \lambda_F' S_t' Q_t S_t \lambda_F \,. \tag{89}$$

By Lemma 3,

$$(S_t \Sigma_F^* S_t' \;+\; \Sigma_\varepsilon)^{-1} S_t \Sigma_F^* S_t' \Sigma_\varepsilon^{-1} \;\leq\; \Sigma_\varepsilon^{-1} S_t \Sigma_F^* S_t' \Sigma_\varepsilon^{-1} \tag{90}$$

and hence

$$q \;=\; \lambda_F' S_t' (S_t \Sigma_F^* S_t' \;+\; \Sigma_\varepsilon)^{-1} S_t \Sigma_F^* S_t' \Sigma_\varepsilon^{-1} S_t \lambda_F \;\leq\; \lambda_F' S_t' \Sigma_\varepsilon^{-1} S_t \Sigma_F^* S_t' \Sigma_\varepsilon^{-1} S_t \lambda_F \,. \tag{91}$$

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 Corollary 9,

$$\begin{aligned} E[\lambda_F' S_t' \Sigma_\varepsilon^{-1} S_t A S_t' \Sigma_\varepsilon^{-1} S_t \lambda_F] \;&=\; \lambda_F' \Big(((\operatorname{tr} \hat\Sigma)^2 + \operatorname{tr}(\hat\Sigma^2)) \Psi A \Psi + \operatorname{tr}(\hat\Sigma^2) \operatorname{tr}(\Psi A) \Psi \\ &+\; \operatorname{tr}(\hat\Sigma^2) \, (\kappa - 2) \Psi^{1/2} \operatorname{diag}(\Psi^{1/2} A \Psi^{1/2}) \Psi^{1/2} \Big) \lambda_F \\ &=\; (\operatorname{tr} \hat\Sigma)^2 \lambda_F' \Big((1 + \frac{\operatorname{tr}(\hat\Sigma^2)}{(\operatorname{tr} \hat\Sigma)^2}) \Psi A \Psi + \frac{\operatorname{tr}(\hat\Sigma^2)}{(\operatorname{tr} \hat\Sigma)^2} \operatorname{tr}(\Psi A) \Psi \\ &+\; \frac{\operatorname{tr}(\hat\Sigma^2)}{(\operatorname{tr} \hat\Sigma)^2} \, (\kappa - 2) \Psi^{1/2} \operatorname{diag}(\Psi^{1/2} A \Psi^{1/2}) \Psi^{1/2} \Big) \lambda_F \end{aligned} \tag{92}$$

with

$$A \;=\; \Sigma_F^* \tag{93}$$

and

$$\hat{\Sigma} \;=\; \Sigma^{1/2}\Sigma_\varepsilon^{-1}\Sigma^{1/2}\,. \tag{94}$$

By Assumption 2, $\frac{\operatorname{tr}(\hat{\Sigma}^2)}{(\operatorname{tr}\hat{\Sigma})^2} \to 0$ and, since λ_F and Ψ and A and $\operatorname{tr}(\hat{\Sigma})$ are uniformly bounded, we get that

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

Since, by Assumption 1, $\operatorname{tr}(A)$ is uniformly bounded, we also get that $\operatorname{tr}(\Psi A \Psi) \le \|\Psi\|^2 \operatorname{tr}(A)$ is uniformly bounded and, hence, $\lambda_F' \Psi A \Psi \lambda_F \to 0$ by (15).

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

$$E[\lambda_F' S_t' \Sigma_\varepsilon^{-1} S_t \lambda_F] \;=\; \operatorname{tr}(\hat{\Sigma}) \lambda_F' \Psi \lambda_F \tag{96}$$

whereas, by Corollary 9,

$$\begin{aligned} E[(\lambda_F' S_t' \Sigma_\varepsilon^{-1} S_t \lambda_F)^2] \;&=\; E[\lambda_F' S_t' \Sigma_\varepsilon^{-1} S_t \lambda_F \lambda_F' S_t' \Sigma_\varepsilon^{-1} S_t \lambda_F] \\ &=\; \lambda_F' \Big(((\operatorname{tr}\hat{\Sigma})^2 + \operatorname{tr}(\hat{\Sigma}^2)) \Psi \lambda_F \lambda_F' \Psi + \operatorname{tr}(\hat{\Sigma}^2) \operatorname{tr}(\Psi \lambda_F \lambda_F') \Psi \\ &+\; \operatorname{tr}(\hat{\Sigma}^2)\,(\kappa - 2) \Psi^{1/2} \operatorname{diag}(\Psi^{1/2} \lambda_F \lambda_F' \Psi^{1/2}) \Psi^{1/2} \Big) \lambda_F \end{aligned} \tag{97}$$

and the same argument as in (92) implies that

$$E[(\lambda_F' S_t' \Sigma_\varepsilon^{-1} S_t \lambda_F)^2] \;\approx\; \operatorname{tr}(\hat{\Sigma})^2 (\lambda_F' \Psi \lambda_F)^2\,. \tag{98}$$

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

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

$$\frac{Z_t}{1+Z_t} \;-\; \frac{\mathrm{tr}(\hat{\Sigma})\lambda_F'\Psi\lambda_F}{1+\mathrm{tr}(\hat{\Sigma})\lambda_F'\Psi\lambda_F} \;\to\; 0 \tag{99}$$

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

$$\begin{aligned} & E[(\pi_t^{MV})' R_{t+1} R_{t+1}' \pi_t^{MV}] \\ &= E[\lambda' S_t'(S_t(\Sigma_F)S_t' \;+\; \Sigma_\varepsilon)^{-1}(S_t(\Sigma_F)S_t' \;+\; \Sigma_\varepsilon)(S_t(\Sigma_F)S_t' \;+\; \Sigma_\varepsilon)^{-1} S_t\lambda] \\ &= E[R_{t+1}'\pi_t^{MV}] \;\to\; \frac{\mathrm{tr}(\hat{\Sigma})\lambda_F'\Psi\lambda_F}{1+\mathrm{tr}(\hat{\Sigma})\lambda_F'\Psi\lambda_F}\,. \end{aligned} \tag{100}$$

Now, for the factor portfolios, we have

$$\begin{aligned} E[F_t] \;&=\; E[S_t'R_{t+1}] \;=\; E[S_t'(S_t\widetilde{F}_{t+1} \;+\; \varepsilon_{t+1})] \\ &=\; E[S_t'S_t\widetilde{F}_{t+1}] \;=\; E[\Psi^{1/2}X_t'\Sigma X_t\Psi^{1/2}\widetilde{F}_{t+1}] \\ &=\; E[\Psi^{1/2}X_t'\Sigma X_t\Psi^{1/2}]\lambda_F \\ &=\; \mathrm{tr}(\Sigma)\, E[\Psi^{1/2}\Psi^{1/2}]\lambda_F \\ &=\; \mathrm{tr}(\Sigma)\,\Psi\lambda_F\,, \end{aligned} \tag{101}$$

and, again by Corollary 9 and the same argument as in (92), we have

$$\begin{aligned} & E[F_tF_t'] \;=\; E[S_t'(S_t\widetilde{F}_{t+1} \;+\; \varepsilon_{t+1})(S_t\widetilde{F}_{t+1} \;+\; \varepsilon_{t+1})'S_t|\lambda] \;=\; E[S_t'(S_t(\Sigma_F)S_t' \;+\; \Sigma_\varepsilon)S_t] \\ &\approx\; \mathrm{tr}(\Sigma\Sigma_\varepsilon)\Psi \;+\mathrm{tr}(\Sigma)^2\Psi(\Sigma_F)\Psi\,. \end{aligned} \tag{102}$$

Then, defining

$$Q \;=\; (\mathrm{tr}(\Sigma\Sigma_\varepsilon)\Psi \;+\; \mathrm{tr}(\Sigma)^2\Psi\Sigma_F^*\Psi)^{-1}\,, \tag{103}$$

we get that the efficient portfolio of factors is given by

$$\begin{aligned}\pi_F &= (\mathrm{tr}(\Sigma\Sigma_\varepsilon)\Psi \ + \ \mathrm{tr}(\Sigma)^2\Psi\Sigma_F\Psi)^{-1}\Psi\lambda_F \\ &= \ (\mathrm{tr}(\Sigma\Sigma_\varepsilon)\Psi \ + \ \mathrm{tr}(\Sigma)^2\Psi(\Sigma_F^* + \lambda_F\lambda_F')\Psi)^{-1}\Psi\lambda_F \\ &\underbrace{=}_{(80)} \ \frac{1}{1+Z}Q\Psi\lambda_F\,,\end{aligned} \tag{104}$$

where

$$Z \ = \ \mathrm{tr}(\Sigma)^2\lambda_F'\Psi Q\Psi\lambda_F\,. \tag{105}$$

By the same argument as above,

$$\lambda_F'(\Psi Q\Psi - \Psi(\mathrm{tr}(\Sigma\Sigma_\varepsilon)\Psi)^{-1}\Psi)\lambda_F \ \to 0 \tag{106}$$

by Assumption 15 because Σ_*^F has a bounded trace. Thus,

$$Z \ \approx \ \frac{\mathrm{tr}(\Sigma)^2}{\mathrm{tr}(\Sigma\Sigma_\varepsilon)}\lambda_F'\Psi\lambda_F \tag{107}$$

and

$$E[\pi_F'F_{t+1}] \ = \ E[\lambda_F'\frac{1}{1+Z}\Psi Q\Psi\lambda_F] \approx \ \frac{Z}{1+Z}\,, \tag{108}$$

while

$$E[\pi_F'F_{t+1}F_{t+1}'\pi_F] \ = \ E[\pi_F'F_{t+1}]\,, \tag{109}$$

and the proof is complete because

$$\mathrm{tr}(\Sigma\Sigma_\varepsilon^{-1})\lambda_F'\Psi\lambda_F \;=\; \frac{\mathrm{tr}(\Sigma)^2}{\mathrm{tr}(\Sigma\Sigma_\varepsilon)}\lambda_F'\Psi\lambda_F \tag{110}$$

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. □

C Auxilliary Results

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

Lemma 4 *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*

$$Y_t \;=\; X_t'A_PX_t\,. \tag{111}$$

Then,

$$(1)\; Y_t = \mathrm{tr}(A_PX_tX_t') \tag{112}$$

$$(2)\; \lim_{P\to\infty} E[(Y_t - \mathrm{tr}(A_P))^2|A_P] = 0 \tag{113}$$

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

$$\lim_{P\to\infty} E[(X_t'B_PX_t - \mathrm{tr}(B_P))^2|B_P]/P^2 \;=\; 0\,. \tag{114}$$

Proof of Lemma 4.

(1):

$$X_t'AX_t \in R \Rightarrow X_t'AX_t = \mathrm{tr}(X_t'AX_t)$$
$$\mathrm{tr}(AB) = \mathrm{tr}(BA) \Rightarrow \mathrm{tr}(X_t'AX_t) = \mathrm{tr}(AX_tX_t')$$

(2): Define $Y_t = X_tA_PX_t$. We have

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

and hence

$$E[(Y_t - \mathrm{tr}(A_P))^2|A_P] \;=\; \mathrm{Var}[Y_t|A_P] \;=\; E[Y_t^2|A_P] - E[Y_t|A_P]^2 \tag{115}$$

and hence it suffices to prove that

$$E[Y_t^2|A_P] \;-(\mathrm{tr}(A_P))^2 \;\to\; 0 \tag{116}$$

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

$$Y_t \;=\; \sum_{i,j} X_iX_jA_{i,j} \tag{117}$$

and therefore

$$Y_t^2 \;=\; \sum_{i_1,j_1,i_2,j_2} X_{i_1}X_{j_1}A_{i_1,j_1}A_{i_2,j_2}X_{i_2}X_{j_2} \tag{118}$$

Now we attempt to compute the expectation:

$$\begin{aligned} E[Y_t^2] &= \sum_{i_1,j_1,i_2,j_2} A_{i_1,j_1} A_{i_2,j_2} E[X_{i_1} X_{j_1} X_{i_2} X_{j_2}] \\ &= (\sum_i A_{i,i}^2 E[X_i^4] + \sum_{i,j} (A_{i,j}^2 + A_{i,i} A_{j,j}) E[X_i^2 X_j^2] \\ &= (\sum_i k A_{i,i}^2 + \sum_{i,j} A_{i,j}^2 + 2 A_{i,i} A_{j,j}) \\ &= ((k-1) \sum_i A_{i,i}^2 + \sum_{i,j} A_{i,j}^2 + \mathrm{tr}(A)^2) \end{aligned} \tag{119}$$

We have

$$\sum_i A_{i,i}^2 \;\leq\; \sum_{i,j} A_{i,j}^2 \;=\; \|A\|_2^2 \,, \tag{120}$$

and therefore

$$|E[Y_t^2] - \mathrm{tr}(A)^2| \;\leq\; k \|A_2\|_2^2 \,, \tag{121}$$

and the proof is complete. □

Recall that

$$F_{t+1} \;=\; S_t' \, R_{t+1} \tag{122}$$

and

$$\hat{\lambda}(z) \;=\; (zI + B_T)^{-1} \frac{1}{T} \sum_{t=1}^{T} F_t \tag{123}$$

where

$$B_T \;=\; \frac{1}{T}\sum_{t=1}^{T} F_t F_t' \,, \tag{124}$$

while

$$\hat{R}^M_{T+1}(z) \;=\; \hat{\lambda}(z)' F_{t+1} \;=\; (S_t\hat{\lambda}(z))' R_{t+1} \,. \tag{125}$$

In the sequel, to simplify some expressions, we often assume that factor risk premia $\lambda_F \sim N(0, \Sigma_\lambda/P)$ for some uniformly bounded sequence of matrices $\Sigma_\lambda = \Sigma_\lambda(P)$. In this case,

$$\lambda_F' A \lambda_F \;\approx P^{-1}\operatorname{tr}(A\Sigma_\lambda) \tag{126}$$

in probability (and in L_2). All our results hold under the more general condition (15), and all expressions can be rewritten without Σ_λ using (126).

Lemma 5 *We have*

$$(\tilde{F}_{t+1}' A_P \tilde{F}_{t+1} \;-\; \operatorname{tr}((\Sigma_{F,t} A_P) + P^{-1}\operatorname{tr}(A_P\Sigma_\lambda))) \;\to\; 0 \tag{127}$$

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

Proof of Lemma 5. The proof follows directly from Lamme 4. □

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

Lemma 6 *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,*

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

We can now prove

Lemma 7 (Expected Factor Moments) *Suppose a normalization* $\operatorname{tr}(\Sigma) = 1$ *and let* $\sigma_* = \operatorname{tr}(\Sigma\Sigma_\varepsilon)$ *and* $E[X_{i,k}^4] = \kappa$ *for all* i, k. *We have*

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

and

$$\begin{aligned} E[F_{t+1}F_{t+1}'] \;&=\; ((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2))\Psi\Sigma_F\Psi \\ &+\; \operatorname{tr}(\Sigma^2)\Psi^{1/2}\operatorname{diag}(\kappa-2)\operatorname{diag}(\Psi^{1/2}\Sigma_F\Psi^{1/2})\Psi^{1/2} + \Psi\Big(\operatorname{tr}(\Sigma\Sigma_\varepsilon) + \operatorname{tr}(\Psi\Sigma_F)\operatorname{tr}(\Sigma^2)\Big) \end{aligned} \tag{128}$$

Thus,

$$\|E[F_{t+1}F_{t+1}'] - (\Psi\Sigma_F\Psi + \sigma_*\Psi)\| \;\to\; 0 \tag{129}$$

Proof of Lemma 7. Recall that, by Assumption 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

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

Defining $\tilde{\beta} = \Psi^{1/2}\tilde{F}_{t+1}$, we get

$$\begin{aligned} E[S_t'S_t\tilde{\beta}\tilde{\beta}'S_t'S_t] \;&=\; E[\Psi^{1/2}X_t'\Sigma X_t\Psi^{1/2}\tilde{\beta}\tilde{\beta}'\Psi^{1/2}X_t'\Sigma X_t\Psi^{1/2}] \;= E[\Psi^{1/2}X_t'\Sigma X_t\tilde{\beta}\tilde{\beta}'X_t'\Sigma X_t\Psi^{1/2}] \\ &=\; \Psi^{1/2}E[\tilde{X}_t'D\tilde{X}_t\tilde{\beta}\tilde{\beta}'\tilde{X}_t'D\tilde{X}_t]\Psi^{1/2}\,, \end{aligned} \tag{130}$$

where we have defined $\Sigma = U'DU$ and D is diagonal and U is orthogonal and $\tilde{X} = UX$ are still have the same moments as X by the assumptions made.

Now,

$$E[\tilde{X}_t' D\tilde{X}_t \tilde{\beta}\tilde{\beta}' \tilde{X}_t' D\tilde{X}_t]_{k_1,k_2} \;=\; E[\sum_{i_1,i_2,l_1,l_2} D_{i_1} D_{i_2}\, \tilde{X}_{i_1,k_1}\tilde{X}_{i_1,l_1}\tilde{\beta}_{l_1}\tilde{\beta}_{l_2}\tilde{X}_{i_2,l_2}\tilde{X}_{i_2,k_2}]\,.$$

First we study the terms with $i_1 \neq i_2$:

$$\sum_{i_1\neq i_2} D_{i_1} D_{i_2}\, E[\sum_{l_1,l_2} \tilde{X}_{i_1,k_1}\tilde{X}_{i_1,l_1}\tilde{\beta}_{l_1}\tilde{\beta}_{l_2}\tilde{X}_{i_2,l_2}\tilde{X}_{i_2,k_2}] \;=\; \sum_{i_1\neq i_2} D_{i_1} D_{i_2}\tilde{\beta}_{k_1}\tilde{\beta}_{k_2} = ((\operatorname{tr}\Sigma)^2 - \operatorname{tr}(\Sigma^2))\tilde{\beta}_{k_1}\tilde{\beta}_{k_2}$$

At the same time,

$$\sum_{i_1=i_2} D_{i_1}^2\; E[\sum_{l_1,l_2} \tilde{X}_{i_1,k_1}\tilde{X}_{i_1,l_1}\tilde{\beta}_{l_1}\tilde{\beta}_{l_2}\tilde{X}_{i_2,l_2}\tilde{X}_{i_2,k_2}]$$

depends on whether $k_1 = k_2$. If $k_1 = k_2$, then we have

$$\sum_{i_1=i_2} D_{i_1}^2\; E[\sum_{l_1,l_2} \tilde{X}_{i_1,k_1}^2\tilde{X}_{i_1,l_1}\tilde{\beta}_{l_1}\tilde{\beta}_{l_2}\tilde{X}_{i_1,l_2}] \;=\; \operatorname{tr}(\Sigma^2)(\kappa\tilde{\beta}_{k_1}^2 + \|\tilde{\beta}\|^2)$$

and if $k_1 \neq k_2$ then we need that ℓ_1, ℓ_2 coincide with k_1, k_2, so that

$$\sum_{i_1=i_2} D_{i_1} D_{i_2} E[\sum_{l_1,l_2} \tilde{X}_{i_1,k_1}\tilde{X}_{i_1,l_1}\tilde{\beta}_{l_1}\tilde{\beta}_{l_2}\tilde{X}_{i_2,l_2}\tilde{X}_{i_2,k_2}] \;=\; 2\sum_{i_1=i_2} D_{i_1}^2 E[\tilde{X}_{i_1,k_1}^2\tilde{X}_{i_1,k_2}^2]\tilde{\beta}_{k_1}\tilde{\beta}_{k_2} \;=\; 2\operatorname{tr}(\Sigma^2)\tilde{\beta}_{k_1}\tilde{\beta}_{k_2}$$

Thus,

$$\begin{aligned}
&E[\tilde{X}_t' D\tilde{X}_t \tilde{\beta}\tilde{\beta}' \tilde{X}_t' D\tilde{X}_t]_{k_1,k_2} \\
&= ((\operatorname{tr}\Sigma)^2 - \operatorname{tr}(\Sigma^2))\tilde{\beta}_{k_1}\tilde{\beta}_{k_2} \;+ 2\operatorname{tr}(\Sigma^2)\tilde{\beta}_{k_1}\tilde{\beta}_{k_2}(1-\delta_{k_1,k_2}) + \;\operatorname{tr}(\Sigma^2)(\kappa\tilde{\beta}_{k_1}^2 + \|\tilde{\beta}\|^2)\delta_{k_1,k_2} \quad (131)\\
&= ((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2))\tilde{\beta}_{k_1}\tilde{\beta}_{k_2} + \operatorname{tr}(\Sigma^2)((\kappa-2)\tilde{\beta}_{k_1}^2 + \|\tilde{\beta}\|^2)\delta_{k_1,k_2}
\end{aligned}$$

Thus, by formula (130), we get

$$E[S_t'S_t\lambda\lambda'S_t'S_t] \;=\; ((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))\Psi\Sigma_F\Psi+\operatorname{tr}(\Sigma^2)((\kappa-2)\Psi^{1/2}\operatorname{diag}(\tilde\beta_{k_1}^2)\Psi^{1/2}+\|\tilde\beta\|^2\Psi) \quad (132)$$

and the claim follows because $\|\tilde\beta\|^2=\lambda'\lambda_F$.

Corollary 9 *We have*

$$\begin{aligned} E[S_t'S_tAS_t'S_t] \;&=\; ((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))\Psi A\Psi+\operatorname{tr}(\Sigma^2)\operatorname{tr}(\Psi A)\Psi \\ &+\; \operatorname{tr}(\Sigma^2)\,\Psi^{1/2}\operatorname{diag}(\kappa-2)\operatorname{diag}(\Psi^{1/2}A\Psi^{1/2})\Psi^{1/2} \end{aligned} \quad (133)$$

where $\operatorname{diag}(\Psi^{1/2}A\Psi^{1/2})$ *is the diagonal matrix with diagonal coinciding with that of* $\operatorname{diag}(\Psi^{1/2}A\Psi^{1/2})$.

Proof. Writing

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

we can apply the calculations for rank-one A. □

The proof of Lemma 7 is complete.

□

D Technical Results from RMT

Theorem 10 *The eigenvalue distribution of* $E[F_tF_t']$ *converges to that of* $\Psi\sigma_*$ *where* $\sigma_*=\lim\operatorname{tr}(\Sigma\Sigma_\varepsilon)$ *in the limit as* $N,P,T\to\infty,\ P/T\to c$, *so that*

$$\frac{1}{P}\operatorname{tr}\left((zI+E[F_tF_t'])^{-1}\right) \;\to\; \sigma_*^{-1}m_\Psi(-z/\sigma_*) \;=\; m_{\sigma_*\Psi}(-z) \;=\; \frac{1}{P}\operatorname{tr}((zI+\sigma_*\Psi)^{-1}), \quad (134)$$

whereas

$$\frac{1}{P}\operatorname{tr}((zI + B_T)^{-1}) \;\to\; m(-z;c)\,, \tag{135}$$

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

$$m(z;c) = \frac{1}{1 \;-\; c \;-\; c\,z\,m(z;c)}\, m_{\sigma_*\Psi}\left(\frac{z}{1 \;-\; c \;-\; c\,z\,m(z;c)}\right). \tag{136}$$

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. See Appendix F for details.

We will also need the following lemma from KMZ.

Lemma 8 *Define* $\xi(z;c)$ *through*

$$\frac{c^{-1}\xi(z;c)}{1+\xi(z;c)} \;=\; 1 \;-\; m(-z;c)z\,. \tag{137}$$

Then,

$$\frac{1}{T}\operatorname{tr}((zI + B_T)^{-1}\Psi) \;\to\; \xi(z;c) \tag{138}$$

almost surely and

$$\frac{1}{T}F_{T+1}'(zI + B_T)^{-1}F_{T+1} \;\to\; \xi(z;c) \tag{139}$$

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

Define the effective shrinkage

$$Z^*(z;c) \;=\; z\,(1+\xi(z;c)) \;\in (z,\; z+c) \tag{140}$$

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

$$Z^*(z;c) \;\to\; \begin{cases} 0, & c<1 \\ 1/\tilde{m}(c), & c>1 \end{cases} \tag{141}$$

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

$$c-1 \;=\; \frac{\int \frac{dH(x)}{\tilde{m}(1+\tilde{m}\,x)}}{\int \frac{x dH(x)}{1+\tilde{m}\,x}} \tag{142}$$

E The Proof that Managed Portfolio Returns Satisfy Assumption of RMT

Lemma 9 *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

$$\begin{aligned} &\mathrm{tr}((zI + A_P + X_P)^{-1} X_P (zI + A_P)^{-1}) \;=\; \mathrm{tr}(X_P (zI + A_P)^{-1} (zI + A_P + X_P)^{-1}) \\ &\leq\; \mathrm{tr}(X_P) \| (zI + A_P)^{-1} (zI + A_P + X_P)^{-1} \| \;\leq\; K z^{-2} \end{aligned} \tag{143}$$

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

We will need the following auxiliary lemma.

Lemma 10 *Let ε be a random vector with independent $N(0,1)$ coordinates. We have*

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

and

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

for any vector Z. Furthermore,

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

for any matrix A. Furthermore,

$$E[\varepsilon_t' B \varepsilon_t \varepsilon_t' B \varepsilon_t] \;=\; (\kappa_\varepsilon - 1)\, 0.5(\mathrm{tr}(BB) + \mathrm{tr}(B'B)) \;+\; \mathrm{tr}(B)^2 \tag{144}$$

and

$$E[\varepsilon_t \varepsilon_t' B \varepsilon_t \varepsilon_t'] \;=\; (\kappa_\varepsilon - 1) 0.5 (B + B') \;+\; \mathrm{tr}(B)$$

where $\kappa_\varepsilon = E[\tilde{\varepsilon}^4]$.

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

$$E[\varepsilon' A \varepsilon] \;=\; \operatorname{tr} E[\varepsilon' A \varepsilon] = \operatorname{tr} E[A\varepsilon\varepsilon'] = \operatorname{tr}(A) \tag{145}$$

For the last claim: first, we do a transformation $\varepsilon_t = \tilde{\varepsilon}_t$ and then we make the observation that, for any matrix B,

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

Since $0.5(B + B')$ is symmetric, we can diagonalize it: $\tilde{B} = (0.5(B + B'))$. Then,

$$E[\varepsilon_t' B \varepsilon_t \varepsilon_t' B \varepsilon_t] \;=\; E[(\sum_i \varepsilon_{i,t}^2 \lambda_i(0.5(B + B')))^2] \;=\; (\kappa_\varepsilon - 1)\, \operatorname{tr}(\tilde{B}^2) \;+\; \operatorname{tr}(\tilde{B})^2 \,, \tag{146}$$

and we have

$$\operatorname{tr}(\tilde{B}^2) \;=\; \operatorname{tr}((0.5(B + B'))(0.5(B + B'))) \;=\; 0.25(\operatorname{tr}(BB) + 2(\operatorname{tr} B'B) + \operatorname{tr}(B'B'))$$

and

$$\operatorname{tr}(B'B') \;=\; \operatorname{tr}(B'B') \;=\; \operatorname{tr}(BB) \,.$$

Let $\varepsilon = \tilde{\varepsilon}$ and $\tilde{B} = U\Lambda U'$ and $\hat{\varepsilon} = U'\tilde{\varepsilon}$

$$\begin{aligned} & E[\varepsilon_t \varepsilon_t' B \varepsilon_t \varepsilon_t'] \\ = \; & E[\tilde{\varepsilon}\tilde{\varepsilon}'\tilde{B}\tilde{\varepsilon}\tilde{\varepsilon}'] \\ = \; & UE[\hat{\varepsilon}\hat{\varepsilon}'\Lambda\hat{\varepsilon}\hat{\varepsilon}']U' \\ = \; & UE[\hat{\varepsilon}\sum_i \hat{\varepsilon}_{i_1}^2 \lambda_{i_1}(\tilde{B})\hat{\varepsilon}']U' \\ = \; & (\kappa_\varepsilon - 1)\tilde{B} \; + \; \mathrm{tr}(\tilde{B}) \\ = \; & (\kappa_\varepsilon - 1)0.5(B + B') \; + \; \mathrm{tr}(B) \end{aligned} \tag{147}$$

□

Lemma 11 (Managed Portfolios Satisfy The RMT Conditions) *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*

$$\mathrm{Var}[\frac{1}{T} F_t' A_P F_t] \; \to \; 0\,, \tag{148}$$

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 7,

$$\frac{1}{T}\,\mathrm{tr}(A_P E[F_t F_t']) \; \approx \; \frac{1}{T}\,\mathrm{tr}(A_P(\Psi\Sigma_F\Psi + \sigma_*\Psi))\,. \tag{149}$$

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

$$\frac{1}{T}\operatorname{tr}(A_P(\Psi\Sigma_F\Psi+\sigma_*\Psi)) \;\approx\; \frac{1}{T}\operatorname{tr}(A_P(\sigma_*\Psi)) \tag{150}$$

Similarly, the following is a direct consequence of Lemma 7.

Lemma 12 *Let A_P, B_P be sequences of symmetric $P\times P$ matrices such that $\|A_P\|,\ \|B_P\|\leq K$, and $A_P,\ B_P$ are independent of F_t. Then,*

$$(\lambda' E[A_P F_t F_t' B_P]\lambda \;-\; \lambda' A_P(\Psi\Sigma_F\Psi+\sigma_*\Psi)B_P\lambda) \;\to\; 0\,.$$

If λ satisfies the technical condition (15), then

$$(\lambda' E[A_P F_t F_t' B_P]\lambda \;-\; \lambda' A_P(\lambda_F\lambda'\Psi+\sigma_*\Psi)B_P\lambda) \;\to\; 0$$

because $\operatorname{tr}(\Sigma_{F,t})$ is bounded.

Note that $\operatorname{tr}(A_P\,F_tF_t') \;=\; F_t'A_PF_t$.

Proof of Lemma 11. For simplicity, we will assume that A_P is deterministic.[42] We can also assume that A_P is symmetric because $F_t'A_PF_t = F_t 0.5(A_P+A_P')F_t$. We need to prove that

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

For simplicity, we will assume that $\Sigma_\varepsilon = I$. We have by Lemma 7 that

$$\begin{aligned} E[F_tF_t'] \;&=\; ((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))\Psi\Sigma_F\Psi \\ &+\operatorname{tr}(\Sigma^2)\Psi^{1/2}\operatorname{diag}(\kappa-2)\operatorname{diag}(\Psi^{1/2}\Sigma_F\Psi^{1/2})\Psi^{1/2} \;+\Psi\Big(\operatorname{tr}(\Sigma)+\operatorname{tr}(\Psi\Sigma_F)\operatorname{tr}(\Sigma^2)\Big) \end{aligned} \tag{151}$$

[42]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

$$\begin{aligned}
\frac{1}{T}E[F_t' A_P F_t] &= \frac{1}{T}\operatorname{tr} E[A_P F_t F_t'] \\
&\approx \frac{1}{T}\operatorname{tr}\Bigg(A_P\Bigg((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \operatorname{tr}(\Sigma^2)\Psi^{1/2}\operatorname{diag}(\kappa-2)\operatorname{diag}(\Psi^{1/2}\Sigma_F\Psi^{1/2})\Psi^{1/2} \\
&+ \Psi\Big(\operatorname{tr}(\Sigma) + \operatorname{tr}(\Psi\Sigma_F)\operatorname{tr}(\Sigma^2)\Big)\Bigg)\Bigg) \\
&\approx T^{-1}\operatorname{tr}(A_P\Psi)
\end{aligned} \tag{152}$$

since

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

and, similarly, the kurtosis term does not matter because it has a uniformly bounded trace.

Now, we have

$$\begin{aligned}
F_t F_t' &= S_{t-1}'(S_{t-1}\beta\beta' S_{t-1}' + \varepsilon_t\beta' S_{t-1}' + S_{t-1}\beta\varepsilon_t' + \varepsilon_t\varepsilon_t')S_{t-1} \\
&= Z_t\beta\beta' Z_t + S_{t-1}'\varepsilon_t\beta' Z_t + Z_t\beta\varepsilon_t' S_{t-1} + S_{t-1}'\varepsilon_t\varepsilon_t' S_{t-1}\,.
\end{aligned} \tag{153}$$

with $Z_t = S_{t-1}' S_{t-1}$. Then, using the fact that ε and all third moments of ε have zero

expectations as well as Lemma 10, we have

$$
\begin{aligned}
\frac{1}{T^2}E[F_t'AF_tF_t'AF_t] \;&=\; \frac{1}{T^2}\operatorname{tr}E[F_tF_t'AF_tF_t'A] \\
&= \frac{1}{T^2}\operatorname{tr}E[(Z_t\beta\beta'Z_t + S_{t-1}'\varepsilon_t\beta'Z_t + Z_t\beta\varepsilon_t'S_{t-1} + S_{t-1}'\varepsilon_t\varepsilon_t'S_{t-1})A \\
&(Z_t\beta\beta'Z_t + S_{t-1}'\varepsilon_t\beta'Z_t + Z_t\beta\varepsilon_t'S_{t-1} + S_{t-1}'\varepsilon_t\varepsilon_t'S_{t-1})A] \\
&= \frac{1}{T^2}\operatorname{tr}E[Z_t\beta\beta'Z_tAZ_t\beta\beta'Z_tA] \\
&+ \frac{1}{T^2}2\operatorname{tr}E[Z_t\beta\beta'Z_tAS_{t-1}'\varepsilon_t\varepsilon_t'S_{t-1}A] \\
&+ \frac{1}{T^2}2\operatorname{tr}E[S_{t-1}'\varepsilon_t\beta'Z_tAS_{t-1}'\varepsilon_t\beta'Z_tA] \\
&+ \frac{1}{T^2}2\operatorname{tr}E[S_{t-1}'\varepsilon_t\beta'Z_tAZ_t\beta\varepsilon_t'S_{t-1}A] \\
&+ \frac{1}{T^2}\operatorname{tr}E[S_{t-1}'\varepsilon_t\varepsilon_t'S_{t-1}AS_{t-1}'\varepsilon_t\varepsilon_t'S_{t-1}A] \\
&= \frac{1}{T^2}\operatorname{tr}E[Z_t\beta\beta'Z_tAZ_t\beta\beta'Z_tA] \\
&+ \frac{1}{T^2}2\operatorname{tr}E[Z_t\beta\beta'Z_tAZ_tA] \\
&+ \frac{1}{T^2}2\operatorname{tr}E[Z_tAZ_t\beta\beta'Z_tA] \\
&+ \frac{1}{T^2}2\operatorname{tr}E[(\beta'Z_tAZ_t\beta)Z_tA] \\
&+ \frac{1}{T^2}((\kappa_\varepsilon-1)\operatorname{tr}E[Z_tAZ_tA] \;+\; \operatorname{tr}E[\operatorname{tr}(Z_tA)Z_tA]) \\
&= \frac{1}{T^2}\operatorname{tr}E[Z_t\beta\beta'Z_tAZ_t\beta\beta'Z_tA] \\
&+ \frac{1}{T^2}4\operatorname{tr}E[Z_t\beta\beta'Z_tAZ_tA] \\
&+ \frac{1}{T^2}2\operatorname{tr}E[(\beta'Z_tAZ_t\beta)Z_tA] \\
&+ \frac{1}{T^2}((\kappa_\varepsilon-1)\operatorname{tr}E[Z_tAZ_tA] \;+\; \operatorname{tr}E[\operatorname{tr}(Z_tA)Z_tA]) \\
&= Term1 \;+\; Term2 \;+\; Term3 \;+\; Term4 \;+\; Term5\,,
\end{aligned}
\tag{154}
$$

where in the last term we have used Lemma 10 to show that

$$\begin{aligned}
&\operatorname{tr} E[S'_{t-1}\varepsilon_t\varepsilon'_t S_{t-1}AS'_{t-1}\varepsilon_t\varepsilon'_t S_{t-1}A] \\
&= \operatorname{tr} E[S'_{t-1}\Big((\kappa_\varepsilon-1)(S_{t-1}AS'_{t-1}) \;+\; \operatorname{tr}((S_{t-1}AS'_{t-1}))\Big)S_{t-1}A] \\
&= (\kappa_\varepsilon-1)\operatorname{tr} E[Z_tAZ_tA] \;+\; \operatorname{tr} E[\operatorname{tr}(Z_tA)Z_tA]\,.
\end{aligned} \tag{155}$$

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

Lemma 13 (Newton's identities) *Let A be a matrix with eigenvalues λ_i. Then,*

$$\begin{aligned}
&\sum_{i_1,i_2,i_1\neq i_2}\lambda_{i_1}\lambda_{i_2} \;=\; (\operatorname{tr}A)^2-\operatorname{tr}(A^2) \\
&\sum_{i_1,i_2,i_3\;\; all\;\; different}\lambda_{i_1}\lambda_{i_2}\lambda_{i_3} \;=\; (\operatorname{tr}A)^3-3\operatorname{tr}(A)\operatorname{tr}(A^2)\;+\;2\operatorname{tr}(A^3) \\
&\sum_{i_1,i_2,i_3,i_4\;\; all\;\; different}\lambda_{i_1}\lambda_{i_2}\lambda_{i_3}\lambda_{i_4} \\
&= (\operatorname{tr}A)^4-6(\operatorname{tr}(A))^2\operatorname{tr}(A^2)\;+\;3(\operatorname{tr}(A^2))^2+8(\operatorname{tr}A)(\operatorname{tr}(A^3))-\;6\operatorname{tr}(A^4)\,.
\end{aligned} \tag{156}$$

We also note that Assumption 2 implies

$$\operatorname{tr}(\Sigma^3) \;\le\; \operatorname{tr}(\Sigma^2)\operatorname{tr}(\Sigma) \;=\; o((\operatorname{tr}\Sigma)^3),\;\; \operatorname{tr}(\Sigma^4)\le(\operatorname{tr}(\Sigma^2))^2 \;=\; o((\operatorname{tr}\Sigma)^4) \tag{157}$$

E.1 *Term*1 in (154)

We start with the first term. We have

$$\frac{1}{T^2}\operatorname{tr} E[Z_t\beta\beta'Z_tAZ_t\beta\beta'Z_tA] \;=\; \frac{1}{T^2}E[(\beta'Z_tAZ_t\beta)^2]\,. \tag{158}$$

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

$$\|\tilde{\beta}\|^2 \;=\; \beta'\Psi\beta \;\sim\; b_*\frac{1}{P}\operatorname{tr}(\Psi)\,.$$

Then, setting $\lambda_k = \lambda_k(\tilde{A})$ we get

$$\begin{aligned}
&\frac{1}{T^2}\operatorname{tr} E[Z_t\beta\beta' Z_t A Z_t\beta\beta' Z_t A] \;=\; \frac{1}{T^2}E[(\beta' Z_t A Z_t\beta)^2] \\
&= \frac{1}{T^2}\|\tilde{\beta}\|^4 E[\left(\sum_{i_1,j_1,k_1} X_{i_1,1}\lambda_{i_1}(\Sigma)X_{i_1,k_1}\lambda_{k_1}X_{j_1,k_1}\lambda_{j_1}(\Sigma)X_{j_1,1}\right)^2] \\
&= \frac{1}{T^2}\|\tilde{\beta}\|^4 E[\left(\sum_{i_1,j_1,k_1} X_{i_1,1}\lambda_{i_1}(\Sigma)X_{i_1,k_1}\lambda_{k_1}X_{j_1,k_1}\lambda_{j_1}(\Sigma)X_{j_1,1}\right)^2] \\
&= \frac{1}{T^2}\|\tilde{\beta}\|^4 E[\sum_{i_2,j_2,k_2}\sum_{i_1,j_1,k_1} X_{i_1,1}\lambda_{i_1}(\Sigma)X_{i_1,k_1}\lambda_{k_1}X_{j_1,k_1}\lambda_{j_1}(\Sigma)X_{j_1,1}X_{i_2,1}\lambda_{i_2}(\Sigma)X_{i_2,k_2}\lambda_{k_2}X_{j_2,k_2}\lambda_{j_2}(\Sigma)X_{j_2,1}]
\end{aligned} \tag{159}$$

- First, consider the terms with $k_1 = k_2$ in (159):

$$\frac{1}{T^2}\|\tilde{\beta}\|^4 E[\sum_{i_2,j_2}\sum_{i_1,j_1,k_1} X_{i_1,1}\lambda_{i_1}(\Sigma)X_{i_1,k_1}\lambda_{k_1}X_{j_1,k_1}\lambda_{j_1}(\Sigma)X_{j_1,1}X_{i_2,1}\lambda_{i_2}(\Sigma)X_{i_2,k_1}\lambda_{k_1}X_{j_2,k_1}\lambda_{j_2}(\Sigma)X_{j_2,1}] \tag{160}$$

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

$$
\begin{aligned}
&\|\tilde{\beta}\|^4 \frac{1}{T^2} E[\sum_{i_2,j_2} \sum_{i_1,j_1} X^2_{i_1,1} \lambda_{i_1}(\Sigma) \lambda_1^2 X^2_{j_1,1} \lambda_{j_1}(\Sigma) X^2_{i_2,1} \lambda_{i_2}(\Sigma) X^2_{j_2,1} \lambda_{j_2}(\Sigma)] \\
&= \|\tilde{\beta}\|^4 \frac{1}{T^2} \lambda_1^2 \Bigg(E[\sum_{i_2,j_2,i_1,j_1 \ all \ different} X^2_{i_1,1} \lambda_{i_1}(\Sigma) X^2_{j_1,1} \lambda_{j_1}(\Sigma) X^2_{i_2,1} \lambda_{i_2}(\Sigma) X^2_{j_2,1} \lambda_{j_2}(\Sigma)] \\
&+ E[\sum_{i_2,j_2,i_1,j_1 \ only \ two \ are \ equal} X^2_{i_1,1} \lambda_{i_1}(\Sigma) X^2_{j_1,1} \lambda_{j_1}(\Sigma) X^2_{i_2,1} \lambda_{i_2}(\Sigma) X^2_{j_2,1} \lambda_{j_2}(\Sigma)] \\
&+ E[\sum_{i_2,j_2,i_1,j_1 \ only \ three \ are \ equal} X^2_{i_1,1} \lambda_{i_1}(\Sigma) X^2_{j_1,1} \lambda_{j_1}(\Sigma) X^2_{i_2,1} \lambda_{i_2}(\Sigma) X^2_{j_2,1} \lambda_{j_2}(\Sigma)] \\
&+ E[\sum_{i_2,j_2,i_1,j_1 \ all \ four \ are \ equal} X^2_{i_1,1} \lambda_{i_1}(\Sigma) X^2_{j_1,1} \lambda_{j_1}(\Sigma) X^2_{i_2,1} \lambda_{i_2}(\Sigma) X^2_{j_2,1} \lambda_{j_2}(\Sigma)] \Bigg) \\
&= \|\tilde{\beta}\|^4 \lambda_1^2 \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) \\
&= \|\tilde{\beta}\|^4 \lambda_1^2 \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\Bigg((\operatorname{tr}\Sigma)^4 (\tilde{\beta}' \tilde{A} \tilde{\beta})^2 / (T^2) \Bigg) \ = \ O(1/T^2)
\end{aligned}
\tag{161}
$$

Here, we have used the fact that

$$(\operatorname{tr}\Sigma)^4(\tilde{\beta}'\tilde{A}\tilde{\beta})^2 \;=\; O()$$

because $(\operatorname{tr}\Sigma)^2 b_*$ converges to a finite limit. The rest 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

$$\begin{aligned} &\|\tilde{\beta}\|^4 E[\sum_{i_1=i_2\neq j_1=j_2;k_1} X^2_{i_1,1}\lambda^2_{i_1}(\Sigma)X^2_{i_1,k_1}\lambda^2_{k_1}X^2_{j_1,k_1}\lambda^2_{j_1}(\Sigma)X^2_{j_1,1}] \\ &\sim\; \|\tilde{\beta}\|^4 \operatorname{tr}(\tilde{A}^2)\left((\operatorname{tr}(\Sigma^2))^2 \;-\operatorname{tr}(\Sigma^4)\right) \;\sim\; \|\tilde{\beta}\|^4 \operatorname{tr}(\tilde{A}^2)\,(\operatorname{tr}(\Sigma^2))^2\,, \end{aligned} \tag{162}$$

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

$$\|\tilde{\beta}\|^4 E[X^4] \operatorname{tr}(\tilde{A}^2)\,(\operatorname{tr}(\Sigma^4)) \;=\; O(P)$$

which is negligible.

- We can now focus on the case $k_1 \neq k_2$ in (159). First, consider the terms with $k_1 = 1$. By symmetry, terms with $k_2 = 1$ give the same contribution. Since $k_2 \neq 1$ and

$\|\tilde{\beta}\|^2\lambda_1 = \tilde{\beta}'\tilde{A}\tilde{\beta}$, Newton's identities imply that

$$
\begin{aligned}
&\lambda_1 \frac{1}{T^2}\|\tilde{\beta}\|^4 E[\sum_{i_2,j_2,k_2\neq 1}\sum_{i_1,j_1} X_{i_1,1}^2\lambda_{i_1}(\Sigma)X_{j_1,1}^2\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 \lambda_1 \frac{1}{T^2}\|\tilde{\beta}\|^4 E[\sum_{i_2,k_2}\sum_{i_1,j_1} X_{i_1,1}^2 X_{j_1,1}^2\lambda_{i_1}(\Sigma)\lambda_{j_1}(\Sigma)X_{i_2,1}^2\lambda_{i_2}(\Sigma)^2 X_{i_2,k_2}^2\lambda_{k_2}] \\
&\sim (\tilde{\beta}'\tilde{A}\tilde{\beta})\,\|\tilde{\beta}\|^2\,\frac{1}{T^2}\operatorname{tr}(\tilde{A})\left(E[\sum_{i_2}\sum_{i_1,j_1} X_{i_1,1}^2 X_{j_1,1}^2\lambda_{i_1}(\Sigma)\lambda_{j_1}(\Sigma)X_{i_2,1}^2\lambda_{i_2}(\Sigma)^2]\right) \\
&= (\tilde{\beta}'\tilde{A}\tilde{\beta})\,\|\tilde{\beta}\|^2\,\frac{1}{T^2}\operatorname{tr}(\tilde{A})\left(\sum_{i_2,i_1,j_1\ all\ different}\lambda_{i_1}(\Sigma)\lambda_{j_1}(\Sigma)\lambda_{i_2}(\Sigma)^2\right. \\
&+ \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 \\
&\left.+ E[X^6]\operatorname{tr}(\Sigma^4)\right) \\
&= (\tilde{\beta}'\tilde{A}\tilde{\beta})\,\|\tilde{\beta}\|^2\,\frac{1}{T^2}\operatorname{tr}(\tilde{A})\left(\sum_{i_2}\lambda_{i_2}(\Sigma)^2((\operatorname{tr}(\Sigma)-\lambda_{i_2})^2-(\operatorname{tr}(\Sigma^2)-\lambda_{i_2}^2))\right. \\
&+ E[X^4]((\operatorname{tr}(\Sigma^2))^2-\operatorname{tr}(\Sigma^4)) \\
&+ 2E[X^4]\sum_{i_2}\lambda_{i_2}(\Sigma)^3(\operatorname{tr}(\Sigma)-\lambda_{i_2}) \\
&\left.+ E[X^6]\operatorname{tr}(\Sigma^4)\right) \\
&= (\tilde{\beta}'\tilde{A}\tilde{\beta})\,\|\tilde{\beta}\|^2\,\frac{1}{T^2}\operatorname{tr}(\tilde{A})\left((\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\right. \\
&+ E[X^4]((\operatorname{tr}(\Sigma^2))^2-\operatorname{tr}(\Sigma^4)) \\
&\left.+ 2E[X^4]((\operatorname{tr}\Sigma)(\operatorname{tr}(\Sigma^3))-\operatorname{tr}(\Sigma^4))+ E[X^6]\operatorname{tr}(\Sigma^4)\right)
\end{aligned}
\tag{163}
$$

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

$$O(\|\lambda\|^4 (\mathrm{tr}(\Sigma))^4 \,\mathrm{tr}(\tilde{A})/(T^2)) \;=\; O(P/T^2)$$

and hence are negligible.

- Now, in the case when $k_1 \neq k_2$ and both are different from 1 in (159), 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

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

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

$$\begin{aligned}
&\|\tilde{\beta}\|^4 E[\sum_{i_2,k_2} \sum_{i_1,k_1} X_{i_1,1}^2 X_{i_1,k_1}^2 \lambda_{k_1} \lambda_{i_1}^2(\Sigma) \lambda_{i_2}^2(\Sigma) X_{i_2,1}^2 X_{i_2,k_2}^2 \lambda_{k_2}] \\
&\sim\; \|\tilde{\beta}\|^4 ((\mathrm{tr}\,\tilde{A})^2 - \mathrm{tr}(\tilde{A}^2)) \left(E[\sum_{i_2} \sum_{i_1} X_{i_1,1}^2 \lambda_{i_1}^2(\Sigma) \lambda_{i_2}^2(\Sigma) X_{i_2,1}^2] \right) \\
&=\; \|\tilde{\beta}\|^4 ((\mathrm{tr}\,\tilde{A})^2 - \mathrm{tr}(\tilde{A}^2))((\mathrm{tr}(\Sigma^2))^2 - \mathrm{tr}(\Sigma^4)) \\
&\sim\; \|\tilde{\beta}\|^4 ((\mathrm{tr}\,\tilde{A})^2 - \mathrm{tr}(\tilde{A}^2))(\mathrm{tr}(\Sigma^2))^2
\end{aligned} \tag{164}$$

Summarizing, the dominant terms are (162) (multiplied by 3) and (164), so that

$$
\begin{aligned}
&Term1 \\
&\sim\ 3\|\tilde{\beta}\|^4 \operatorname{tr}(\tilde{A}^2)\,(\operatorname{tr}(\Sigma^2))^2\frac{1}{T^2}\ +\ \|\tilde{\beta}\|^4 E[X^4]\operatorname{tr}(\tilde{A}^2)\,(\operatorname{tr}(\Sigma^4))\frac{1}{T^2} \\
&+\ 2(\tilde{\beta}'\tilde{A}\tilde{\beta})\,\|\tilde{\beta}\|^2\,\frac{1}{T^2}\operatorname{tr}(\tilde{A})\Bigg(\operatorname{tr}(\Sigma^2)(\operatorname{tr}(\Sigma))^2 - 2(\operatorname{tr}\Sigma)(\operatorname{tr}(\Sigma^3)) + 2\operatorname{tr}(\Sigma^4) - (\operatorname{tr}(\Sigma^2))^2 \\
&+\ E[X^4]((\operatorname{tr}(\Sigma^2))^2 - \operatorname{tr}(\Sigma^4)) \\
&+\ 2E[X^4]((\operatorname{tr}\Sigma)(\operatorname{tr}(\Sigma^3)) - \operatorname{tr}(\Sigma^4)) +\ E[X^6]\operatorname{tr}(\Sigma^4)\Bigg)\frac{1}{T^2} \\
&+\ \|\tilde{\beta}\|^4 E[X^4]\,(\operatorname{tr}(\tilde{A})^2 - \operatorname{tr}(\tilde{A}^2))\ \operatorname{tr}(\Sigma^4)\frac{1}{T^2} \\
&+\ \|\tilde{\beta}\|^4((\operatorname{tr}\tilde{A})^2 - \operatorname{tr}(\tilde{A}^2))(\operatorname{tr}(\Sigma^2))^2\frac{1}{T^2} \\
&\sim\ \|\tilde{\beta}\|^4((\operatorname{tr}\tilde{A})^2 + 2\operatorname{tr}(\tilde{A}^2))(\operatorname{tr}(\Sigma^2))^2/(T^2)\ \sim\ \|\tilde{\beta}\|^4(\operatorname{tr}\tilde{A})^2(\operatorname{tr}(\Sigma^2))^2/(T^2)
\end{aligned}
\tag{165}
$$

because $\operatorname{tr}(\tilde{A}^2) = O(P)$.

E.2 *Term2* in (154)

We now proceed with the second term (note that *it comes with a factor of four*). We have

$$
E[\lambda' Z_t A Z_t A Z_t \lambda]\ =\ \|\tilde{\beta}\|^2 E[\sum X_{i_1,1}\lambda_{i_1}(\Sigma)X_{i_1,k_1}\lambda_{k_1}X_{i_2,k_1}\lambda_{i_2}(\Sigma)X_{i_2,k_2}\lambda_{k_2}X_{i_3,k_2}\lambda_{i_3}(\Sigma)X_{i_3,1}]\,. \tag{166}
$$

- Suppose first that $k_1 = k_2 \neq 1$ in (166). The respective contribution is

$$
\|\tilde{\beta}\|^2 E[\sum X_{i_1,1}\lambda_{i_1}(\Sigma)X_{i_1,k_1}\lambda_{k_1}X^2_{i_2,k_1}\lambda_{i_2}(\Sigma)\lambda_{k_1}X_{i_3,k_1}\lambda_{i_3}(\Sigma)X_{i_3,1}]\,, \tag{167}
$$

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

$$\begin{aligned}
&\|\tilde{\beta}\|^2 E[\sum X_{i_1,1}^2 \lambda_{i_1}(\Sigma)^2 X_{i_1,k_1}^2 \lambda_{k_1}^2 X_{i_2,k_1}^2 \lambda_{i_2}(\Sigma)] \\
&= \|\tilde{\beta}\|^2 \left(\sum_{i_1 \neq i_2, k_1 \neq 1} \lambda_{i_1}(\Sigma)^2 \lambda_{k_1}^2 \lambda_{i_2}(\Sigma) \;+\; E[X^4] \sum_{i_1, k_1 \neq 1} \lambda_{i_1}(\Sigma)^3 \lambda_{k_1}^2 \right) \\
&\sim \|\tilde{\beta}\|^2 \operatorname{tr}(\tilde{A}^2)((E[X^4]-1)\operatorname{tr}(\Sigma^3) + \operatorname{tr}(\Sigma)\operatorname{tr}(\Sigma^2)) \;=\; O(P(b_*(\operatorname{tr}\Sigma)^2)\operatorname{tr}\Sigma) \;=\; O(P)
\end{aligned} \tag{168}$$

- The terms with $k_1 = k_2 = 1$ in (166) give

$$\begin{aligned}
&\lambda_1^2 \|\tilde{\beta}\|^2 E[\sum X_{i_1,1}^2 \lambda_{i_1}(\Sigma) X_{i_2,1}^2 \lambda_{i_2}(\Sigma) X_{i_3,1}^2 \lambda_{i_3}(\Sigma)] \\
&\sim \lambda_1^2 \|\tilde{\beta}\|^2 \left(\sum_{i_1,i_2,i_3 \;\; pairwise \;\; different} \lambda_{i_1}(\Sigma)\lambda_{i_2}(\Sigma)\lambda_{i_3}(\Sigma) \right. \\
&\left. + \; 3 \sum_{i_1,i_2 \;\; different} E[X^4]\lambda_{i_1}^2(\Sigma)\lambda_{i_2}(\Sigma) \;+\; E[X^6]\operatorname{tr}(\Sigma^3) \right) \\
&= (\tilde{\beta}'\tilde{A}\tilde{\beta})^2 \|\tilde{\beta}\|^2 \left((\operatorname{tr}\Sigma)^3 - 3(\operatorname{tr}\Sigma)\operatorname{tr}(\Sigma^2) \;+\; 2\operatorname{tr}(\Sigma^3) \right. \\
&\left. + \; 3E[X^4]((\operatorname{tr}\Sigma)\operatorname{tr}(\Sigma^2) - \operatorname{tr}(\Sigma^3)) + E[X^6]\operatorname{tr}(\Sigma^3) \right) \;=\; O(b_*(\operatorname{tr}\Sigma)^2 \operatorname{tr}\Sigma) \;=\; O()
\end{aligned} \tag{169}$$

  by Newton's identities, where $3\sum_{i_1,i_2 \;\; different}$ appears because there are three possibilities for a coincidence of of pair among i_1, i_2, i_3, and where we have used that $\|\tilde{\beta}\|^2 \lambda_1 = \tilde{\beta}'\tilde{A}\tilde{\beta}$.

- For the terms with $k_1 \neq k_2$ and none of them equal to 1 in in (166), we must have

$i_1 = i_2 = i_3$ for them to be non-zero, giving

$$\begin{aligned} \|\tilde{\beta}\|^2 E[\sum X^2_{i_1,1}\lambda_{i_1}(\Sigma)^3 X^2_{i_1,k_1}\lambda_{k_1}X^2_{i_1,k_2}\lambda_{k_2}] \;&\sim\; \|\tilde{\beta}\|^2((\mathrm{tr}(\tilde{A}))^2 - \mathrm{tr}(\tilde{A}^2))\,\mathrm{tr}(\Sigma^3) \\ &=\; o(P^2) \end{aligned} \tag{170}$$

since $((\mathrm{tr}(\tilde{A}))^2 - \mathrm{tr}(\tilde{A}^2)) = O(P^2)$.

- If $k_1 \neq k_2 = 1$ in (166), then we get the contribution

$$\begin{aligned} &\|\tilde{\beta}\|^2 E[\sum X_{i_1,1}\lambda_{i_1}(\Sigma)X_{i_1,k_1}\lambda_{k_1}X_{i_2,k_1}\lambda_{i_2}(\Sigma)X_{i_2,1}\lambda_1\lambda_{i_3}(\Sigma)X^2_{i_3,1}] \\ &= \tilde{\beta}'\tilde{A}\tilde{\beta}E[\sum X_{i_1,1}\lambda_{i_1}(\Sigma)X_{i_1,k_1}\lambda_{k_1}X_{i_2,k_1}\lambda_{i_2}(\Sigma)X_{i_2,1}\lambda_{i_3}(\Sigma)X^2_{i_3,1}] \\ &=\; \{only\ terms\ with\ i_1 = i_2\ survive\} \\ &=\; \tilde{\beta}'\tilde{A}\tilde{\beta}E[\sum X^2_{i_1,1}\lambda^2_{i_1}(\Sigma)X^2_{i_1,k_1}\lambda_{k_1}\lambda_{i_3}(\Sigma)X^2_{i_3,1}] \\ &\sim\; \tilde{\beta}'\tilde{A}\tilde{\beta}\,(\mathrm{tr}\,\tilde{A})\Big(\,\mathrm{tr}(\Sigma)(\mathrm{tr}(\Sigma^2))\;+\;(E[X^4]-1)\,\mathrm{tr}(\Sigma^3)\Big)\;=\;O(Pb_*(\mathrm{tr}(\Sigma))^3)\;=\;O(P) \end{aligned} \tag{171}$$

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

Thus,

$$\begin{aligned} &\frac{1}{4}Term2\;\sim\;\|\tilde{\beta}\|^2\,\mathrm{tr}(\tilde{A}^2)((E[X^4]-1)\,\mathrm{tr}(\Sigma^3) + \mathrm{tr}(\Sigma)\,\mathrm{tr}(\Sigma^2)) \\ &+ \|\tilde{\beta}\|^2((\mathrm{tr}(\tilde{A}))^2 - \mathrm{tr}(\tilde{A}^2))\,\mathrm{tr}(\Sigma^3) \\ &+\;2\tilde{\beta}'\tilde{A}\tilde{\beta}\,(\mathrm{tr}\,\tilde{A})\Big(\,\mathrm{tr}(\Sigma)(\mathrm{tr}(\Sigma^2))\;+\;(E[X^4]-1)\,\mathrm{tr}(\Sigma^3)\Big) \\ &\sim o(T^2)\,. \end{aligned} \tag{172}$$

E.3 *Term*3 **in** (154)

We now proceed with the third term. We have

$$
\begin{aligned}
&2\frac{1}{T^2}E[\mathrm{tr}(AZ_t)\,\lambda' Z_t A Z_t \lambda] \\
&= \; 2\|\tilde{\beta}\|^2\frac{1}{T^2}E[\sum_k \lambda_k(\tilde{A})\sum_i \lambda_i(\Sigma)X_{i,k}^2 \sum_{i_1,k_1,i_2} X_{i_1,1}\lambda_{i_1}(\Sigma)X_{i_1,k_1}\lambda_{k_1}(\tilde{A})X_{i_2,k_1}\lambda_{i_2}(\Sigma)X_{i_2,1}]
\end{aligned} \tag{173}
$$

- First consider the terms with $k_1 = 1$ in (173). This gives

$$
\begin{aligned}
&2\|\tilde{\beta}\|^2\frac{1}{T^2}E[\sum_k \lambda_k(\tilde{A})\sum_i \lambda_i(\Sigma)X_{i,k}^2 \sum_{i_1,i_2} X_{i_1,1}^2\lambda_{i_1}(\Sigma)\lambda_1(\tilde{A})\lambda_{i_2}(\Sigma)X_{i_2,1}^2] \\
&\sim \; 2\frac{1}{T^2}(\tilde{\beta}'\tilde{A}\tilde{\beta})\,(\mathrm{tr}\,\tilde{A})\,(\mathrm{tr}\,\Sigma)\,E[\sum_{i_1,i_2} X_{i_1,1}^2\lambda_{i_1}(\Sigma)\lambda_{i_2}(\Sigma)X_{i_2,1}^2] \\
&= \; 2\frac{1}{T^2}(\tilde{\beta}'\tilde{A}\tilde{\beta})\,(\mathrm{tr}\,\tilde{A})\,(\mathrm{tr}\,\Sigma)((\mathrm{tr}(\Sigma))^2 + (E[X^4]-1)\,\mathrm{tr}(\Sigma^2)) \\
&= \; O(Pb_*(\mathrm{tr}\,\Sigma)^3) \;=\; O(P)
\end{aligned} \tag{174}
$$

  where in the transition from the first to the second line we have used that λ_1 is a negligible fraction of $\mathrm{tr}\,\tilde{A}$.

- If $k_1 \neq 1$ in in (173), the only non-zero terms are with $i_1 = i_2$ and they give

$$
\begin{aligned}
& 2\|\tilde{\beta}\|^2 \frac{1}{T^2} E[\sum_k \lambda_k(\tilde{A}) \sum_i \lambda_i(\Sigma) X_{i,k}^2 \sum_{i_1,k_1 \neq 1} X_{i_1,1}^2 \lambda_{i_1}^2(\Sigma) X_{i_1,k_1}^2 \lambda_{k_1}(\tilde{A})] \\
\sim\ & 2\|\tilde{\beta}\|^2 \frac{1}{T^2} E[\sum_{k \neq 1} \lambda_k(\tilde{A}) \sum_i \lambda_i(\Sigma) X_{i,k}^2 \sum_{i_1,k_1 \neq 1} X_{i_1,1}^2 \lambda_{i_1}^2(\Sigma) X_{i_1,k_1}^2 \lambda_{k_1}(\tilde{A})] \\
=\ & 2\|\tilde{\beta}\|^2 \frac{1}{T^2} E[\sum_{k \neq 1} \lambda_k(\tilde{A}) \sum_i \lambda_i(\Sigma) X_{i,k}^2 \sum_{i_1,k_1 \neq 1} \lambda_{i_1}^2(\Sigma) X_{i_1,k_1}^2 \lambda_{k_1}(\tilde{A})] \\
=\ & 2\|\tilde{\beta}\|^2 \frac{1}{T^2} \Bigg(E[\sum_{k \neq 1} \lambda_k^2(\tilde{A}) \sum_i \lambda_i(\Sigma) X_{i,k}^2 \sum_{i_1} \lambda_{i_1}^2(\Sigma) X_{i_1,k}^2] \\
+\ & E[\sum_{k \neq 1} \lambda_k(\tilde{A}) \sum_{i,k_1 \neq 1,k} \lambda_i(\Sigma) X_{i,k}^2 \sum_{i_1} \lambda_{i_1}^2(\Sigma) X_{i_1,k_1}^2 \lambda_{k_1}(\tilde{A})] \Bigg) \\
=\ & 2\|\tilde{\beta}\|^2 \frac{1}{T^2} \Bigg(E[X^4] \operatorname{tr}(\tilde{A}^2) \operatorname{tr}(\Sigma^3) + \sum_{k \neq 1} \lambda_k^2(\tilde{A}) \sum_i \lambda_i(\Sigma) \sum_{i_1 \neq i} \lambda_{i_1}^2(\Sigma) \\
+\ & \sum_{k \neq 1} \lambda_k(\tilde{A}) \sum_{i,k_1 \neq 1,k} \lambda_i(\Sigma) \sum_{i_1} \lambda_{i_1}^2(\Sigma) \lambda_{k_1}(\tilde{A}) \Bigg) \\
\sim\ & 2\|\tilde{\beta}\|^2 \frac{1}{T^2} \Bigg(\operatorname{tr}(\tilde{A}^2) \Big((E[X^4] - 1) \operatorname{tr}(\Sigma^3) + \operatorname{tr}(\Sigma) \operatorname{tr}(\Sigma^2) \Big) \\
+\ & ((\operatorname{tr} \tilde{A})^2 - \operatorname{tr}(\tilde{A}^2)) \operatorname{tr}(\Sigma) \operatorname{tr}(\Sigma^2) \Bigg) \ \sim\ 2\|\tilde{\beta}\|^2 \frac{1}{T^2} (\operatorname{tr} \tilde{A})^2 \operatorname{tr}(\Sigma) \operatorname{tr}(\Sigma^2) \,.
\end{aligned}
\tag{175}
$$

Thus,

$$
\begin{aligned}
Term3 \sim\ & 2\frac{1}{T^2} (\tilde{\beta}' \tilde{A} \tilde{\beta}) (\operatorname{tr} \tilde{A}) (\operatorname{tr} \Sigma) ((\operatorname{tr}(\Sigma))^2 \\
& + (E[X^4] - 1) \operatorname{tr}(\Sigma^2)) + 2\|\tilde{\beta}\|^2 \frac{1}{T^2} (\operatorname{tr} \tilde{A})^2 \operatorname{tr}(\Sigma) \operatorname{tr}(\Sigma^2) \\
\sim\ & 2\frac{1}{T^2} (\tilde{\beta}' \tilde{A} \tilde{\beta}) (\operatorname{tr} \tilde{A}) (\operatorname{tr} \Sigma)^3 + 2\|\tilde{\beta}\|^2 \frac{1}{T^2} (\operatorname{tr} \tilde{A})^2 \operatorname{tr}(\Sigma) \operatorname{tr}(\Sigma^2) \\
\sim\ & 2\|\tilde{\beta}\|^2 \frac{1}{T^2} (\operatorname{tr} \tilde{A})^2 \operatorname{tr}(\Sigma) \operatorname{tr}(\Sigma^2)
\end{aligned}
\tag{176}
$$

E.4 $Term4$ and $Term5$ in (154)

We have

$$\begin{aligned}
&E[(E[\varepsilon^4]-1)\operatorname{tr}(AZ_tAZ_t) \ + \ (\operatorname{tr}(AZ_t))^2] \\
&= \ (E[\varepsilon^4]-1)E[\sum \lambda_k(\tilde{A})X_{i,k}\lambda_i(\Sigma)X_{i,k_1}\lambda_{k_1}(\tilde{A})X_{i_1,k_1}\lambda_{i_1}(\Sigma)X_{i_1,k}] \\
&+ E[(\sum_k \lambda_k(\tilde{A})\sum_i \lambda_i(\Sigma)X_{i,k}^2)^2]
\end{aligned} \tag{177}$$

We have

$$\begin{aligned}
&E[(\sum_k \lambda_k(\tilde{A})\sum_i \lambda_i(\Sigma)X_{i,k}^2)^2] \\
&= \ E[\sum_{k,k_1,i,i_1} \lambda_k(\tilde{A})\lambda_{k_1}(\tilde{A})\lambda_{i_1}(\Sigma)X_{i_1,k_1}^2\lambda_{i_2}(\Sigma)X_{i_2,k_2}^2] \\
&= \ E[\sum_k \lambda_k^2(\tilde{A})\sum_{i_1,i_2}\lambda_{i_1}\lambda_{i_2}X_{i_1,k}^2X_{i_2,k}^2] \ + \ \sum_{k_1\neq k_2}\lambda_{k_1}(\tilde{A})\lambda_{k_2}(\tilde{A})(\operatorname{tr}(\Sigma))^2 \\
&\sim \ \operatorname{tr}(\tilde{A}^2)((E[X^4]-1)\operatorname{tr}(\Sigma^2)+(\operatorname{tr}\Sigma)^2) \ + \ ((\operatorname{tr}(\tilde{A}))^2-\operatorname{tr}(\tilde{A}^2))(\operatorname{tr}\Sigma)^2
\end{aligned} \tag{178}$$

Similarly,

$$\begin{aligned}
&(E[\varepsilon^4]-1)E[\sum \lambda_k(\tilde{A})X_{i,k}\lambda_i(\Sigma)X_{i,k_1}\lambda_{k_1}(\tilde{A})X_{i_1,k_1}\lambda_{i_1}(\Sigma)X_{i_1,k}] \\
&= \ (E[\varepsilon^4]-1)E[\sum_{k_1=k} \lambda_k(\tilde{A})^2X_{i,k}^2\lambda_i(\Sigma)\lambda_{i_1}(\Sigma)X_{i_1,k}^2] \\
&+ \ (E[\varepsilon^4]-1)E[\sum_{k\neq k_1}\sum_i \lambda_k(\tilde{A})X_{i,k}^2\lambda_i^2(\Sigma)X_{i,k_1}^2\lambda_{k_1}(\tilde{A})] \\
&\sim \ (E[\varepsilon^4]-1)\operatorname{tr}(\tilde{A}^2)((E[X^4]-1)\operatorname{tr}(\Sigma^2)+(\operatorname{tr}\Sigma)^2) \ + \ (E[\varepsilon^4]-1)((\operatorname{tr}(\tilde{A}))^2-\operatorname{tr}(\tilde{A}^2))\operatorname{tr}(\Sigma^2)
\end{aligned} \tag{179}$$

Thus,

$$
\begin{aligned}
Term4 + Term5 \;&\sim\; \operatorname{tr}(\tilde{A}^2)((E[X^4]-1)\operatorname{tr}(\Sigma^2) + (\operatorname{tr}\Sigma)^2) \;+\; ((\operatorname{tr}(\tilde{A}))^2 - \operatorname{tr}(\tilde{A}^2))(\operatorname{tr}\Sigma)^2 \\
&+\; (E[\varepsilon^4]-1)\operatorname{tr}(\tilde{A}^2)((E[X^4]-1)\operatorname{tr}(\Sigma^2) + (\operatorname{tr}\Sigma)^2) \;+\; (E[\varepsilon^4]-1)((\operatorname{tr}(\tilde{A}))^2 - \operatorname{tr}(\tilde{A}^2))\operatorname{tr}(\Sigma^2) \\
&\sim\; (\operatorname{tr}(\tilde{A}^2)(\operatorname{tr}\Sigma)^2 \;+\; ((\operatorname{tr}(\tilde{A}))^2 - \operatorname{tr}(\tilde{A}^2))(\operatorname{tr}\Sigma)^2)\frac{1}{T^2} \\
&+ (E[\varepsilon^4]-1)\Big(\operatorname{tr}(\tilde{A}^2)(\operatorname{tr}\Sigma)^2 + ((\operatorname{tr}(\tilde{A}))^2 - \operatorname{tr}(\tilde{A}^2))\operatorname{tr}(\Sigma^2)\Big)\frac{1}{T^2} \\
&=\; (\operatorname{tr}(\tilde{A}))^2(\operatorname{tr}\Sigma)^2\frac{1}{T^2} \\
&+ (E[\varepsilon^4]-1)\Big(\operatorname{tr}(\tilde{A}^2)(\operatorname{tr}\Sigma)^2 + ((\operatorname{tr}(\tilde{A}))^2 - \operatorname{tr}(\tilde{A}^2))\operatorname{tr}(\Sigma^2)\Big)\frac{1}{T^2} \\
&\sim (\operatorname{tr}(\tilde{A}))^2(\operatorname{tr}\Sigma)^2/(T^2)
\end{aligned}
\tag{180}
$$

because $\operatorname{tr}(\Sigma^2)/(\operatorname{tr}(\Sigma))^2 \to 0$.

E.5 Equating the terms

By (152),

$$
\begin{aligned}
(\frac{1}{T}\operatorname{tr} E[A_P F_t F_t'])^2 \;&\sim \frac{1}{T^2}\operatorname{tr}(\tilde{A})^2(\operatorname{tr}\Sigma + \|\tilde{\beta}\|^2\ \operatorname{tr}(\Sigma^2))^2 \\
&=\; \frac{1}{T^2}\operatorname{tr}(\tilde{A})^2\Big((\operatorname{tr}\Sigma)^2 + 2\|\tilde{\beta}\|^2(\operatorname{tr}\Sigma)\ \operatorname{tr}(\Sigma^2) + \|\tilde{\beta}\|^4\,(\operatorname{tr}(\Sigma^2))^2\Big)
\end{aligned}
\tag{181}
$$

and the claim follows from (165), (172), (176), and (180).

The proof of Lemma 11 is complete. □

F Proof of Theorem 10

Proof of Theorem 10. The first claim follows because, by Lemma 9, the other contributions do not impact eigenvalue distribution.

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

(1) $E[Z_t Z_t'] = A_P$ for some matrix A_P

(2) $E[(Z_t' B Z_t - \mathrm{tr}(A_P B_P))^2] \;=\; o(T^2)$ 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 (3), which in turn follows from Lemma 11. The proof of Theorem 10 is complete. □

G Technical Lemmas for Computing Higher Moments

The following lemma is a direct consequence of (154) and the polarization identity

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

Lemma 14 *Let $Z_t = S_{t-1}' S_{t-1}$. Recall also that*

$$R_{t+1} \;=\; S_t \beta \;+\; \varepsilon_{t+1}, \tag{182}$$

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

$$F_t \;=\; Z_t \beta \;+\; S_{t-1}' \varepsilon_t\,. \tag{183}$$

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

$$\begin{aligned}
&\frac{1}{T}E[F_t'AF_tF_t'BF_t] \\
&= \frac{1}{T}\operatorname{tr} E[Z_t\beta\beta' Z_tAZ_t\beta\beta' Z_tB] \\
&+ \frac{1}{T}2\operatorname{tr}(E[\beta' Z_tAZ_tBZ_t\beta] + E[\beta' Z_tBZ_tAZ_t\beta]) \\
&+ \frac{1}{T}\operatorname{tr}(E[(\beta' Z_tAZ_t\beta)Z_tB] + E[(\beta' Z_tBZ_t\beta)Z_tA]) \\
&+ \frac{1}{T}((\kappa_\varepsilon - 1)\operatorname{tr} E[Z_tAZ_tB] + E[\operatorname{tr}(Z_tA)\operatorname{tr}(Z_tB)]) \\
&= Term1 + Term2 + Term3 + Term4 + Term5\,.
\end{aligned} \tag{184}$$

Proof. When A, B are symmetric, (154) implies

$$\begin{aligned}
&\frac{1}{T}E[F_t'AF_tF_t'BF_t] \\
&= \frac{1}{T}\operatorname{tr} E[Z_t\beta\beta' Z_tAZ_t\beta\beta' Z_tB] \\
&+ \frac{1}{T}2\operatorname{tr}(E[Z_t\beta\beta' Z_tAZ_tB] + E[Z_t\beta\beta' Z_tBZ_tA]) \\
&+ \frac{1}{T}\operatorname{tr}(E[(\beta' Z_tAZ_t\beta)Z_tB] + E[(\beta' Z_tBZ_t\beta)Z_tA]) \\
&+ \frac{1}{T}((\kappa_\varepsilon - 1)\operatorname{tr} E[Z_tAZ_tB] + E[\operatorname{tr}(Z_tA)\operatorname{tr}(Z_tB)])
\end{aligned} \tag{185}$$

The general case follows because

$$
\begin{aligned}
&\frac{1}{T}E[F_t'AF_tF_t'BF_t] \;=\; \frac{1}{T}E[F_t'0.5(A+A')F_tF_t'0.5(B+B')F_t] \\
&=\; \frac{1}{T}\operatorname{tr}E[Z_t\beta\beta'Z_t0.5(A+A')Z_t\beta\beta'Z_t0.5(B+B')] \\
&+\; \frac{1}{T}2\operatorname{tr}(E[Z_t\beta\beta'Z_t0.5(A+A')Z_t0.5(B+B')]+E[Z_t\beta\beta'Z_t0.5(B+B')Z_t0.5(A+A')]) \\
&+\frac{1}{T}\operatorname{tr}(E[(\beta'Z_t0.5(A+A')Z_t\beta)Z_t0.5(B+B')]+E[(\beta'Z_t0.5(B+B')Z_t\beta)Z_t0.5(A+A')]) \\
&+\frac{1}{T}((\kappa_\varepsilon-1)\operatorname{tr}E[Z_t0.5(A+A')Z_t0.5(B+B')] \;+\; E[\operatorname{tr}(Z_t0.5(A+A'))\operatorname{tr}(Z_t0.5(B+B'))]) \\
&=\; \frac{1}{T}\operatorname{tr}E[Z_t\beta\beta'Z_tAZ_t\beta\beta'Z_tB] \\
&+\; \frac{1}{T}\operatorname{tr}(E[\beta'Z_tAZ_tBZ_t\beta]+E[\beta'Z_tBZ_tAZ_t\beta]+E[\beta'Z_tA'Z_tBZ_t\beta]+E[\beta'Z_tAZ_tB'Z_t\beta]) \\
&+\frac{1}{T}\operatorname{tr}(E[(\beta'Z_tAZ_t\beta)Z_tB]+E[(\beta'Z_tBZ_t\beta)Z_tA]) \\
&+\frac{1}{T}((\kappa_\varepsilon-1)0.5\operatorname{tr}(E[Z_tAZ_tB]+E[Z_tA'Z_tB])+\; E[\operatorname{tr}(Z_tA)\operatorname{tr}(Z_tB)])
\end{aligned}
\tag{186}
$$

□

Lemma 15 *For any two matrices A, B, we have*

$$
\begin{aligned}
&\frac{1}{T}\operatorname{tr} E[Z_t\beta\beta' Z_t A Z_t\beta\beta' Z_t B] \\
&\sim \ \Big((\tilde\beta'\tilde A\tilde\beta)\operatorname{tr}(\tilde B) \ + \ (\tilde\beta'\tilde B\tilde\beta)\operatorname{tr}(\tilde A)\Big)\,\|\tilde\beta\|^2 \ \operatorname{tr}(\Sigma^2)(\operatorname{tr}(\Sigma))^2\frac{1}{T} \\
&+ \ \|\tilde\beta\|^4((\operatorname{tr}\tilde A)(\operatorname{tr}\tilde B)+2\operatorname{tr}(\tilde A\tilde B))(\operatorname{tr}(\Sigma^2))^2\frac{1}{T} \\
&+ \|\tilde\beta\|^4 E[X^4]\ \operatorname{tr}(\tilde A)\operatorname{tr}(\tilde B)\operatorname{tr}(\Sigma^4)\frac{1}{T} \\
&\frac{1}{T}2\operatorname{tr}(E[Z_t\beta\beta' Z_t A Z_t B]+E[Z_t\beta\beta' Z_t B Z_t A]) \\
&\sim \ \frac{1}{T}4\|\tilde\beta\|^2\operatorname{tr}(\tilde A\tilde B)\operatorname{tr}(\Sigma)\operatorname{tr}(\Sigma^2) \\
&+ \frac{1}{T}4\|\tilde\beta\|^2(\operatorname{tr}(\tilde A)\operatorname{tr}(\tilde B)-\operatorname{tr}(\tilde A\tilde B))\operatorname{tr}(\Sigma^3) \\
&+ \ \frac{1}{T}4\Big(\tilde\beta'\tilde A\tilde\beta\,(\operatorname{tr}\tilde B)+\tilde\beta'\tilde B\tilde\beta\,(\operatorname{tr}\tilde A)\Big)\operatorname{tr}(\Sigma)(\operatorname{tr}(\Sigma^2)) \\
&\frac{1}{T}\operatorname{tr}(E[(\beta' Z_t A Z_t\beta)Z_t B]+E[(\beta' Z_t B Z_t\beta)Z_t A]) \\
&\sim \ \frac{1}{T}\Big(\tilde\beta'\tilde A\tilde\beta\,(\operatorname{tr}\tilde B)+\tilde\beta'\tilde B\tilde\beta\,(\operatorname{tr}\tilde A)\Big)\,(\operatorname{tr}\Sigma)^3+2\|\tilde\beta\|^2\frac{1}{T}(\operatorname{tr}\tilde A)(\operatorname{tr}\tilde B)\operatorname{tr}(\Sigma)\operatorname{tr}(\Sigma^2) \\
&\frac{1}{T}((\kappa_\varepsilon-1)\operatorname{tr} E[Z_t A Z_t B] \ + \ E[\operatorname{tr}(Z_t A)\operatorname{tr}(Z_t B)]) \\
&\sim \ \Big((\operatorname{tr}\tilde A)(\operatorname{tr}\tilde B) \ + \ (E[\varepsilon^4]-1)\operatorname{tr}(\tilde A\tilde B)\Big)(\operatorname{tr}\Sigma)^2\frac{1}{T}
\end{aligned}
\tag{187}
$$

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

Proof of Lemma 15. Using (165), (172), (176), and (180) , 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\; 3\|\tilde\beta\|^4 \operatorname{tr}(\tilde A\tilde B)\,(\operatorname{tr}(\Sigma^2))^2\frac{1}{T} \;+\; \|\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)\operatorname{tr}(\tilde A)\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}\\
&+\; \|\tilde\beta\|^4 E[X^4]\,(\operatorname{tr}(\tilde A)\operatorname{tr}(\tilde B) - \operatorname{tr}(\tilde A\tilde B))\;\operatorname{tr}(\Sigma^4)\frac{1}{T}\\
&+\; \|\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(\operatorname{tr}(\tilde A)\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 + 2\|\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) \;+\; (E[\varepsilon^4]-1)\operatorname{tr}(\tilde A\tilde B)\Big)(\operatorname{tr}\Sigma)^2\frac{1}{T}\\
&+(E[\varepsilon^4]-1)\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{188}
$$

where we have used that

$$\begin{aligned}
&\Big(\operatorname{tr}(\Sigma^2)(\operatorname{tr}(\Sigma))^2 - 2(\operatorname{tr}\Sigma)(\operatorname{tr}(\Sigma^3)) + 2\operatorname{tr}(\Sigma^4) - (\operatorname{tr}(\Sigma^2))^2 \\
&+ \; E[X^4]((\operatorname{tr}(\Sigma^2))^2 - \operatorname{tr}(\Sigma^4)) \\
&+ \; 2E[X^4]((\operatorname{tr}\Sigma)(\operatorname{tr}(\Sigma^3)) - \operatorname{tr}(\Sigma^4)) + \; E[X^6]\operatorname{tr}(\Sigma^4)\Big) \;\sim\; \operatorname{tr}(\Sigma^2)(\operatorname{tr}(\Sigma))^2
\end{aligned} \tag{189}$$

□

Lemma 16 *Define $\psi_{*,1}$ through the equation*

$$b_*\psi_{*,1} \;=\; \operatorname{tr}((\Sigma_{F,t}\Psi) + \lambda_F'\Psi\lambda_F)\,. \tag{190}$$

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 16. We have

$$\frac{1}{T}\operatorname{tr} E[\beta\beta' F_{t_1}F_{t_1}'F_{t_1}F_{t_1}'Q] \;=\; \frac{1}{T}\operatorname{tr} E[F_{t_1}'F_{t_1}F_{t_1}'Q\beta\beta'F_{t_1}] \tag{191}$$

and hence we are in a position to apply Lemmas 14 and 15 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, (191) 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} \\
&+ \ \|\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 \\
&+ 2\|\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}((E[\varepsilon^4]-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}\Psi)(\operatorname{tr}\Psi^{1/2}Q\beta\beta'\Psi^{1/2}) \ + \ (E[\varepsilon^4]-1)\operatorname{tr}(\Psi\Psi^{1/2}Q\beta\beta'\Psi^{1/2})\Big)(\operatorname{tr}\Sigma)^2\frac{1}{T}
\end{aligned}
\tag{192}
$$

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

$$
\begin{aligned}
&\frac{1}{T}\operatorname{tr} E[Z_t\beta\beta' Z_t A Z_t\beta\beta' Z_t B] \\
&\sim\ (\tilde\beta'\Psi^{1/2}Q\beta\beta'\Psi^{1/2}\tilde\beta)\operatorname{tr}(\Psi)\,\|\tilde\beta\|^2\ \operatorname{tr}(\Sigma^2)(\operatorname{tr}(\Sigma))^2\frac{1}{T} \\
&\frac{1}{T}2\operatorname{tr}(E[Z_t\beta\beta' Z_t A Z_t B]+E[Z_t\beta\beta' Z_t B Z_t A]) \\
&\sim\ \frac{1}{T}4\tilde\beta'\Psi^{1/2}Q\beta\beta'\Psi^{1/2}\tilde\beta\,(\operatorname{tr}\Psi)\operatorname{tr}(\Sigma)(\operatorname{tr}(\Sigma^2)) \\
&\frac{1}{T}\operatorname{tr}(E[(\beta' Z_t A Z_t\beta)Z_t B]+E[(\beta' Z_t B Z_t\beta)Z_t A]) \\
&\sim\ \frac{1}{T}\tilde\beta'\Psi^{1/2}Q\beta\beta'\Psi^{1/2}\tilde\beta\,(\operatorname{tr}\Psi)(\operatorname{tr}\Sigma)^3 \\
&\frac{1}{T}((\kappa_\varepsilon-1)\operatorname{tr} E[Z_t A Z_t B]\ +\ E[\operatorname{tr}(Z_t A)\operatorname{tr}(Z_t B)]) \\
&\sim\ (\operatorname{tr}\Psi)(\operatorname{tr}\Psi^{1/2}Q\beta\beta'\Psi^{1/2})(\operatorname{tr}\Sigma)^2\frac{1}{T}
\end{aligned}
\tag{193}
$$

Recall that $b_* = \operatorname{tr} E[\beta\beta'] = \operatorname{tr}((\Sigma_{F,t}\Psi) + \lambda'\lambda_F))$. 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

$$
\begin{aligned}
&\frac{1}{T}\tilde\beta'\Psi^{1/2}Q\beta\beta'\Psi^{1/2}\tilde\beta\,(\operatorname{tr}\Psi)(\operatorname{tr}\Sigma)^3+(\operatorname{tr}\Psi)(\operatorname{tr}\Psi^{1/2}Q\beta\beta'\Psi^{1/2})(\operatorname{tr}\Sigma)^2\frac{1}{T} \\
&\sim\frac{1}{T}b_*\psi_{*,1}\operatorname{tr}(\Psi)\,(\operatorname{tr}(\Sigma))^3 E[\beta'\Psi Q\beta]\ +\ (\operatorname{tr}\Psi)E[\beta'\Psi Q\beta](\operatorname{tr}\Sigma)^2\frac{1}{T}
\end{aligned}
\tag{194}
$$

where we have used that, by Lemma 5, $\beta'\Psi^{1/2}\tilde\beta \approx \operatorname{tr}((\Sigma_{F,t}\Psi) + \lambda'\lambda_F))$ The proof of Lemma 16 is complete. □

H The Martingale Lemma and $\xi(z;c)$

We start with the following Lemma from KMZ.

Lemma 17 *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 18 *Let*

$$\frac{1}{T}\operatorname{tr}((zI+B_T)^{-1}\Psi\sigma_*) \;\to\; \xi(z;c) \tag{195}$$

almost surely and

$$\frac{1}{T}F_t'(zI+B_{T,t})^{-1}F_t \;\to\; \xi(z;c)\,, \tag{196}$$

in probability, where

$$\frac{c^{-1}\xi(z;c)}{1+\xi(z;c)} \;=\; 1 \;-\; m(-z;c)z \tag{197}$$

Proof. First, Lemma 11 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 17 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 7,

$$\begin{aligned} E[F_t F_t'] &= ((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2))\Psi\Sigma_F\Psi \\ &+ \operatorname{tr}(\Sigma^2)(\kappa-2)\Psi^{1/2}\operatorname{diag}(\Psi^{1/2}\Sigma_F\Psi^{1/2})\Psi^{1/2} + \Psi\Big(\operatorname{tr}(\Sigma\Sigma_\varepsilon) + \operatorname{tr}(\Psi\Sigma_F)\operatorname{tr}(\Sigma^2)\Big) \end{aligned} \tag{198}$$

We have

$$\frac{1}{T}\operatorname{tr}((zI + B_{T,t})^{-1}(\operatorname{tr}\Sigma)^2\Psi\Sigma_{F,t}\Psi) = O(1/T) \tag{199}$$

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

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

is also uniformly bounded. Thus, we get

$$\begin{aligned} &\frac{1}{T}F_t'(zI + B_{T,t})^{-1}F_t \sim \frac{1}{T}\operatorname{tr}((zI + B_{T,t})^{-1}E[F_tF_t']) \\ &\sim T^{-1}\operatorname{tr}[(zI + B_{T,t})^{-1}\Psi\sigma_*] \to \xi(z;c)\,. \end{aligned} \tag{200}$$

Now, we have

$$\begin{aligned} 1 &= P^{-1}\operatorname{tr}E[(zI + B_T)^{-1}(zI + B_T)] \\ &= zm(-z;c) + \frac{1}{P}\operatorname{tr}\frac{1}{T}\sum_t E[(zI + B_T)^{-1}F_tF_t'] \\ &= zm(-z;c) + \frac{1}{P}\operatorname{tr}E[(zI + B_T)^{-1}F_tF_t'] \end{aligned} \tag{201}$$

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 (200).[43] □

I Expected Return on the Feasible Portfolio

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

Proposition 11 *We have*

$$E[R^F_{t+1}(z)] \;=\; \frac{\Gamma_{1,1}(z)}{1+\xi(z;c)}\,, \tag{202}$$

where

$$\Gamma_{1,1}(z) \;=\; \lim_{T,P\to\infty} \lambda' E[\Psi(zI + B_T)^{-1}\Psi]\lambda\,. \tag{203}$$

[43] 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.

Proof of Proposition 11. We start by computing

$$E[F_{t+1}] \;=\; E[S_t' R_{t+1}] = E[S_t'(S_t \widetilde{F}_{t+1} + \varepsilon_{t+1})] \;=\; \mathrm{tr}(\Sigma)\lambda_F \tag{204}$$

and therefore, by (125), we have

$$\begin{aligned} &E[R^F_{t+1}(z)] \;=\; E[\hat{\beta}(z)' F_{t+1}] \\ &= \; \mathrm{tr}(\Sigma) E[\frac{1}{T}\sum_t F_t'(zI + B_T)^{-1}]\lambda_F \;\sim\; E[\frac{1}{T}\sum_t F_t'(zI + B_T)^{-1}]\lambda_F \,, \end{aligned} \tag{205}$$

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

$$\begin{aligned} &E[\frac{1}{T}\sum_t F_t'(zI + B_T)^{-1}]\lambda_F \\ &= \; E[F_t'(zI + B_T)^{-1}\Psi]\lambda \;=\; E[F_t'(zI + B_{T,t})^{-1}\frac{1}{1 + (T)^{-1}F_t'(zI + B_{T,t})^{-1}F_t}\Psi]\lambda \,, \end{aligned} \tag{206}$$

where

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

By Lemma 18,

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

is probability and therefore

$$E[F_t'(zI + B_{T,t})^{-1}\frac{1}{1 + (T)^{-1}F_t'(zI + B_{T,t})^{-1}F_t}\Psi]\lambda \;\sim\; \frac{E[F_t'(zI + B_{T,t})^{-1}\lambda_F]}{1 + \xi(z;c)} \,, \tag{207}$$

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

$$E[F_t'(zI + B_{T,t})^{-1}\lambda_F] \;=\; \mathrm{tr}(\Sigma\Sigma_\varepsilon)\lambda' E[\Psi(zI + B_{T,t})^{-1}\lambda_F] \;\sim\; \Gamma_{1,1}(z)\,. \tag{208}$$

The proof of Proposition 11 is complete.

□

J Computing the Quasi-Moments

Lemma 19 *Let*

$$\psi_{*,k} \;=\; \lim P^{-1}\,\mathrm{tr}(\Psi^k\Sigma_\lambda) \tag{209}$$

and

$$\Gamma_{k,l,T}(z) \;\equiv\; \lambda' E[\Psi^k(zI + B_T)^{-1}\Psi^\ell]\lambda\,. \tag{210}$$

We have

$$\psi_{*,k+\ell} \sim\; z\,\Gamma_{k,\ell,T}(z) \;+\; \left(\psi_{*,k+1}\Gamma_{1,\ell,T}(z) \;+\; \sigma_*\Gamma_{k+1,\ell,T}\right)(1 + \xi(z;c))^{-1} \tag{211}$$

Proof of Lemma 19. Using the Sherman-Morrison formula and Lemma 18, 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

$$\begin{aligned} E[F_t F_t'] \;&=\; ((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2/))\Psi\Sigma_F\Psi \\ &+\; \operatorname{tr}(\Sigma^2/)(\kappa-2)\Psi^{1/2}\operatorname{diag}(\Psi^{1/2}\Sigma_F\Psi^{1/2})\Psi^{1/2} + \Psi\Big(\operatorname{tr}(\Sigma\Sigma_\varepsilon) + \operatorname{tr}(\Psi\Sigma_F)\operatorname{tr}(\Sigma^2)\Big) \qquad (212) \\ &=\; \widehat{\Sigma}_F \;+ \Psi\Sigma_F\Psi \;+\; \sigma_*\Psi\,, \end{aligned}$$

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

$$\Sigma_F \;=\; \lambda\lambda' \;+\; \Sigma_F^*\,. \qquad (213)$$

We will need the following important observation:

Lemma 20 *For any sequence*

$$\lambda' A_P Q_P \lambda \;\to\; 0 \qquad (214)$$

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

$$\begin{aligned} &\lambda' A_P Q_P \lambda \;=\; \operatorname{tr}(\lambda\lambda' A_P Q_P) \\ &\le\; \|\lambda\lambda' A_P Q_P\|_1 \;\le\; \|Q_P\|_\infty \|\lambda\lambda' A_P\|_1 \\ &=\; \|Q_P\|_\infty \operatorname{tr}((\lambda\lambda' A_P A_P'\lambda\lambda')^{1/2}) \;=\; \|Q_P\|_\infty (\lambda' A_P A_P'\lambda)^{1/2}\operatorname{tr}((\lambda\lambda')^{1/2}) \;=\; (\lambda' A_P A_P'\lambda)^{1/2}\|\lambda\| \\ &=\; (\operatorname{tr}(A_P A_P'\lambda\lambda'))^{1/2}\|\lambda\| \;\to\; (P^{-1}\operatorname{tr}(\Sigma_\lambda))^{1/2}(P^{-1}\operatorname{tr}(A_P A_P'\Sigma_\lambda))^{1/2} \\ &\le\; (P^{-1}\operatorname{tr}(\Sigma_\lambda))^{1/2}\|\Sigma_\lambda\|^{1/2}(P^{-1}\operatorname{tr}(A_P A_P'))^{1/2} \;\to\; 0 \end{aligned} \qquad (215)$$

The proof of Lemma 20 is complete. □

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

$$\begin{aligned}
\psi_{*,k+\ell} &= P^{-1}\operatorname{tr}(\Psi^{k+\ell}\Sigma_\lambda) \approx \lambda'\Psi^{k+\ell}\lambda = \lambda' E[\Psi^k(zI+B_T)(zI+B_T)^{-1}\Psi^\ell]\lambda \\
&= z\Gamma_{k,\ell,T}(z) + \lambda' E[\Psi^k B_T(zI+B_T)^{-1}\Psi^\ell]\lambda \\
&\underbrace{=}_{\text{symmetry over } t} z\Gamma_{k,\ell,T}(z) + \lambda' E[\Psi^k F_t F_t'(zI+B_T)^{-1}\Psi^\ell]\lambda \\
&\underbrace{=}_{(80)} z\Gamma_{k,\ell,T}(z) + \lambda' E[\Psi^k F_t F_t'(zI+B_{T,t})^{-1}(1+(T)^{-1}F_t'(zI+B_{T,t})^{-1}F_t)^{-1}\Psi^\ell]\lambda \\
&\underbrace{\sim}_{\text{Lemma 18}} z\Gamma_{k,\ell,T}(z) + \lambda' E[\Psi^k F_t F_t'(zI+B_{T,t})^{-1}\Psi^\ell]\lambda(1+\xi(z;c))^{-1} \\
&\underbrace{\sim}_{(212)} z\Gamma_{k,\ell,T}(z) + \lambda' E[\Psi^k(\widehat{\Sigma}_F + \Psi\Sigma_F\Psi + \sigma_*\Psi)(zI+B_{T,t})^{-1}\Psi^\ell]\lambda(1+\xi(z;c))^{-1} \qquad (216) \\
&\sim z\Gamma_{k,\ell,T}(z) + \lambda' E[\Psi^k(\Psi(\Sigma_F+\lambda\lambda')\Psi + \sigma_*\Psi)(zI+B_T)^{-1}\Psi^\ell]\lambda(1+\xi(z;c))^{-1} \\
&\underbrace{\sim}_{(214)} z\Gamma_{k,\ell,T}(z) + \lambda' E[\Psi^k(\lambda_F\lambda_F' + \sigma_*\Psi)(zI+B_T)^{-1}\Psi^\ell]\lambda(1+\xi(z;c))^{-1} \\
&= z\Gamma_{k,\ell,T}(z) + \lambda'\Psi^{k+1}\lambda E[\lambda_F'(zI+B_T)^{-1}\Psi^\ell]\lambda(1+\xi(z;c))^{-1} \\
&+ \lambda_F'^{k+1}\sigma_*(zI+B_T)^{-1}\Psi^\ell\lambda(1+\xi(z;c))^{-1} \\
&\sim z\,\Gamma_{k,\ell,T}(z) + \left(\psi_{*,k+1}\Gamma_{1,\ell,T}(z) + \sigma_*\Gamma_{k+1,\ell,T}\right)(1+\xi(z;c))^{-1}
\end{aligned}$$

□

Lemma 21 *Let*

$$\delta(z) = -\sigma_* z^{-1}(1+\xi(z;c))^{-1}. \qquad (217)$$

Then,

$$\Gamma_{1,l}(z) = \frac{z^{-1}P^{-1}\operatorname{tr}(\Psi^{1+\ell}(I-\Psi\delta(z))^{-1}\Sigma_\lambda)}{1-\delta(z)P^{-1}\operatorname{tr}(\Psi^2(I-\Psi\delta(z))^{-1}\Sigma_\lambda)} \qquad (218)$$

and

$$\Gamma_{k,\ell} \;=\; z^{-1}P^{-1}\,\mathrm{tr}(\Psi^{k+\ell}(I-\Psi\delta(z))^{-1}\Sigma_\lambda)-z^{-1}P^{-1}\,\mathrm{tr}(\Psi^{k+1}(I-\Psi\delta(z))^{-1}\Sigma_\lambda)\Gamma_{1,\ell}(1+\xi(z;c))^{-1} \tag{219}$$

Proof. We have

$$\Gamma_{k,\ell} \;=\; a_{k+1}+\delta\,\Gamma_{k+1,\ell} \tag{220}$$

where

$$a_{k+1,\ell} \;=\; z^{-1}(\psi_{*,k+\ell}-\psi_{*,k+1}\Gamma_{1,\ell}(1+\xi(z;c))^{-1}),\;\delta(z) \;=\; -\sigma_* z^{-1}(1+\xi(z;c))^{-1}\,. \tag{221}$$

Let us pick $z>\max(1,\|\Psi\|)$ sufficiently large, so that $\sigma_* z^{-1}(1+\xi(z;c))^{-1}<1$ and[44]

$$|\delta^k\Gamma_{k,\ell}(z)| \;\le\; z^{-k+1}\|\lambda\|^2\|\Psi\|^{k+\ell} \;\to_{k\to\infty}\; 0\,. \tag{222}$$

Then, since iterating forward, we get

$$\Gamma_{k,\ell} \;=\; \sum_{\tau=0}^{\infty} a_{k+\tau+1,\ell}\delta^\tau\,. \tag{223}$$

Now,

$$a_{k+\tau+1,\ell} \;=\; z^{-1}(\psi_{*,k+\tau+\ell}-\psi_{*,k+\tau+1}\Gamma_{1,\ell}(1+\xi(z;c))^{-1}),\;\delta(z) \;=\; -\sigma_* z^{-1}(1+\xi(z;c))^{-1}\,. \tag{224}$$

[44]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.

$$
\begin{aligned}
\Gamma_{1,\ell} &= \sum_{\tau=0}^{\infty} a_{\tau+2,\ell}\delta^{\tau} \\
&= \sum_{\tau=0}^{\infty} z^{-1}(\psi_{*,1+\tau+\ell} - \psi_{*,1+\tau+1}\Gamma_{1,\ell}(1+\xi(z;c))^{-1})\delta^{\tau} \\
&= \sum_{\tau=0}^{\infty} (z^{-1}(P^{-1}\operatorname{tr}(\Psi^{\tau+\ell+1}\Sigma_{\lambda}) - P^{-1}\operatorname{tr}(\Psi^{\tau+2}\Sigma_{\lambda})\Gamma_{1,\ell}(1+\xi(z;c))^{-1}))\delta^{\tau} \\
&= z^{-1}P^{-1}\operatorname{tr}(\Psi^{1+\ell}(I-\Psi\delta(z))^{-1}\Sigma_{\lambda}) - z^{-1}P^{-1}\operatorname{tr}(\Psi^{2}(I-\Psi\delta(z))^{-1}\Sigma_{\lambda})\Gamma_{1,\ell}(1+\xi(z;c))^{-1},
\end{aligned}
\tag{225}
$$

implying that

$$
\Gamma_{1,l} = \frac{z^{-1}P^{-1}\operatorname{tr}(\Psi^{1+\ell}(I-\Psi\delta(z))^{-1}\Sigma_{\lambda})}{1-\delta(z)P^{-1}\operatorname{tr}(\Psi^{2}(I-\Psi\delta(z))^{-1}\Sigma_{\lambda})} \tag{226}
$$

Then, the same argument implies

$$
\Gamma_{k,\ell} = z^{-1}P^{-1}\operatorname{tr}(\Psi^{k+\ell}(I-\Psi\delta(z))^{-1}\Sigma_{\lambda}) - z^{-1}P^{-1}\operatorname{tr}(\Psi^{k+1}(I-\Psi\delta(z))^{-1}\Sigma_{\lambda})\Gamma_{1,\ell}(1+\xi(z;c))^{-1} \tag{227}
$$

Furthermore,

$$
\delta(z) = -\sigma_{*}z^{-1}(1+\xi(z;c))^{-1}, \tag{228}
$$

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

$$
\Gamma_{1,1}(z) \approx \frac{z^{-1}\tilde{\lambda}'(I-\Psi\delta(z))^{-1}\tilde{\lambda}}{1-\delta(z)\tilde{\lambda}'(I-\Psi\delta(z))^{-1}\tilde{\lambda}} \tag{229}
$$

□

K Proof of Theorem 4: Second Moment of the Feasible Efficient Portfolio

We start with

Lemma 22 *We have*

$$G(z;c) \;=\; \frac{d}{dz}(z\xi(z;c)) \;\in (0, cz^{-2}] \tag{230}$$

satisfies

$$G(z;c) \;=\; \mathcal{M}(z; Z_*(z;c))\,, \tag{231}$$

where

$$\mathcal{M}(z;Z) \;=\; -1 \;+\; \frac{Z}{z + c\phi(Z)Z^2},\;\; \phi(z) \;=\; P^{-1}\,\mathrm{tr}(E[FF'](zI + E[FF'])^{-2})\,. \tag{232}$$

Proof of Lemma 22. By the master equation,

$$m(z;c) = \frac{1}{1 \;-\; c \;-\; c\,z\,m(z;c)}\, m_{\sigma_*\Psi}\left(\frac{z}{1 \;-\; c \;-\; c\,z\,m(z;c)}\right)\,. \tag{233}$$

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

$$\frac{c^{-1}\xi(z;c)}{1+\xi(z;c)} \;=\; 1 \;-\; m(-z;c)z\,. \tag{234}$$

and hence

$$\xi(z;c) \;=\; \frac{1 - zm(-z;c)}{c^{-1} - 1 + zm(-z;c)} \tag{235}$$

and hence

$$1+\xi(z;c) \;=\; \frac{c^{-1}}{c^{-1}-1+zm(-z;c)} \;=\; \frac{1}{1-c+czm(-z;c)} \tag{236}$$

Differentiating this identity, we get

$$\xi'(z;c) \;=\; -c(zm(-z;c))'(1+\xi(z;c))^2 \tag{237}$$

Furthermore, differentiating the identity

$$zm(-z;c) \;=\; Z_*(z;c)m(-Z_*(z;c)), \tag{238}$$

we get

$$(zm(-z;c))' \;=\; (zm(-z))'(Z_*)Z'_* \;=\; (zm(-z))'(Z_*)(1+\xi(z;c)+z\xi'(z;c)) \tag{239}$$

so that

$$\xi'(z;c) \;=\; -c(zm(-z))'(Z_*)(1+\xi(z;c)+z\xi'(z;c))(1+\xi(z;c))^2\,, \tag{240}$$

implying that

$$\xi'(z;c) \;=\; \frac{-c(zm(-z))'(Z_*)(1+\xi(z;c))^3}{1+c(zm(-z))'(Z_*)z(1+\xi(z;c))^2} \tag{241}$$

and hence

$$1+\xi(z;c)+z\xi'(z;c) \;=\; \frac{1+\xi(z;c)}{1+c(zm(-z))'(Z_*)z(1+\xi(z;c))^2} \;=\; \frac{Z_*(z;c)}{z+c(zm(-z))'(Z_*)Z_*^2(z;c)} \tag{242}$$

□

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 7, we get[45]

$$\begin{aligned}
&E[(R^F_{t+1}(z))^2] \;=\; E[\frac{1}{T}\overline{F_t}'(zI+B_T)^{-1}F_{t+1}F_{t+1}'(zI+B_T)^{-1}\frac{1}{T}\overline{F_t}]\\
&=\; E[\frac{1}{T}\overline{F_t}'(zI+B_T)^{-1}E_{t-}[F_{t+1}F_{t+1}'](zI+B_T)^{-1}\frac{1}{T}\overline{F_t}]\\
&\underbrace{=}_{Lemma\ 7}\; E[\frac{1}{T}\overline{F_t}'(zI+B_T)^{-1}\bigg(((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))\Psi\Sigma_F\Psi+\Psi\Big(\operatorname{tr}(\Sigma\Sigma_\varepsilon)+\operatorname{tr}(\Psi\Sigma_F)\operatorname{tr}(\Sigma^2)\Big)\bigg)\\
&(zI+B_T)^{-1}\frac{1}{T}\overline{F_t}]\\
&\approx\; E[\frac{1}{T}\overline{F_t}'(zI+B_T)^{-1}\bigg((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)\bigg)\\
&(zI+B_T)^{-1}\frac{1}{T}\overline{F_t}]\\
&=\; \frac{1}{T^2}\sum_{t_1,t_2}E[F_{t_1}(zI+B_T)^{-1}\bigg((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_T)^{-1}F_{t_2}]\\
&\sim\; Term1\;+\;Term2
\end{aligned}\tag{243}$$

with

$$Term1 \;=\; \frac{1}{T}E[F_{t_1}'(zI+B_T)^{-1}\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_T)^{-1}F_{t_1}]\tag{244}$$

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

and

$$Term2 \;=\; \frac{T(T-1)}{T^2} E[F_{t_1}'(zI+B_T)^{-1}\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_T)^{-1}F_{t_2}] \tag{245}$$

for any $t_1 \neq t_2$.

K.1 $Term1$ in (244)

We first deal with the first term. Using the Sherman-Morrison formula and Lemma 18, and Lemma 7, we get

$$\begin{aligned} Term1 \;&=\; \frac{1}{T}\operatorname{tr} E[\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_T)^{-1}F_{t_1}F_{t_1}'(zI+B_T)^{-1}] \\ &\sim\; \frac{1}{T}\operatorname{tr} E[\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_{T,t_1})^{-1}F_{t_1}F_{t_1}'(zI+B_{T,t_1})^{-1}](1+\xi(z;c))^{-2} \\ &\sim \frac{1}{T}\operatorname{tr} E[\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_{T,t})^{-1} \\ &\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_{T,t_1})^{-1}](1+\xi(z;c))^{-2} \end{aligned} \tag{246}$$

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

$$\begin{aligned} &\frac{1}{T}\operatorname{tr} E[(\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi(zI+B_{T,t})^{-1}(\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi(zI+B_{T,t})^{-1}](1+\xi(z;c))^{-2} \\ &=\; \frac{1}{T}\operatorname{tr} E[\Psi\Sigma_F\Psi(zI+B_{T,t})^{-1}\Psi\Sigma_F\Psi(zI+B_{T,t})^{-1}](1+\xi(z;c))^{-2} \;\to 0 \end{aligned} \tag{247}$$

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

$$\frac{1}{T}\operatorname{tr} E[(\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi(zI+B_{T,t})^{-1}\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Psi(zI+B_{T,t})^{-1}]/(1+\xi(z;c))^2 \;=\; O(T^{-1}) \tag{248}$$

by the same argument. Finally, the last term is

$$(\operatorname{tr}(\Sigma\Sigma_\varepsilon))^2\frac{1}{T}\operatorname{tr} E[\Psi(zI+B_{T,t})^{-1}\Psi(zI+B_{T,t})^{-1}]/(1+\xi(z;c))^2 \tag{249}$$

and it needs to be evaluated directly.

Lemma 23 *We have*

$$\begin{aligned}
&\frac{1}{P}\operatorname{tr} E[F_{t_1}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_2}F_{t_2}'(zI+B_{T,t_1,t_2})^{-1}]\\
&\sim\; \sigma_*^2\frac{1}{P}\operatorname{tr} E[\Psi(zI+B_T)^{-1}\Psi(zI+B_T)^{-1}]\\
&\to\; \Gamma_3(z) \;=\; \left(1-(-z^2m'(-z;c)+2zm(-z;c)+c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2)\right)(1+\xi(z;c))^4\\
&=\; c^{-1}(\xi(z;c)\;+\;z\xi'(z;c))(1+\xi(z;c))^2
\end{aligned} \tag{250}$$

Proof. We have by the Sherman-Morrison formula that

$$\begin{aligned}
&\frac{1}{P}\frac{1}{T}\operatorname{tr} E[F_{t_1}F_{t_1}'(zI+B_T)^{-1}F_{t_1}F_{t_1}'(zI+B_T)^{-1}]\\
&\sim \frac{1}{c}\frac{1}{T^2}E[F_{t_1}'(zI+B_T)^{-1}F_{t_1}F_{t_1}'(zI+B_T)^{-1}F_{t_1}]\\
&=\; c^{-1}E\left[\left(\frac{\frac{1}{T}F_{t_1}'(zI+B_{T,t_1})^{-1}F_{t_1}}{1+\frac{1}{T}F_{t_1}'(zI+B_{T,t_1})^{-1}F_{t_1}}\right)^2\right]\\
&\sim\; c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2
\end{aligned} \tag{251}$$

by Lemma 18. Now,

$$m'(-z;c) \;=\; \lim P^{-1}\operatorname{tr} E[(zI+B_T)^{-2}] \tag{252}$$

and hence

$$
\begin{aligned}
1 \;&=\; \frac{1}{P}\operatorname{tr} E[(zI+B_T)(zI+B_T)^{-1}(zI+B_T)(zI+B_T)^{-1}] \\
&=\; \frac{1}{P}z^2 \operatorname{tr} E[(zI+B_T)^{-2}] \;+\; 2z\,\frac{1}{P}\operatorname{tr} E[(zI+B_T)^{-2}B_T] \\
&+\; \frac{1}{P}\operatorname{tr} E[B_T(zI+B_T)^{-1}B_T(zI+B_T)^{-1}] \\
&\sim\; z^2 m'(-z;c) \;+\; 2z\,\frac{1}{P}\operatorname{tr} E[(zI+B_T)^{-2}(B_T+zI-zI)] \\
&+\; \frac{1}{P}\frac{1}{T^2}\sum_{t_1,t_2}\operatorname{tr} E[F_{t_1}F_{t_1}'(zI+B_T)^{-1}F_{t_2}F_{t_2}'(zI+B_T)^{-1}] \\
&=\; -z^2 m'(-z;c) + 2zm(-z;c) + \frac{1}{P}\frac{1}{T}\operatorname{tr} E[F_{t_1}F_{t_1}'(zI+B_T)^{-1}F_{t_1}F_{t_1}'(zI+B_T)^{-1}] \\
&+\; \frac{1}{P}\frac{T(T-1)}{T^2}\operatorname{tr} E[F_{t_1}F_{t_1}'(zI+B_T)^{-1}F_{t_2}F_{t_2}'(zI+B_T)^{-1}] \\
&\sim\; -z^2 m'(-z;c) + 2zm(-z;c) + c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2 \qquad (253) \\
&+\; \frac{1}{P}\operatorname{tr} E[F_{t_1}F_{t_1}'(zI+B_T)^{-1}F_{t_2}F_{t_2}'(zI+B_T)^{-1}] \\
&\sim\; -z^2 m'(-z;c) + 2zm(-z;c) + c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2 \\
&+\; \frac{1}{P}\operatorname{tr} E[F_{t_1}F_{t_1}'(zI+B_{T,t_1})^{-1}F_{t_2}F_{t_2}'(zI+B_{T,t_2})^{-1}]/(1+\xi(z;c))^2 \\
&\sim\; -z^2 m'(-z;c) + 2zm(-z;c) + c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2 \\
&+\; \frac{1}{P}E[F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_2}F_{t_2}'(zI+B_{T,t_1,t_2})^{-1}F_{t_1}]/(1+\xi(z;c))^4 \\
&=\; -z^2 m'(-z;c) + 2zm(-z;c) + c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2 \\
&+\; \frac{1}{P}\operatorname{tr} E[F_{t_1}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_2}F_{t_2}'(zI+B_{T,t_1,t_2})^{-1}]/(1+\xi(z;c))^4
\end{aligned}
$$

where we have defined

$$
B_{T,t_1,t_2} \;=\; \frac{1}{T}\sum_{\tau\notin\{t_1,t_2\}} F_\tau F_\tau'\,. \qquad (254)
$$

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 18 and the Sherman-Morrison formula.

Now,

$$\begin{aligned}
&\frac{1}{P}\operatorname{tr} E[F_{t_1}F'_{t_1}(zI + B_{T,t_1,t_2})^{-1}F_{t_2}F'_{t_2}(zI + B_{T,t_1,t_2})^{-1}] \\
&= \frac{1}{P}\operatorname{tr} E[\Bigg(((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2))\Psi\Sigma_F\Psi \\
&+ \Psi\Big(\operatorname{tr}(\Sigma\Sigma_\varepsilon) + \operatorname{tr}(\Sigma_F\Psi)\operatorname{tr}(\Sigma^2)\Big)\Bigg)(zI + B_{T,t_1,t_2})^{-1}\Bigg(((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2))\Psi\Sigma_F\Psi \\
&+ \Psi\Big(\operatorname{tr}(\Sigma\Sigma_\varepsilon) + \operatorname{tr}(\Sigma_F\Psi)\operatorname{tr}(\Sigma^2)\Big)\Bigg)(zI + B_{T,t_1,t_2})^{-1}]
\end{aligned} \tag{255}$$

which coincides with the expression in (246). By the derivations in formulas (247) and (248), we get

$$\begin{aligned}
&\frac{1}{P}\operatorname{tr} E[F_{t_1}F'_{t_1}(zI + B_{T,t_1,t_2})^{-1}F_{t_2}F'_{t_2}(zI + B_{T,t_1,t_2})^{-1}] \\
&\sim \sigma_*^2\frac{1}{P}\operatorname{tr} E[\Psi(zI + B_T)^{-1}\Psi(zI + B_T)^{-1}]\,,
\end{aligned} \tag{256}$$

and hence

$$\begin{aligned}
1 \;&=\; -z^2m'(-z;c) + 2zm(-z;c) + c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2 \\
&+\; \sigma_*^2\frac{1}{P}\operatorname{tr} E[\Psi(zI + B_T)^{-1}\Psi(zI + B_T)^{-1}]/(1+\xi(z;c))^4
\end{aligned} \tag{257}$$

Finally,

$$\frac{\xi(z;c)}{1+\xi(z;c)} \;=\; c(1 - zm(-z;c)) \tag{258}$$

$$
\begin{aligned}
&(1+z^2m'(-z;c)-2zm(-z;c)-c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2)(1+\xi(z;c))^4\\
&= (\frac{d}{dz}(z(1-zm(-z;c)))-c^{-1}\left(\frac{\xi(z;c)}{1+\xi(z;c)}\right)^2)(1+\xi(z;c))^4\\
&= c^{-1}\Big(\frac{d}{dz}(\frac{z\xi(z;c)}{1+\xi(z;c)})(1+\xi(z;c))^2-(\xi(z;c))^2\Big)(1+\xi(z;c))^2\\
&= c^{-1}\Big(\frac{d}{dz}(z-\frac{z}{1+\xi(z;c)})(1+\xi(z;c))^2-(\xi(z;c))^2\Big)(1+\xi(z;c))^2\\
&= c^{-1}\Big(\Big(1-\frac{1}{1+\xi(z;c)}+\frac{z\xi'(z;c)}{(1+\xi(z;c))^2}\Big)(1+\xi(z;c))^2-(\xi(z;c))^2\Big)(1+\xi(z;c))^2\\
&= c^{-1}(\xi(z;c)\ +\ z\xi'(z;c))(1+\xi(z;c))^2
\end{aligned}
\tag{259}
$$

The proof of Lemma 23 is complete. □

We conclude that the first term from (243) characterized in (246) satisfies

$$
\begin{aligned}
Term1\ &=\ \frac{1}{T}E[F'_{t_1}(zI+B_T)^{-1}\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_T)^{-1}F_{t_1}]\\
&\sim\ (1+\xi(z;c))^{-2}c\Gamma_3(z)
\end{aligned}
\tag{260}
$$

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

K.2 $Term2$ in (245)

We now proceed with the second term (245). By the Sherman-Morrison formula and Lemma 18,

$$
\begin{aligned}
& E[F_{t_1}'(zI+B_T)^{-1}\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_T)^{-1}F_{t_2}] \\
& \sim\ E[F_{t_1}'(zI+B_{T,t_1})^{-1}\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_{T,t_2})^{-1}F_{t_2}]/(1+\xi(z;c))^2 \\
& \sim\ E[F_{t_1}'\Big((zI+B_{T,t_1,t_2})^{-1}-\frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_2}F_{t_2}'(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F_{t_2}'(zI+B_{T,t_1,t_2})^{-1}F_{t_2}}\Big) \\
& \Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)\Big((zI+B_{T,t_1,t_2})^{-1} \\
& -\frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}\Big)F_{t_2}]/(1+\xi(z;c))^2 \\
& =\ Term1\ +\ Term2\ +\ Term3
\end{aligned}
\tag{261}
$$

where

$$
\begin{aligned}
Term1\ &=\ E[F_{t_1}'(zI+B_{T,t_1,t_2})^{-1} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)(zI+B_{T,t_1,t_2})^{-1}F_{t_2}]/(1+\xi(z;c))^2 \\
Term2\ &=\ -2E[F_{t_1}'(zI+B_{T,t_1,t_2})^{-1} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big) \\
&\times\frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}F_{t_2}]/(1+\xi(z;c))^2 \\
Term3\ &=\ E[F_{t_1}'\frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_2}F_{t_2}'(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F_{t_2}'(zI+B_{T,t_1,t_2})^{-1}F_{t_2}} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)\frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}F_{t_2}]/(1+\xi(z;c))^2
\end{aligned}
\tag{262}
$$

We now analyze each term separately.

K.3 *Term*1 in (262)

We will need the following lemma.

Lemma 24 *We have*

$$F(A) \;=\; \lambda' E[(zI+B_T)^{-1}A(zI+B_T)^{-1}]\lambda \;\to\; 0 \tag{263}$$

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,

$$\begin{aligned}
&\lambda' E[A(zI+B_T)^{-1}]\lambda \;=\; \lambda' E[(zI+B_T)^{-1}(zI+B_T)A(zI+B_T)^{-1}]\lambda \\
&\underbrace{=}_{symmetry} \; z\lambda' E[(zI+B_T)^{-1}A(zI+B_T)^{-1}]\lambda \;+\; \frac{1}{T}\lambda' E[(zI+B_T)^{-1}F_tF_t'A(zI+B_T)^{-1}\lambda] \\
&= \; z\lambda' E[(zI+B_T)^{-1}A(zI+B_T)^{-1}]\lambda \\
&+ \; E[\Big((zI+B_{T,t})^{-1} - \frac{\frac{1}{T}(zI+B_{T,t})^{-1}F_tF_t'(zI+B_{T,t})^{-1}}{1+\frac{1}{T}F_t'(zI+B_{T,t})^{-1}F_t}\Big)F_tF_t'A(zI+B_T)^{-1}\lambda] \\
&\approx \; z\lambda' E[(zI+B_T)^{-1}A(zI+B_T)^{-1}]\lambda \;+\; (1+\xi(z;c))^{-1}\lambda' E[(zI+B_{T,t})^{-1}F_tF_t'A \\
&\times \Big((zI+B_{T,t})^{-1} - \frac{\frac{1}{T}(zI+B_{T,t})^{-1}F_tF_t'(zI+B_{T,t})^{-1}}{1+\xi(z;c)}\Big)\lambda] \\
&= \; z\lambda' E[(zI+B_T)^{-1}A(zI+B_T)^{-1}]\lambda \\
&+ \; (1+\xi(z;c))^{-1}\lambda' E[(zI+B_{T,t})^{-1}\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)A(zI+B_{T,t})^{-1}]\lambda \\
&- (1+\xi(z;c))^{-2}\lambda' E[(zI+B_{T,t})^{-1}F_tF_t'A\frac{1}{T}(zI+B_{T,t})^{-1}F_tF_t'(zI+B_{T,t})^{-1}]\lambda \\
&\approx \; z\lambda' E[(zI+B_T)^{-1}A(zI+B_T)^{-1}]\lambda \\
&+ \; (1+\xi(z;c))^{-1}\lambda' E[(zI+B_{T,t})^{-1}\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)A(zI+B_{T,t})^{-1}]\lambda \\
&- Q(z)(1+\xi(z;c))^{-2}\lambda' E[(zI+B_{T,t})^{-1}F_tF_t'(zI+B_{T,t})^{-1}]\lambda
\end{aligned} \tag{264}$$

where

$$Q(z) \;=\; F_t' A \frac{1}{T}(zI + B_{T,t})^{-1} F_t \;\to\; T^{-1} \operatorname{tr} E[\Psi A (zI + B_{T,t})^{-1}] \;\to 0 \tag{265}$$

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

$$\begin{aligned} &\lambda' E[(zI + B_{T,t})^{-1} F_t F_t' (zI + B_{T,t})^{-1}] \lambda \\ &= \lambda' E[(zI + B_{T,t})^{-1} \Big((\operatorname{tr} \Sigma)^2 \Psi \Sigma_F \Psi + \Psi \operatorname{tr}(\Sigma \Sigma_\varepsilon) \Big) (zI + B_{T,t})^{-1}] \lambda \;=\; O(1)\,. \end{aligned} \tag{266}$$

Thus, we get

$$o(1) \;\approx\; zF(A) \;+\; (1 + \xi(z;c))^{-1} F((\Psi \Sigma_F \Psi + \Psi) A) \tag{267}$$

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, $E[F_t] = \operatorname{tr}(\Sigma \Sigma_\varepsilon) \lambda_F$ and therefore

$$\begin{aligned} &(1 + \xi(z;c))^2 Term1 \;=\; E[F_{t_1}' (zI + B_{T,t_1,t_2})^{-1} \\ &\Big((\operatorname{tr} \Sigma)^2 \Psi \Sigma_F \Psi + \Psi \operatorname{tr}(\Sigma \Sigma_\varepsilon) \Big) (zI + B_{T,t_1,t_2})^{-1} F_{t_2}] \\ &\sim\; \frac{1}{N^3} (\operatorname{tr}(\Sigma))^2 \lambda_F' \, E[(zI + B_{T,t_1,t_2})^{-1} \\ &\Big((\operatorname{tr} \Sigma)^2 \Psi (\Sigma_{F,t} + \lambda \lambda') \Psi + \Psi \operatorname{tr}(\Sigma \Sigma_\varepsilon) \Big) (zI + B_{T,t_1,t_2})^{-1}] \lambda_F \\ &= \; \frac{1}{N^4} (\operatorname{tr}(\Sigma))^2 \lambda_F' \, E[(zI + B_{T,t_1,t_2})^{-1} (\operatorname{tr} \Sigma)^2 \Psi \Sigma_{F,t} \Psi (zI + B_{T,t_1,t_2})^{-1}] \lambda_F \\ &+ \frac{1}{N^4} (\operatorname{tr}(\Sigma))^2 \lambda_F' \, E[(zI + B_{T,t_1,t_2})^{-1} (\operatorname{tr} \Sigma)^2 \Psi \lambda \lambda_F' (zI + B_{T,t_1,t_2})^{-1}] \lambda_F \\ &+ \frac{1}{N^3} (\operatorname{tr}(\Sigma))^2 \lambda_F' \, E[(zI + B_{T,t_1,t_2})^{-1} (\operatorname{tr} \Sigma \Sigma_\varepsilon) \Psi (zI + B_{T,t_1,t_2})^{-1}] \lambda_F \\ &\sim\; \Gamma_{1,1}(z)^2 \;+\; \Gamma_{4,T}(z)\,, \end{aligned} \tag{268}$$

where Γ_4 is defined in the following lemma.

Lemma 25 *We have*

$$\begin{aligned} \sigma_* \lambda_F' E[(zI + B_{T,t_1,t_2})^{-1} \Psi (zI + B_{T,t_1,t_2})^{-1}] \lambda_F \;&=\; \Gamma_{4,T}(z) \\ \to\; \Gamma_4(z) \;&=\; \frac{\Gamma_{1,1}(z) + z\Gamma_{1,1}'(z) - (\Gamma_{1,1}(z))^2 (1 + \xi(z;c))^{-2}}{(1 + \xi(z;c))^{-2}} \end{aligned} \tag{269}$$

Proof. We have by the symmetry across t and the Sherman-Morrison formula and Lemma 18 that

$$\begin{aligned} \Gamma_{1,1}(z) \;&\sim\; \lambda' E[\Psi (zI + B_T)^{-1} \Psi] \lambda \;=\; \lambda' E[\Psi (zI + B_T)^{-1} (zI + B_T)(zI + B_T)^{-1} \Psi] \lambda \\ &=\; z\, \lambda' E[\Psi (zI + B_T)^{-1} (zI + B_T)^{-1} \Psi] \lambda \;+\; \lambda' E[\Psi (zI + B_T)^{-1} B_T (zI + B_T)^{-1} \Psi] \lambda \\ &=\; -z\, \Gamma_{1,1,T}'(z) \;+\; \lambda' E[\Psi (zI + B_T)^{-1} \frac{1}{T} \sum_t F_t F_t' (zI + B_T)^{-1} \Psi] \lambda \\ &=\; -z\, \Gamma_{1,1,T}'(z) \;+\; \lambda' E[\Psi (zI + B_T)^{-1} F_t F_t' (zI + B_T)^{-1} \Psi] \lambda \\ &\sim\; -z\, \Gamma_{1,1,T}'(z) \;+\; \lambda' E[\Psi (zI + B_{T,t})^{-1} F_t F_t' (zI + B_{T,t})^{-1} \Psi] \lambda (1 + \xi(z;c))^{-2} \\ &=\; -z\, \Gamma_{1,1,T}'(z) \\ &+\; \lambda' E[\Psi (zI + B_{T,t})^{-1} \Big(((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2)) \Psi \Sigma_F \Psi \\ &+\; \Psi \Big(\operatorname{tr}(\Sigma \Sigma_\varepsilon) + \operatorname{tr}(\Sigma_F \Psi) \operatorname{tr}(\Sigma^2) \Big) \Big) (zI + B_{T,t})^{-1} \Psi] \lambda (1 + \xi(z;c))^{-2} \\ &\sim\; -z\, \Gamma_{1,1,T}'(z) \;+\; (\Gamma_{1,1}(z))^2 (1 + \xi(z;c))^{-2} \\ &+\; \Gamma_{4,T}(z) (1 + \xi(z;c))^{-2} \end{aligned} \tag{270}$$

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 *Term*2 in (262)

The next term in (262) is (note the factor of 2 as it appears two times):

$$
\begin{aligned}
Term2 \;&=\; -2E[F_{t_1}'(zI+B_{T,t_1,t_2})^{-1} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big) \\
&\times \frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}F_{t_2}]/(1+\xi(z;c))^2 \\
&=\; -2E[F_{t_1}'(zI+B_{T,t_1,t_2})^{-1} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big) \\
&\times \frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}\lambda_F]\operatorname{tr}(\Sigma)/(1+\xi(z;c))^2 \\
&\sim\; -2E[F_{t_1}'(zI+B_{T,t_1,t_2})^{-1} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big) \\
&\times \frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}\lambda_F]/(1+\xi(z;c))^2 \\
&= -2(1+\xi(z;c))^{-2}E[X_TY_T],
\end{aligned}
\tag{271}
$$

where we have used that

$$
E[F_{t_2}] \;=\; \lambda_F\,, \tag{272}
$$

and where

$$
\begin{aligned}
Y_T \;&=\; F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}\lambda \\
X_T \;&=\; F_{t_1}'(zI+B_{T,t_1,t_2})^{-1} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big) \\
&\times \frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}{1+\frac{1}{T}F_{t_1}'(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}
\end{aligned}
\tag{273}
$$

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

$$E[Y_T] = E[F'_{t_1}(zI + B_{T,t_1,t_2})^{-1}\lambda] \ \to \ \Gamma_{1,1}(z)\,. \tag{274}$$

Proof. Recall that

$$\lambda'\Psi^k(zI + B_T)^{-1}\Psi^\ell\lambda \ \to \ \Gamma_{k,l}(z) \tag{275}$$

by Lemma 21.

We have

$$
\begin{aligned}
& E[\big(F_{t_1}'(zI + B_{T,t_1,t_2})^{-1}\lambda\big)^2] \\
&= E[F_{t_1}'(zI + B_{T,t_1,t_2})^{-1}\lambda\lambda'(zI + B_{T,t_1,t_2})^{-1}F_{t_1}] \\
&= \operatorname{tr} E[(zI + B_{T,t_1,t_2})^{-1}\lambda\lambda'(zI + B_{T,t_1,t_2})^{-1}F_{t_1}F_{t_1}'] \\
&\sim \operatorname{tr} E[(zI + B_{T,t_1,t_2})^{-1}\lambda\lambda'(zI + B_{T,t_1,t_2})^{-1} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)] \;\leq\; Kz^{-2}
\end{aligned}
\tag{276}
$$

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

$$
\begin{aligned}
X_T \;&=\; F_{t_1}'(zI + B_{T,t_1,t_2})^{-1} \\
&\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big) \\
&\times \frac{\frac{1}{T}(zI + B_{T,t_1,t_2})^{-1}F_{t_1}}{1 + \frac{1}{T}F_{t_1}'(zI + B_{T,t_1,t_2})^{-1}F_{t_1}}
\end{aligned}
\tag{277}
$$

Now, we know from the proof of Lemma 11 that

$$
\frac{1}{T}F_t'AF_t - \frac{1}{T}\operatorname{tr}(AE[F_tF_t']) \;\to\; 0
$$

in L_2 and

$$
\begin{aligned}
& F'_{t_1}(zI + B_{T,t_1,t_2})^{-1} \\
& \Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi + \Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)\frac{1}{T}(zI + B_{T,t_1,t_2})^{-1}F_{t_1} \\
& \sim \ \frac{1}{T}\operatorname{tr} E[(zI + B_{T,t_1,t_2})^{-1}\Big(\Psi(\Sigma_{F,t} + \lambda\lambda')\Psi + \sigma_*\Psi\Big) \\
& \times (zI + B_{T,t_1,t_2})^{-1}\Big(\Psi(\Sigma_{F,t} + \lambda\lambda')\Psi + \sigma_*\Psi\Big)] \\
& \overbrace{\qquad\qquad}^{\sim} \\
& \text{(214) and Lemma 24} \\
& \frac{1}{T}\operatorname{tr} E[(zI + B_{T,t_1,t_2})^{-1}\Big(\Psi\lambda\lambda'_F + \sigma_*\Psi\Big) \\
& \times (zI + B_{T,t_1,t_2})^{-1}\Big(\Psi\lambda\lambda'\Psi + \sigma_*\Psi\Big)] \\
& \sim \ \frac{1}{T}\operatorname{tr} E[(zI + B_{T,t_1,t_2})^{-1}\Psi\lambda\lambda'\Psi(zI + B_{T,t_1,t_2})^{-1}\Psi\lambda\lambda'_F] \\
& + \ 2\frac{1}{T}\operatorname{tr} E[(zI + B_{T,t_1,t_2})^{-1}\Psi\lambda\lambda'\Psi(zI + B_{T,t_1,t_2})^{-1}\Psi\sigma_*] \\
& + \ \sigma_*^2\frac{1}{T}\operatorname{tr} E[(zI + B_{T,t_1,t_2})^{-1}\Psi(zI + B_{T,t_1,t_2})^{-1}\Psi] \\
& \sim \ c\Gamma_3(z)
\end{aligned}
\tag{278}
$$

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 (271) imply that

$$Term2 \;\sim -2\frac{c\Gamma_3(z)\Gamma_{1,1}(z)}{(1+\xi(z;c))^3}\,. \tag{279}$$

K.5 $Term3$ in (262)

Finally, we now deal with $Term3$ in (262).

Lemma 28 *$Term3$ in (262) converges to zero.*

Proof of Lemma 28. We have

$$\begin{aligned} Term3 \;&=\; E[F'_{t_1}\frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_2}F'_{t_2}(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F'_{t_2}(zI+B_{T,t_1,t_2})^{-1}F_{t_2}} \\ &\Big((\operatorname{tr}\Sigma)^2\Psi\Sigma_F\Psi+\Psi\operatorname{tr}(\Sigma\Sigma_\varepsilon)\Big)\frac{\frac{1}{T}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}F'_{t_1}(zI+B_{T,t_1,t_2})^{-1}}{1+\frac{1}{T}F'_{t_1}(zI+B_{T,t_1,t_2})^{-1}F_{t_1}}F_{t_2}]/(1+\xi(z;c))^2 \\ &=\; E[X_TY_T]\,/(1+\xi(z;c))^2\,, \end{aligned} \tag{280}$$

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

$$\begin{aligned} & E[(F_{t_1}' A F_{t_2})^2] \;=\; N^{-2} E[F_{t_1}' A F_{t_2} F_{t_2}' A F_{t_1}] \\ & =\; N^{-2} \, \mathrm{tr} \, E[A F_{t_2} F_{t_2}' A F_{t_1} F_{t_1}'] \\ & \sim\; \mathrm{tr} \, E[A \left(\Psi \Sigma_F \Psi \;+\; \sigma_* \Psi \right) \\ & \times A \left(\Psi \Sigma_F \Psi \;+\; \sigma_* \Psi \right))] \end{aligned} \tag{281}$$

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

$$R_t \;=\; S_{t-1}\beta_t \;+\; \varepsilon_t, \tag{282}$$

and

$$F_t \;=\; S'_{t-1}R_t \;=\; S'_{t-1}S_{t-1}\beta_t + S'_{t-1}\varepsilon_t \;=\; Z_t\beta \;+\; S'_{t-1}\varepsilon_t \tag{283}$$

and therefore

$$F_tF'_t \;=\; Z_t\beta\beta' Z_t + S'_{t-1}\varepsilon_t\beta' Z_t + Z_t\beta\varepsilon'_t S_{t-1} + S'_{t-1}\varepsilon_t\varepsilon'_t S_{t-1}\,. \tag{284}$$

and formula (154) applied to $t = t_1$ implies

$$\begin{aligned}
E[(F'_{t_1}AF_{t_2})^4] &= E[F'_{t_1}AF_{t_2}F'_{t_2}AF_{t_1}F'_{t_1}AF_{t_2}F'_{t_2}AF_{t_1}] \\
&= \operatorname{tr} E[F_{t_1}F'_{t_1}AF_{t_2}F'_{t_2}AF_{t_1}F'_{t_1}AF_{t_2}F'_{t_2}A] \\
&= \operatorname{tr} E[Z_{t_1}\beta\beta'Z_{t_1}AF_{t_2}F'_{t_2}AZ_{t_1}\beta\beta'Z_{t_1}AF_{t_2}F'_{t_2}A] \\
&+ \operatorname{tr} E[Z_{t_1}\beta\beta'Z_{t_1}AF_{t_2}F'_{t_2}AZ_{t_1}AF_{t_2}F'_{t_2}A] \\
&+ 2\operatorname{tr} E[(\beta'Z_{t_1}AF_{t_2}F'_{t_2}AZ_{t_1}\beta)Z_{t_1}AF_{t_2}F'_{t_2}A] \\
&+ ((\kappa_\varepsilon - 1)\operatorname{tr} E[Z_{t_1}AF_{t_2}F'_{t_2}AZ_{t_1}AF_{t_2}F'_{t_2}A] \\
&+ E[\operatorname{tr}(Z_{t_1}AF_{t_2}F'_{t_2}A)^2]
\end{aligned} \tag{285}$$

We then again apply (154) to $t = t_2$. It is then straightforward to show that the leading contribution will be

$$\begin{aligned}
E[\operatorname{tr}(Z_{t_1}AZ_{t_2}A)^2] &= E[\left(\sum X_{i_1,k_1,t_1}\lambda_{i_1}(\Sigma)X_{i_1,k_2,t_1}\lambda_{k_2}(\tilde{A})X_{i_2,k_2,t_2}\lambda_{i_2}(\Sigma)X_{i_2,k_1,t_2}\lambda_{k_1}(\tilde{A})\right)^2] \\
&= E[\sum X_{i_1,k_1,t_1}\lambda_{i_1}(\Sigma)X_{i_1,k_2,t_1}\lambda_{k_2}(\tilde{A})X_{i_2,k_2,t_2}\lambda_{i_2}(\Sigma)X_{i_2,k_1,t_2}\lambda_{k_1}(\tilde{A}) \\
&\times X_{\tilde{i}_1,\tilde{k}_1,t_1}\lambda_{\tilde{i}_1}(\Sigma)X_{\tilde{i}_1,\tilde{k}_2,t_1}\lambda_{\tilde{k}_2}(\tilde{A})X_{\tilde{i}_2,\tilde{k}_2,t_2}\lambda_{i_2}(\Sigma)X_{\tilde{i}_2,\tilde{k}_1,t_2}\lambda_{\tilde{k}_1}(\tilde{A})]
\end{aligned} \tag{286}$$

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 $\operatorname{tr}(\Sigma)^4(\operatorname{tr}(\tilde{A}^2))^2$. 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, (280) converges to zero.

The proof of Lemma 28 is complete. □

Summarizing, we get from (271) and (268), (279), that

$$Term2 \;=\; (1+\xi(z;c))^{-2}(\Gamma_{1,1}(z)^2+\Gamma_4(z)) \;-2\frac{c\Gamma_3(z)\Gamma_{1,1}(z)}{(1+\xi(z;c))^3} \tag{287}$$

and (243) implies

$$\begin{aligned} &E[(R^F_{t+1}(z))^2] \underbrace{\sim}_{(243)} Term1 \;+\; Term2 \\ &\underbrace{\sim}_{} (260)\; (1+\xi(z;c))^{-2}c\Gamma_3(z) \;+\; Term2 \\ &\underbrace{\sim}_{(287)} (1+\xi(z;c))^{-2}c\Gamma_3(z) \;+\; (1+\xi(z;c))^{-2}(\Gamma_{1,1}(z)^2+\Gamma_4(z)) \;-2\frac{c\Gamma_3(z)\Gamma_{1,1}(z)}{(1+\xi(z;c))^3} \end{aligned} \tag{288}$$

and the final expression follows from Lemma 25:

$$\Gamma_{1,1}(z)^2+\Gamma_4(z) \;=\; \Gamma_{1,1}(z)^2+\frac{\Gamma_{1,1}(z)+z\Gamma'_{1,1}(z)-(\Gamma_{1,1}(z))^2(1+\xi(z;c))^{-2}}{(1+\xi(z;c))^{-2}} \tag{289}$$

L Pricing Errors

Proof of Proposition 6. We have

$$\begin{aligned} &PricingError(z;cq;q) \;=\; E[F'(1-\lambda(z;q)'F(q))]\,E[FF']^{-1}E[(1-\lambda(z;q)'F)F] \\ &=\; (E[F]-E[FF(q)']\lambda(z;q))'E[FF']^{-1}(E[F]-E[FF(q)']\lambda(z;q)) \\ &=\; E[F]'E[FF']^{-1}E[F]-2\underbrace{E[R^F(z;q)F']E[FF']^{-1}E[F]}_{directional} \\ &+\underbrace{E[R^F(z;q)F']E[FF']^{-1}E[R^F(z;q)F]}_{risk} \\ &=\; E[F]'E[FF']^{-1}E[F]-2E[R^F(z;q)] \;+\; E[(R^F(z;q))^2] \end{aligned} \tag{290}$$

We have

$$E\left[\hat{\lambda}(z;q)'\left(\frac{1}{\hat{T}}\sum_{\tau}(F_\tau(q))F_\tau'\right)((0+)I+\hat{B}_{\hat{T}})^{-1}\left(\frac{1}{\hat{T}}\sum_{\tau}F_\tau\right)\right] \tag{291}$$

Now, all matrices here have a block structure:

$$\left(\frac{1}{\hat{T}}\sum_{\tau}(F_\tau(q))F_\tau'\right) \;=\; [\hat{B}_{\hat{T}}(q)+(0+)I\,,\hat{\Psi}_{1,2}] \tag{292}$$

where $\hat{\Psi}_{1,2}\in\mathbb{R}^{P_1\times(P-P_1)}$ and, assuming for simplicity that

$$\left(\frac{1}{\hat{T}}\sum_{\tau}(F_\tau(q))F_\tau'\right)((0+)I+\hat{B}_{\hat{T}})^{-1} \;=\; [I_{P_1\times P_1},0_{P_1\times(P-P_1)}] \tag{293}$$

by the definition of the inverse matrix. Namely,

$$(A,B)\begin{pmatrix}A & B\\ C & D\end{pmatrix}^{-1} \;=\; (I,0) \tag{294}$$

Thus,

$$E[R^F(z;q)F']E[FF']^{-1} \;=\; \hat{\lambda}(z;q)'(I,0) \tag{295}$$

and hence

$$\begin{aligned}&E[R^F(z;q)F']E[FF']^{-1}E[R^F(z;q)F]\\ &=\; E[R^F(z;q)F']E[FF']^{-1}E[FF']E[FF']^{-1}E[R^F(z;q)F]\\ &=\; \hat{\lambda}(z;q)'E[F(q)F(q)']\hat{\lambda}(z;q)\,.\end{aligned} \tag{296}$$

Finally, the last identity follows from

$$\mathcal{D} \;=\; 1 \;-\; 2E[\hat{R}^M] \;+\; E[(\hat{R}^M)^2] \;=\; 1 - 2\mathcal{E}(Z^*) \;+\; \mathcal{V}(Z^*) + G(z;c)\mathcal{R}(Z^*) \;=\; \mathcal{R}(Z^*) + G(z;c)\mathcal{R}(Z^*) \tag{297}$$

□

M Robustness Check: Turnover

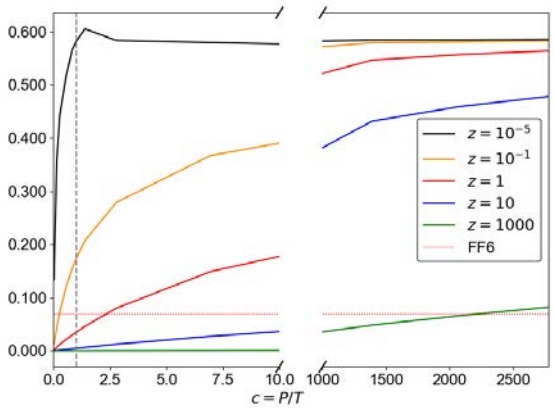
To establish that our fundamental findings are not contingent on signals with exceptionally high turnover, we replicate our main experiments, excluding the 20 signals exhibiting the highest turnover. We define turnover for each characteristic i as the temporal and cross-sectional mean of the absolute variation in the rank-standardized characteristic from one month to the next. Formally, this can be expressed as:

$$Turnover_i = \frac{1}{\tau}\sum_{t=1}^{\tau}\frac{1}{N_t}\sum_{k=1}^{N_t}|X_{i,k,t} - X_{i,k,t}| \tag{298}$$

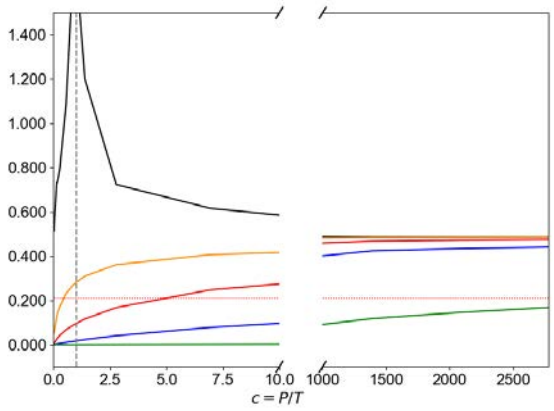
Table 1 displays the turnover values for the top 20 characteristics with the highest turnover rates. We exclude these characteristics from our original set of 130, as detailed in Section 5, and then rerun our principal experiment. The estimated VoC curves are depicted in Figure 8. Remarkably, the outcomes closely mirror those illustrated in Figure 2, thereby confirming that our primary empirical findings are not influenced by characteristics with high turnover rates.

X_i	$Turnover_i$	X_i	$Turnover_i$
ret_1_0	0.351	rmax1_21d	0.230
seas_1_1an	0.337	rmax5_21d	0.211
rskew_21d	0.322	bidaskhl_21d	0.201
coskew_21d	0.322	ivol_ff3_21d	0.196
iskew_ff3_21d	0.321	ivol_capm_21d	0.192
iskew_capm_21d	0.320	ivol_hxz4_21d	0.192
iskew_hxz4_21d	0.308	rvol_21d	0.180
rmax5_rvol_21d	0.301	resff3_6_1	0.151
beta_dimson_21d	0.283	prc_highprc_252d	0.151
ret_3_1	0.241	ret_6_1	0.150

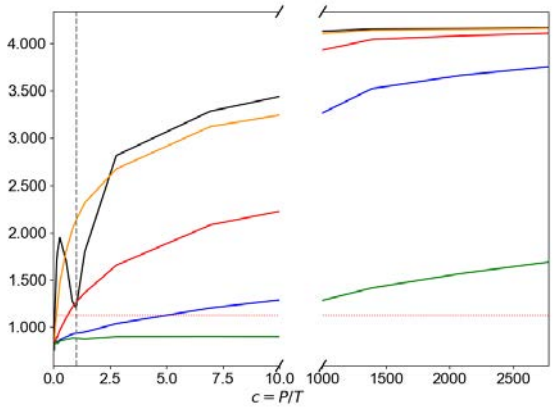
Table 1: This table enumerates the 20 characteristics with the highest turnover in our sample. The first column designates the names of the characteristics, as articulated in Jensen et al. (Forthcoming). The second column delineates the associated turnover values, as formalized in Equation (298).



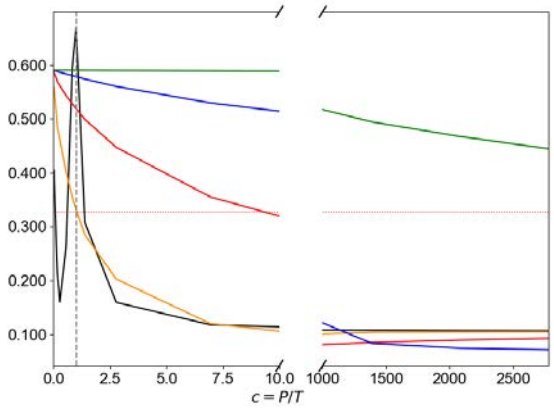
(a) Expected Return



(b) Second Moment



(c) Sharpe Ratio



(d) Pricing Error

Figure 8: Out-of-sample Performance of Complex SDF Model built on low turnover characteristics.