Complexity in the Cross-Section

Antoine Didisheim, Shikun Ke, Bryan Kelly, and Semyon Malamud*

September 21, 2022

Abstract

We provide evidence that the dependence of expected stock returns on characteristics is highly non-linear and complex: Out-of-sample performance of predictive models increases in the number of model parameters even when this number is comparable with the panel size. Contrary to conventional wisdom, we find that small and micro-cap stocks have the same complexity and predictability as mega stocks. The differences in the performance of cross-sectional strategies between size groups can be attributed to different signal correlation structures. We prove that complexity creates a significant wedge between the feasible out-of-sample Sharpe ratio and the hypothetical one, attainable by an econometrician with access to an infinitely large panel. Our estimates suggest that the two Sharpe ratios differ by a factor of at least five. The calibrated wedge is significant even for models with shallow complexity levels.

Keywords: Portfolio choice, machine learning, random matrix theory, benign overfit, overparameterization, asset pricing, stock cross-section
JEL: C3, C58, C61, G11, G12, G14

*Antoine Didisheim is at the University of Lausanne. 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. AQR Capital Management is a global investment management firm, which 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.

1 Introduction

A growing literature uses machine learning methods to study the cross-section of stock returns. (Gu et al., 2020) outline the rationale for machine learning in the cross-section, which may be summarized as follows: A fundamental goal of asset pricing research is to understand the behavior of expected returns and measuring expected returns boils down to a prediction problem. Machine learning methods are well adapted to deal with i) high-dimensional predictor sets and ii) unknown and potentially non-linear functional forms. Thus, machine learning confronts the large "zoo" of candidate anomaly signals as well as the lack of theoretical guidance regarding the functional form for conditional expected returns.

The main empirical conclusion in the literature is that machine learning methods achieve large out-of-sample gains in forecasting the cross section of stock returns relative to traditional, linear, low-dimensional models.[1] But how do these models manage to avoid deleterious overfit and predict so well out-of-sample? Little is known about the theoretical properties of such heavily parameterized models or about the portfolios derived from them. In this paper, we seek to fill this gap. We provide a theoretical characterization of both the cross-section of expected returns estimated from machine learning models and of optimal portfolios built from these estimates.

In traditional econometrics of panel data, the number of model parameters P is small relative to the number of training observations, $T \cdot N$, with T being the number of time periods and N the number of assets in the panel.[2] Therefore, estimators are aptly characterized by the law of large numbers and central limit theorem. This form of asymptotic theory is ill-suited for analyzing machine learning models, which attempt to provide accurate approximations of empirical phenomena through extensive parameterization. Instead, characterizations

[1] See, (Lewellen, 2014) for an overview of linear predictive models, and (Gu et al., 2020), (Kelly et al., 2020), Chinco et al. (2019), (Han et al., 2019), (Liu et al., 2020), (Bryzgalova et al., 2020), (Chen et al., 2019), Chen et al. (2021), Avramov et al. (2021), Guijarro-Ordonez et al. (2021), Ma (2021), Leippold et al. (2022), Gagliardini and Ma (2019) for non-linear, machine-learning-based models.

[2] See, e.g., Gagliardini et al. (2016) where both $N, T \to \infty$, and Chen et al. (2021) where $N \to \infty$.

of machine learning models must accommodate a parameter dimension that is large relative to sample size—the apparatus of random matrix theory is ideally suited for this task.

In particular, our theoretical development characterizes machine learning estimates and related portfolio strategies using asymptotic analysis of a T by N panel when parameterization P grows with the number of observations at a fixed rate: $\lim_{P,TN\to\infty} P/(TN) = c > 0$. We refer to c as the "model complexity" and investigate (both theoretically and empirically) the behavior of machine learning models in practical settings, by gradually increasing c.

A natural reaction to this framework might be: "In practice, the number of return predictors is not very big relative to the size of the panel, so why do I care about model behavior in the large P limit?" After all, the panel is large—it has thousands of stocks times hundreds of months—while the factor zoo amounts to only a few hundred predictors. Our answer to this question is: If the "true" prediction model is non-linear, with an unknown functional form, we need a large number P of parameters to account for the relevant non-linearities.

Numerous machine learning models have been developed to efficiently approximate unknown non-linearities, with deep neural networks being the most prominent and popular example (see, e.g., (Gu et al., 2020) for an overview). The general conclusion is that, unless the researcher knows the correct functional form a priori—which requires heroic assumptions for complicated systems like financial markets—complex (large P) models are needed to accurately approximate the unknown truth.[3] These models achieve their superior performance by engineering thousands of non-linear transformations (features) of the original signals. In the lingua franca of asset pricing, forty years of research has produced a "factor zoo" of a few hundred of stock predictors. Machine learning methods expand this small zoo into an arbitrarily large Noah's ark of signals by building extremely complex models. Our

[3]One known example is a single neuron function $f(x) = \frac{\exp(\theta' x)}{1+\exp(\theta' x)}$, which can only be learned with deep neural networks and requires exponentially large training samples to be learned with shallow models. See (Ghorbani et al., 2020).

goal in this paper is to understand (both theoretically and empirically) the role of complexity for return prediction and the fundamental properties of returns and signals that shape the relationship between complexity and model performance.

To maintain tractability in our theoretical derivations, we work in the confines of linear least squares prediction models (while allowing for ridge shrinkage). The tractability of linear least squares sacrifices little in terms of generality, owing to a number of recent papers that establish equivalence between sophisticated models such as deep neural networks and high-dimensional linear models.[4] We use the method of random features regression (RFR) introduced by Rahimi and Recht (2007), which is a tractable neural network method that can be estimated with nothing more than ridge regression with the original signals replaced by thousands of random, non-linear features. Thus, at the end of the day, our estimation is still a penalised linear regression. It simply occurs inside a highly non-linear space. We apply this methodology with a fixed set of 153 raw predictors from (Jensen et al., Forthcoming) and four subgroups of US stocks grouped by market capitalization (size): mega, large, small, and micro.

Following KMZ, we use a short rolling window of only 12 months to estimate our predictive model.[5] Our empirical and theoretical results demonstrate the existence of a "virtue of complexity" (VoC) in predicting the cross-section of stock returns: The out-of-sample Sharpe ratios monotonically increasing in P and flatten out when $P \approx 10,000$, corresponding to a complexity of $c \approx 2.9$ with a Sharpe Ratio of 0.98 for the "mega" group; $c \approx 1.9$ with a Sharpe Ratio of 1.39 for the "large" group; $c \approx 1.5$ with a Sharpe Ratio of 1.86 for the "small" group; and $c \approx 0.7$ with a Sharpe Ratio of 2.45 for the "micro" group.[6] The

[4] See (Jacot et al., 2018; Hastie et al., 2019; Allen-Zhu et al., 2019).

[5] This window choice is justified by several considerations: (i) computational cost: A window of 120 months and a 1000 stocks gives a panel size of 120000, and matrix of 120000×10000 of random features, which would quickly lead to limitations due to operating memory (RAM) constraints. RAM is a major constraints of many random features and kernel models, where a very large set of random features needs to be fully kept in memory to perform regression. (ii) non-stationarity: We find evidence that the predictive power of the model fluctuates a lot over time, and hence using a short rolling window (as in KMZ) allows us to react more quickly to changing market conditions.

[6] This result generalizes the analysis of Kelly, Malamud, and Zhou (2022, henceforth KMZ), who document

performance is robust, consistent over time and size groups, and the strategies have a large and statistically significant α against the standard six factor model (Fama-French five factor model from Fama and French (2015) plus momentum from Jegadeesh and Titman (1993)). In addition, the model also delivers significant alpha relative to two linar, low-complexity models: The standard, equal weighted combination of all 153 factors and linear panel ridge regression on the 153 characteristics estimated over the same rolling window of 12 months.

We show that model performance and the shape of the VoC depend on a novel object that we introduce in this paper and that only arises in models of the cross-section returns. This novel object is the average covariance matrix of signals *across assets*.[7] It is different from the covariance matrix of signals across signal types (e.g., the covariance between value and momentum signals for every stock, averaged across stocks) that plays a key role in Kelly et al. (2022).

We show explicitly how the eigenvalue distributions of the two matrices affect the Sharpe ratio of the complex trading strategy and the predictive R^2 of the panel regression. Specifically, we highlight the key role in our explicit expressions played by the Herfindahl index of the eigenvalue distribution.

Indeed, we show that a cross-asset signal covariance matrix with a high eigenvalue Herfindahl index implies that signals are aligned across stocks. In such a context, gains from diversification across a large cross-section are limited, and the optimal trading strategy with a high Herfindahl index will be based on a small set of stock portfolios responsible for most of the variation in conditional returns. By contrast, a low Herfindahl index allows much

a theoretical and empirical virtue of complexity for the prediction of the US market equity risk premium (see, also, Kelly et al. (2022) for document the virtue of complexity in return prediction for other asset classes). Our Sharpe ratios are comparable with those obtained in Gu et al. (2020), they obtain Sharpe Ratios as high as 1.35 on a large cross-section of stocks (average number of stocks per month exceeding 6,200). Contrary to our analysis, Gu et al. (2020) use an expanding window to estimate their complex models.

[7]More precisely, for each stock we have multiple characteristics (features). For each characteristic type (e.g., value and momentum), we compute their cross-sectional covariance matrix, and then sum these covariance matrices across characteristic types (e.g., covariance of value across stocks plus covariance of momentum across stocks).

larger benefits from diversification. We stress that the cross-asset signal covariance matrix is estimated in the feature space and is, therefore, different across levels of complexity. More importantly, the benefits of diversification discussed above are defined in the non-linear space. This nuance is key to understanding how complex models outperform linear approximations. While the raw (153-dimensional) signals are often linearly related, non-linear combinations of these same factors tend to be much less correlated. Thus, portfolios built on forecasts from complex models can be diversified in the non-linear space, leading to superior performance. Interestingly, this result implies that complex models can produce higher Sharpe ratios without significantly improving the individual stock-per-stock prediction accuracy of a linear model.

The empirical evidence discussed above suggests that the true relationship between stock returns and characteristics is highly complex. The c (the number of parameters of the true model divided by the number of observations) is so high that the true relationship between raw signals and returns can never be estimated. Indeed, in such a situation the law of large numbers breaks down and the theoretical expected performance of the model based on the unrealistic assumption of unlimited data (corresponding to $c = P/(NT) \approx 0$) becomes infeasible. Using random matrix theory, we compute the theoretically feasible Sharpe Ratio, $SR(c)$ that depends on c in an explicit manner. The wedge, $1 - SR(c)/SR(0)$, represents the "Limits to learning" driven by the insufficient amount of data.

This novel concept of "Limits to learning" is distinct from the well-known concept of over-fitting, which refers to the harmful effect of over-parameterization. While over-fitting is a problem of bad modelization or bad optimization, limits to learning capture a more fundamental concept for economists. It is the difference between the accessible level of predictability in the cross-section and the true predictability that can never be attained. As such, it carries natural implications about limits to arbitrage, the discussion of which are beyond the scope of this paper. Empirically, we find that the infeasible Sharpe ratio, $SR(0)$, is

at least five times higher than the feasible one, $SR(c)$, for realistic complexity levels. Perhaps surprisingly, even for models with low complexity (corresponding to $P = 100$ parameters and a $c \approx 100/20000 = 0.005$), the calibrated wedge is still very high ($1 - SR(c)/SR(0) \approx 0.5$). The reason is that the wedge is determined jointly by c and the signal-to-noise ratio. Low complexity models also have low signal-to-noise ratios, and learning the true relationship becomes impossible even in large panels.

This paper relates to the emergent literature that studies the theoretical properties of machine learning models. A number of recent papers show that linear models combined with random matrix theory help characterize the behavior of neural networks trained by gradient descent.[8] Recent research also emphasizes the phenomenon of benign overfit and "double descent," in which expected forecast error drops in the high complexity regime.[9] Since standard machine learning datasets do not have a natural panel structure, little is known about the behavior of panel regressions in the machine learning regime. To the best of our knowledge, our paper is first one to derive explicit formulas for out of sample Rsquared and Sharpe ratios for panel data. As in Hastie et al. (2019), our key interest is in a mis-specified model, where the "true" model has a large number P of parameters, and we consider an approximating sequence of models of increasing complexity, indexed by $P_1 < P$. While Hastie et al. (2019) focus on a specific form of mis-specification and its ridgeless limit, we derive general expressions for asymptotic expected returns and volatility in terms of signal correlations. The impact of the panel structure (through the eigenvalue distribution of the cross-asset covariances) on the out-of-sample performance is non-trivial, and we characterize it in closed form using novel mathematical techniques.

Our research question is motivated by a growing empirical literature that uses machine learning methods to analyze stock returns. While too large to survey here, a partial reference list includes Feng et al. (2018), Rossi (2018), Rapach and Zhou (2020), Kozak et al. (2020),

[8] See, for example, Jacot et al. (2018); Hastie et al. (2019); Du et al. (2018, 2019); Allen-Zhu et al. (2019).

[9] See, for example, Spigler et al. (2019); Belkin et al. (2018, 2019, 2020); Bartlett et al. (2020), Hastie et al. (2019), Richards et al. (2021).

Freyberger et al. (2020), Gu et al. (2020), Chen et al. (2019), and Bianchi et al. (2021), Filipović and Khalilzadeh (2021), Filipović et al. (2022). Our paper is also related to Martin and Nagel (2021) and Da et al. (2022) who examine market efficiency implications of the high-dimensional prediction problem faced by investors, to Fan et al. (2022b) who touch upon the "double descent" phenomenon in their analysis of structural machine learning models, and to financial econometrics applications of random matrix theory such as Fan et al. (2008), Ledoit and Wolf (2020), and Fan et al. (2022a).

The remainder of the paper is organized as follows. In section 2, we outline our modeling assumptions and environment. Section 3 derives the main theoretical results using random matrix theory. In section 4 we present the empirical frameworks and results. Section 5 concludes.

2 Environment

This section describes our modeling assumptions and outlines the criteria by which we evaluate machine learning portfolios. We refer readers to KMZ for background and a fuller discussion of intuition for our framework. Here, we emphasize the new aspects of model complexity that emerge from introducing a cross-section dimension to the asset space.

2.1 Assumptions on Data Generating Process

Assumption 1 *There are N assets whose excess returns behave according to*

$$\tilde{R}_{t+1} = \tilde{S}_t \beta + \tilde{\varepsilon}_{t+1} \tag{1}$$

with $\tilde{\varepsilon}_{t+1}$ i.i.d., $E[\tilde{\varepsilon}_{i,t+1}] = E[\tilde{\varepsilon}_{i,t+1}\tilde{\varepsilon}_{j,t+1}\tilde{\varepsilon}_{k,t+1}] = 0$ for any i, j, k. Furthermore, $E[\tilde{\varepsilon}_{t+1}\tilde{\varepsilon}'_{t+1}] = \Sigma_{\varepsilon,t} \in \mathbb{R}^{N\times N}$ is a positive-definite matrix, and $E[\|\tilde{\varepsilon}_{t+1}\|^4] < \infty$, and $\tilde{S}_t = (S_{i,k,t})$ is an

$N \times P$-matrix of predictor variables. We define

$$R_{t+1} \;=\; \Sigma_{\varepsilon,t}^{-1/2}\tilde{R}_{t+1} \;=\; S_t\beta \;+\; \varepsilon_{t+1}\,,\; S_t \;=\; \Sigma_{\varepsilon,t}^{-1/2}\tilde{S}_t\,, \tag{2}$$

and note that $E[\varepsilon_{t+1}\varepsilon_{t+1}'] \;=\; I\,.$

Assumption 1 establishes the basic return generating process. Its key feature is that conditional expected returns depend on a potentially high-dimensional information set embodied by the predictors, S. This interpretation of this assumption is *not* that asset returns are subject to many fundamental driving forces. Instead, it espouses the machine learning perspective discussed in the introduction: The DGP's functional form is unknown but may be approximated with richly parameterized models using high-dimensional nonlinear expansion S of some underlying feature set. The assumption does not take a stand on the source of predictability. Predictors S_t may capture effects ranging from rational risk compensation to irrational mispricing.

The covariance structure of S is central to understanding the behavior of machine learning predictions and portfolios. The following assumption imposes basic regularity conditions on this covariance.

Assumption 2 *There exist independent random matrices $X_t \in \mathbb{R}^{N\times P}$ with four finite first moments, and two symmetric, positive definite matrices $\Sigma \in \mathbb{R}^{N\times N}$ and $\Psi \in \mathbb{R}^{P\times P}$, such that*

$$S_t = \Sigma^{1/2}X_t\Psi^{1/2} \tag{3}$$

Furthermore, $E[X_{i,k,t}] = 0$, all (mixed) third-order moments of X_t are zero, and $E[X_t'X_t] = PI_{P\times P}$. Furthermore, the fourth moments $E[X_{i,k,t}^4]$ are uniformly bounded and $X_{i,k,t}$ satisfy

the Lindenberg condition

$$\lim_{P\to\infty} \frac{1}{P} \sum_{i=1}^{N} \sum_{k=1}^{P} E[X_{i,k,t}^2 I_{|X_{i,k,t}|>\varepsilon\sqrt{P}}] = 0 \ for \ all \ \varepsilon > 0.$$

Assumption 1 implies that Ψ and Σ are identifiable only up to a multiplicative constant. Indeed, multiplying Σ by a constant and dividing Ψ by the same constant does not change S_t. Up to this constant, Ψ and Σ can be identified using the identities

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

While Ψ is the direct analog of Ψ in the (Kelly et al., 2021) (the time series covariance matrix of signals), the Σ matrix is a key new object in this paper. It captures the covariances of signals across assets.[10]

The theoretical properties of machine learning portfolios depend heavily on the *distributions of eigenvalues* of both Ψ and Σ. We are interested in limiting behavior in the high model complexity regime, i.e. as $P, T \to \infty$, with $P/(NT) \to c > 0$. Assumption 3 ensures that estimates of Ψ are well-behaved in this limit. We assume that N is finite and fixed, and impose no assumptions on the matrix Σ.

Assumption 3 *We will use $\lambda_k(\Psi)$, $k = 1, \cdots, P$, to denote the eigenvalues of an arbitrary matrix Ψ. In the limit as $P \to \infty$, the spectral distribution G^Ψ of the eigenvalues of Ψ ,*

$$G^\Psi(x) = \frac{1}{P} \sum_{k=1}^{P} \mathbf{1}_{\lambda_k(\Psi)\leq x} \tag{5}$$

converges to a non-random probability distribution H supported on $(0, +\infty)$.[11] *Furthermore,*

[10] In our empirical analysis, signals are not just raw stock characteristics (such as value and momentum). Instead, they are complex non-linear transformations (features) of raw characteristics . See Section 4.2 for details. In a complex model, Σ accounts for covariances of these features across assets.

[11] If 0 is in the support of H, then Ψ is strictly degenerate, meaning that some signals are redundant.

Ψ *is uniformly bounded as* $P \to \infty$. *We will use*

$$\psi_{*,k} = \lim_{P\to\infty} P^{-1} \operatorname{tr}(\Psi^k) \ , \ k \geq 1$$

to denote asymptotic moments of the eigenvalues of Ψ.

Our last DGP assumption governs the behavior of the true predictive coefficient, β, and is a natural panel extension of the β assumption in KMZ.

Assumption 4 *We assume* $\beta = \beta_P$ *is random,* $\beta = (\beta_i)_{i=1}^P \in \mathbb{R}^P$, *independent of* S *and* R, *and satisfies* $E[\beta] = 0$, *and* $E[\beta\beta'] = P^{-1}b_{*,P}I$ *for some constant* $b_{*,P} = E[\|\beta\|^2]$,[12] *and satisfies* $b_{*,P} \to b_*$ *almost surely, for some* $b_* > 0$. *Furthermore,* $E[\beta_i^4] \leq cP^{-2}$ *for some* $c > 0$, *and* β *satisfy the same Lindenberg condition as* X.

Assumption 4 predictive content of signals is rotationally symmetric; i.e., predictability is uniformly distributed across signals in expectation. From a technical standpoint, it is possible to derive explicit expressions for portfolio performance without this assumption, but the expressions become more complex. In this case, the asymptotic behavior depends on the joint distribution of β and Ψ. $E[\beta\beta']$ to P^{-1}, and the finite limiting ℓ_2 norm of β, controls the "true" Sharpe ratio, ensuring that Sharpe ratios remain bounded as the number of predictors grows. In other words, our setting is one in which additional predictors are not distinct drivers of expected returns but represent terms that improve the model's approximation of a nonlinear functional form. While commonly used stock characteristics clearly do not satisfy Assumption 4, we use a "random features" methodology for constructing nonlinear features from "raw" stock characteristics that satisfy rotational symmetry by construction. See Section 4.2 for details. Note that, intuitively, we expect the random features to be relatively less correlated between themselves than the raw characteristics, implying that Ψ

[12] This identity follows because $b_* = \operatorname{tr} E[\beta\beta'] = E[\operatorname{tr}(\beta\beta')] = E[b_*]$.

is roughly proportional to the identity matrix. However, this structure of random features does not necessarily imply that signals are less correlated across assets.[13]

2.2 Modelling Mis-specification

We consider a true DGP with P predictors per stock. Following KMZ, we consider an expanding set of *mis-specified* empirical models to approximate the DGP. Each model is indexed by $P_1 = 1, \cdots, P$ and corresponds to the model used by an economic agent observing only a subset of the signals, $S_t^{(1)} = (S_{i,k,t})_{k=1}^{P_1}$. We use $S_t^{(2)} = (S_{i,k,t})_{k=P_1+1}^{P}$ to denote the remaining unobserved signals. Similarly, $\beta^{(1)}$ is the component of the β vector corresponding to observed signals, while $\beta^{(2)}$ corresponds to unobserved signals. The special case $P_1 = P$ corresponds to a correctly specified empirical model.

The signal covariance matrix corresponding to this partition is

$$\Psi = \begin{pmatrix} \Psi_{1,1} & \Psi_{1,2} \\ \Psi_{1,2}' & \Psi_{2,2} \end{pmatrix}. \tag{6}$$

Naturally, misspecified estimator behavior depends on the correlation structure of observed and unobserved signals. This correlation structure is captured by the off-diagonal blocks of Ψ.

We make the following technical assumption which ensures that estimators in the machine learning regime have well-behaved limits.

Assumption 5 *For any sequence $P_1 \to \infty$ such that $P_1/P = q \in (0, 1]$, the eigenvalue distribution of the matrix $\Psi_{1,1}$ converges to a non-random probability distribution $H(x; q)$.*

[13] As an illustration, two identical assets k_1, k_2 with identical signal vectors $S(k_1) = S(k_2) \in \mathbb{R}^P$ may have $E[S(k_1)S(k_1)'] = E[S(k_2)S(k_2)'] = I_{P\times P}$. However, $Corr(f(S(k_1)), f(S(k_2))) = 1$, no matter how many non-linear transformations f of signals we consider.

We will use

$$\psi_{*,k}(q) = \lim_{P_1 \to \infty} P_1^{-1} \operatorname{tr}(\Psi_{1,1}^k) \;=\; \int x^k dH(x;q) \;,\; k \geq 1$$

to denote asymptotic moments of the eigenvalues of $\Psi_{1,1}$.

The empirical model we analyze is the ridge-regularized least squares estimator for the model in (1). As elaborated in KMZ, ridge regression is a remarkably powerful and tractable framework for analyzing the behavior of machine learning portfolios due to its conjugacy with the key limit results of random matrix theory (RMT).

The panel ridge estimator with shrinkage parameter z is

$$\hat{\beta}(z;q) = \left(zI + \hat{\Psi}_{1,1}\right)^{-1} \frac{1}{NT} \sum_t S_t^{(1)\prime} R_{t+1} \;\in\; \mathbb{R}^{P_1} \,,$$

where

$$\hat{\Psi}_{1,1} = (NT)^{-1} \sum_t S_t^{(1)\prime} S_t^{(1)} \;\in\; \mathbb{R}^{P_1 \times P_1} \,.$$

The ridge-regularized form is necessary for characterizing $\hat{\beta}(z;q)$ in the high complexity regime, $P/(NT) \to c > 1$ and $P_1/P \to q \in (0,1)$. The sample covariance matrix of signals, $\hat{\Psi}_{1,1} = (NT)^{-1} \sum_t S_t^{(1)\prime} S_t^{(1)} \in \mathbb{R}^{P_1 \times P_1}$, is central to understanding its behavior. RMT characterizes one aspect of $\hat{\Psi}$—the distribution of its eigenvalues—and fortunately, the prediction and portfolio performance properties of ridge estimators can be fully understood in terms of this distribution.

The portfolio strategy that we analyze is a linear timing strategy that utilizes predictions from the ridge regression as portfolio weights:

$$R_{t+1}^{\pi} = \pi_t' R_{t+1}, \quad \pi_t(\beta) = S_t^{(1)} \beta^{(1)}.$$

That is, it allocates across assets in proportion to their estimated expected return.[14] We study this strategy not only because it aligns with cross-sectional strategies commonly studied in the empirical literature,[15] but also because it is the most theoretically tractable strategy that exploits time variation and cross-sectional variation in expected returns arising from ridge forecasts. Furthermore, despite its simplicity, it is a close approximation to the conditional mean-variance optimal Markowitz strategy.[16]

3 Theoretical Properties of Machine Learning Estimators and Portfolios

Having defined our environment, we now present our main theoretical results, including: i) the theoretical expected return, variance, and Sharpe ratio of the timing strategy, ii) the theoretical out-of-sample R^2 given a level of complexity and penalty term, iii) the optimal penalty term z_* for a given level of complexity, iv) the characterization of the "Limits to Learning."

We are ultimately interested in understanding the portfolio properties of a feasible timing strategy,

$$\hat{\pi}_t \;=\; \pi_t(\hat{\beta}) = S_t^{(1)}\hat{\beta}^{(1)} \tag{7}$$

for some estimator $\hat{\beta}^{(1)}$ of the $\beta^{(1)}$ vector. The performance of this strategy is, of course, intimately tied to the prediction accuracy of the estimator $\hat{\beta}^{(1)}$, summarized by its expected mean square forecast error (MSE) on an independent test sample. This is the fundamental

[14] Absent information about $\beta^{(2)}$, Assumption 4 implies that $E[R_{t+1}|S_t^{(1)}] = S_t^{(1)}\beta^{(1)}$.

[15] For example, it is analogous to Fama-MacBeth cross-sectional regressions, which also construct portfolio weights that are linear in predictive signals. And the standard practice of conducting portfolio sorts based on predictions and then forming long-short spreads is qualitatively similar to using predictions as weights after applying a non-parametric binning step.

[16] We elaborate on this point in Appendix C. Improving upon this strategy would require the knowledge of the full covariance structure of observable signals with the unobservable ones, which is clearly unrealistic.

notion of estimator "risk" from statistical theory, though we use the term "MSE" here to avoid confusion with portfolio riskiness. We can write MSE as

$$MSE(\hat{\beta}) = E\left[\|R_{t+1} - S_t^{(1)}\hat{\beta}^{(1)}\|^2|\hat{\beta}^{(1)}\right] = \operatorname{tr} E[R_{t+1}R_{t+1}'] - 2\underbrace{E[\hat{\pi}_t'R_{t+1}|\hat{\beta}^{(1)}]}_{\substack{\text{Timing}\\ \text{Expected Return}}} + \underbrace{\operatorname{tr} E[\hat{\pi}_t\hat{\pi}_t'|\hat{\beta}^{(1)}]}_{\substack{\text{Timing}\\ \text{Leverage}}}. \tag{8}$$

In other words, the higher the strategy's expected return, the lower the MSE. And the larger the positions—or "leverage"—of the strategy, the larger the MSE. A timing strategy with a higher expected return corresponds to more predictive power, while higher leverage gives the strategy higher variance. These two objects, expected return and leverage of the timing strategy, appear repeatedly throughout our analysis, and are key to understand how complexity translates into higher strategy performance. The expected return/leverage tradeoff in (8) is a financial decomposition of MSE analogous to its statistical decomposition into a bias/variance tradeoff.

Next we turn to a characterization of the infeasible strategy performance and its dependence on the properties of Ψ and Σ matrices, the key determinants of the data generating process. The following is true.

Proposition 1 (Zero Complexity Limit of the Infeasible Strategy) *Let*

$$H(\Sigma) \;=\; \frac{\operatorname{tr}(\Sigma^2)}{(\operatorname{tr}(\Sigma))^2} \;\in\; [N^{-1}, 0]\,. \tag{9}$$

In the limit as $P \to \infty$, the conditional mean and variance to the infeasible market timing

strategy $\pi_t = S_t^{(1)}\beta^{(1)}$ *converge to*

$$E_t[\pi_t' R_{t+1}] \to_{P\to\infty} \operatorname{tr}(\Sigma) q b_* \psi_{*,1}(q) \tag{10}$$

$$\operatorname{Var}_t\left[(\pi_t' R_{t+1})^2\right] \to_{P\to\infty} \tag{11}$$

$$(q\psi_{*,1}(q) b_* \operatorname{tr}(\Sigma))^2 \left(H(\Sigma) \ + (q\psi_{*,1}(q) b_* \operatorname{tr}(\Sigma))^{-1} (1 + H(\Sigma) b_* \operatorname{tr}(\Sigma) \psi_{*,1}(1)) \right). \tag{12}$$

The infeasible market timing conditional Sharpe ratio is given by

$$SR_t(b_*) \to_{P\to\infty} \frac{1}{\sqrt{H(\Sigma) \ + (q\psi_{*,1}(q) b_* \operatorname{tr}(\Sigma))^{-1} (1 + H(\Sigma) b_* \operatorname{tr}(\Sigma) \psi_{*,1}(1))}} \tag{13}$$

At the same time,

$$R^2 \to \frac{\operatorname{tr}(\Sigma)(\beta^{(1)})' \Psi \beta^{(1)}}{\operatorname{tr}(\Sigma) \beta' \Psi \beta + N} \to \frac{\operatorname{tr}(\Sigma) b_* q \psi_{*,1}(q)}{\operatorname{tr}(\Sigma) b_* \psi_{*,1} + N}. \tag{14}$$

Both SR_t *and* R^2 *are monotone increasing in* b_* *and converge to*

$$\begin{aligned} \lim_{b_*\to\infty} SR_t(b_*) \ &= \ \frac{1}{((1 + \psi_{*,1}(1)/(q\psi_{*,1}(q))) H(\Sigma))^{1/2}} \ \leq \ \left(\frac{N}{2}\right)^{1/2} \\ \lim_{b_*\to\infty} R^2 \ &= \ \frac{q\psi_{*,1}(q)}{\psi_{*,1}(1)}. \end{aligned} \tag{15}$$

Proposition 1 shows that the out-of-sample performance of the infeasible strategy $\pi_t = S_t^{(1)}\beta^{(1)}$ depends only on four quantities:

- The size of β, as captured by b_*;
- The predictable variation in returns, due to observable signals,

$$E_t[\pi_t' R_{t+1}] \ = \ E[\pi_t' \pi_t] \ = \ \operatorname{tr}(\Sigma) q b_* \psi_{*,1}(q); \tag{16}$$

- The Herfindahl index $H(\Sigma)$ of the eigenvalues of Σ.

The quantity

$$\frac{(\text{tr}(\Sigma))^2}{\text{tr}(\Sigma^2)} \;=\; 1/H(\Sigma) \tag{17}$$

is commonly known in the machine learning literature as the *effective rank* of a matrix Σ. See, (Bartlett et al., 2020). For example, Σ has rank K and all its non-zero eigenvalues are identical, then $\frac{(\text{tr}(\Sigma))^2}{\text{tr}(\Sigma^2)} = K$. In general, $\frac{(\text{tr}(\Sigma))^2}{\text{tr}(\Sigma^2)}$ captures whether the assets have a factor structure when identified with their vectors of characteristics. When there is no factor structure, $H(\Sigma) \approx 1/N$ and hence vanishes when N is large. As we show below (see Section 4.2), Σ estimated from standard stock characteristics does have a low effective rank, representing natural limits to the performance of a simple, "linear-in-characteristics" strategy. By contrast, in the world of high complexity where signals S_t are constructed using thousands of non-linear transformations of characteristics (see Section 4.2), Σ has a very large effective rank, implying that $H(\Sigma)$ is close to zero.[17]

3.1 Out-of-sample performance

In this section, we derive the theoretical out-of-sample performance of the $\hat{\pi}_t(z)$ strategy as a function of it's complexity and shrinkage. We start with the expected return of the timing strategy. The following quantities will play a key role in our analysis:[18]

$$\begin{aligned} m(-z;cq;q) &= \lim_{T,P,P_1\to\infty,\ P/(NT)\to c,\ P_1/P\to q} P_1^{-1}\,\text{tr}((zI+\hat{\Psi}_{1,1})^{-1}) \\ m_\Sigma(-z) &\;=\; N^{-1}\,\text{tr}(zI+\Sigma)^{-1}\,. \end{aligned} \tag{18}$$

These are the so-called Stieltjes transforms of the eigenvalue distributions of the two matrices, $\hat{\Psi}_{1,1}$ and Σ.

[17] Note that, under Assumptions 1 to 4, $E[R_{t+1}] = 0$.

[18] See Section D where we explain the link between (18) and random matrix theory.

Proposition 2 *Suppose that* $\mathrm{tr}(\Psi_{2,1}\Psi_{1,2}) = o(P)$. *Then,*[19] *the expected return on the market timing strategy converges in probability to*

$$\mathcal{E}(z;cq;q) \coloneqq \lim E[\hat{\pi}_t(z)'R_{t+1}|\hat{\beta}] \;=\; \mathrm{tr}(\Sigma)b_* q\,\nu(z;cq;q)\,, \tag{19}$$

where

$$\nu(z;cq;q) \;=\; \psi_{*,1}(q) - (cq)^{-1}z\xi(z;cq;q)\,, \tag{20}$$

where $\xi(z;cq;q)$ *is the unique solution to*

$$1 \;=\; zm(-z;cq;q) + c^{-1}\left(1 - \xi(z;cq;q)^{-1}m_\Sigma(-\xi(z;cq;q)^{-1})\right)\,. \tag{21}$$

Expected leverage

$$\mathcal{L}(z;cq;q) \coloneqq \lim \mathrm{tr}\, E[\hat{\pi}_t(z)\hat{\pi}_t(z)'|\hat{\beta}] \tag{22}$$

is given by an explicit expression that only involves the function ξ *and its derivative. See Proposition 12 in th Appendix.*

Formula (19) shows that the expected return on the strategy depends on the bias introduced through the ridge penalty $z > 0$:

$$E[\hat{\pi}_t(z)'R_{t+1}|\hat{\beta}] \;=\; \mathrm{tr}(\Sigma)b_* q(\;\underbrace{\psi_{*,1}(q)}_{predictability} - \underbrace{(cq)^{-1}z\xi(z;cq;q))}_{bias}\,. \tag{23}$$

As $q \to 1$, we recover full predictability, as captured by $b_*\psi_{*,1}(1)$, net of the bias, $c^{-1}z\xi(z)$ which is monotone increasing in z. Thus, ridge penalization unambiguously reduces the

[19] Absent this condition, the expected return depends on the covariance structure of the signals. See Proposition 11 in the Appendix.

expected return of the timing strategy. However, at the same time, it significantly reduces the risk.

It follows directly from the definition that $\operatorname{tr} E[\hat{\pi}_t(z)\hat{\pi}_t(z)'|\hat{\beta}]$ is always monotone decreasing in z. Thus, ridge shrinkage introduces a bias-variance tradeoff whereby large z could potentially lead to higher out-of-sample performance. In the Appendix (see (22)) we show how the expected leverage (the variance part of the tradeoff) depends on the signal risk structure. When the model is highly misspecified, and the variance of the strategy depends on the missing signal variance capturing the *missing non-linearity of signals*: The inability of the mis-specified model to correctly approximate the true, highly non-linear structure of expected returns. Furthermore, additional variance of the strategy arises from the covariance of observed non-linearities with missing non-linearities, and may significantly increase leverage and, as a result, reduce the out-of-sample R^2. Proposition 12 in the Appendix shows how the magnitude of these effects can be computed explicitly in terms of the eigenvalue distributions of $\hat{\Psi}$ and Σ.

We can now compute the expected out-of-sample Sharpe ratio of the strategy. Perhaps surprisingly, it depends on only three quantities: The expected return $\mathcal{E}$, expected leverage $\mathcal{L}$, and the Herfindahl index $H(\Sigma)$.

Proposition 3 *In the limit as $T, P \to \infty$, $P/(NT) \to c$, we have:*

i) The conditional variance of the timing strategy converges in probability to

$$\begin{aligned} \mathcal{V}(z;cq;q) &:= \lim E\left[(\hat{\pi}_t(z)'R_{t+1})^2|\hat{\beta}\right] \;-\; (\mathcal{E}(z;cq;q))^2 \\ &= H(\Sigma)(\mathcal{E}(z;cq;q))^2 \;+\; (1 + b_*\psi_{*,1}(1)\operatorname{tr}(\Sigma)H(\Sigma))\,\mathcal{L}(z;cq;q) \end{aligned} \tag{24}$$

ii) The Sharpe ratio of the timing strategy converges to

$$SR(z;cq;c) = \frac{\mathcal{E}(z;cq;c)}{\sqrt{\mathcal{V}(z;cq;c)}} = \frac{1}{\sqrt{H(\Sigma) + (1 + b_*\psi_{*,1}(1)\operatorname{tr}(\Sigma)H(\Sigma))\,\frac{\mathcal{L}(z;cq;q)}{(\mathcal{E}(z;cq;q))^2}}}. \tag{25}$$

In the limit as $b_* \to \infty$, it converges to

$$\lim_{b_* \to \infty} SR(z; cq; c) \;=\; \frac{1}{(H(\Sigma)(1 + \frac{\psi_{*,1}\hat{\nu}(z;cq;c)}{\nu^2(z;cq;c)}))^{1/2}}\,. \tag{26}$$

When the single stocks signals have no factor structure, $H(\Sigma) \approx \frac{1}{N} \approx 0$ and idiosyncratic risk in the managed portfolios is almost perfectly diversifiable. As a result, portfolio variance is fully driven by leverage: $\mathcal{V}(z; cq; q) \approx \mathcal{L}(z; cq; q)$. However, factor structure in the signals, as captured by $H(\Sigma)$, increases portfolio variance. It does it trough two channels: Fluctuations in conditional expected returns; and fluctuations in the conditional expected variance, resperesented by the two terms in (24). In our empirical exercise, the total predictable variation in stock returns, $\beta' S_t' S_t \beta \approx b_* \psi_{*,1}(1) \operatorname{tr}(\Sigma)$ is large. As a result, changes in $H(\Sigma)$ have a large impact on the portfolio variance, $\mathcal{V}(z; cq; q)$.

Proposition 3 establishes a non-trivial link between the predictive performance (measured by the mean squared error) and out-of-sample Sharpe ratio: Both quantities are increasing in the expected return, $\mathcal{E}$, and decreasing in leverage, $\mathcal{L}$:

$$\begin{aligned} E[\|R_{t+1} - S_t\hat{\beta}(z)\|^2] \;&=\; E[\|R_{t+1}\|^2] - 2\mathcal{E}(z; cq; q) + \mathcal{L}(z; cq; q) \\ SR(z; cq; c) \;&=\; \frac{1}{\sqrt{H(\Sigma) + (1 + b_*\psi_{*,1}(1)\operatorname{tr}(\Sigma)H(\Sigma))\,\frac{\mathcal{L}(z;cq;q)}{(\mathcal{E}(z;cq;q))^2}}}. \end{aligned} \tag{27}$$

However, in general, a change in the data generating process that modifies both $\mathcal{E}$ and $\mathcal{L}$ may have an ambiguous effect on the two performance criteria, pushing one up and another down. Thus, there is generally no direct link between the model predictive ability and its out-of-sample Sharpe ratio. We illustrate this phenomenon in our empirical analysis (Section 4, Figures 4 and 5), where our best performing strategies often happen to have extremely poor prediction errors.

In general, the behavior of quantities $\mathcal{L}$ and $\mathcal{E}$ depends in a complex fashion on the correlations between observable and unobservable signals, as captured by the quantities

(89). However, when both quantities (89) are zero, expressions significantly simplify. For example, this is the case if the matrices $\Psi_{1,2}, \Psi_{2,1}$ have uniformly bounded traces; e.g., when $\Psi_{1,2}$ has a finite, uniformly bounded rank as P, $P_1 \to \infty$ (due to, say, a finite-dimensional factor structure in the signals). We thus obtain the following result.

Proposition 4 *Suppose that*[20] $\mathrm{tr}(\Psi_{1,2}\Psi_{2,1}) = o(P)$ *and that* $\frac{d}{dz}\gamma(z; cq; q)$ *is small.*[21] *Then,* $\widehat{\xi}_{2,1} = \xi_{2,1} = 0$. *Furthermore,*

(i) We have $\mathcal{E}(z; cq; c)$ *is monotone decreasing in* z *and, hence,* $0 < \mathcal{E}(z; cq; c) < \mathcal{E}(0; cq; c) < \mathcal{E}(0, 0; 0)$, *and*

(ii) both $R^2(z; cq; c)$ *and* $SR(z; cq; c)$ *are monotone increasing in* z *for* $z < z_*$ *where*

$$z_* \;=\; \frac{c}{b_*}\Big(1 \;+\; b_*\frac{\mathrm{tr}(\Psi_{2,2})}{P}\frac{\mathrm{tr}(\Sigma^2)}{\mathrm{tr}(\Sigma)}\Big)\,, \tag{28}$$

and monotone decreasing in z *for* $z > z_*$.

It is instructive to understand the nature of the optimal shrinkage parameter, z_*. As one can see from (28), it is determined by three forces: model complexity, c, the amount of predictability, b_*, and the degree of mis-specification, captured by excess volatility of the optimal portfolio due to omitted signals, $\frac{\mathrm{tr}(\Psi_{2,2})}{P}\frac{\mathrm{tr}(\Sigma^2)}{\mathrm{tr}(\Sigma)}$. Interestingly enough, the latter also increases in the dispersion of the eigenvalues of Σ. When the model is correctly specified (i.e., $P_1 = P$ and $q = 1$), we get a particularly simple and intuitive expression: $z_* = c/b_*$. One could think about z_* as measuring the *effective model complexity.* Even if c is low, learning the true β coefficient is hard when b_* is small and returns are dominated by noise. In this case, contrary to the conventional wisdom, even models with low c may require a very large shrinkage to achieve best out-of-sample performance.

[20] This is the case for example, when $\Psi_P = D_P + Q_P$ where $\limsup_{P\to\infty} \mathrm{rank}\, Q_P < \infty$, while D_P are diagonal matrices, and D_P, Q_P are uniformly bounded.

[21] This is the case, for example, when Σ has a small dispersion of eigenvalues or when ξ is small (which is true when either z or c is large).

As Proposition 4 shows, we can only estimate the optimal shrinkage z_* correctly if we assume negligible correlations between observable and unobservable signals (the missing non-linearities). In the data (see Section 4.2), this assumption does not hold. Computing the true optimal z_* thus requires estimating the signal correlation structure and the corresponding random matrix theory-based corrections. We leave this important direction for future research.

We complete this section with an extension of an important theoretical result from KMZ establishing the *virtue of complexity*. As KMZ explain, raising cq brings the usual statistical challenges of overparameterization while at the same time improving the empirical model's approximation of the true DGP. As KMZ show, the improvement in approximation justifies the statistical cost of heavy parameterization, leading to higher out-of-sample portfolio performance. The following is true.

Theorem 5 (Virtue of Complexity) *Suppose that signals are sufficiently mixed (so that $H(x;q)$ does not depend on q) and $\operatorname{tr}(\Psi_{1,2}\Psi_{2,1}) = o(P)$. Suppose also that $H(\Sigma)$ is sufficiently small. Then, with the optimal amount of shrinkage z_*, the Sharpe ratio $SR(z_*(q;c);cq;c)$ and $R^2(z_*(q;c);cq;c)$ are strictly monotone increasing and concave in $q \in [0,1]$.*

While we cannot establish the virtue of complexity theoretically when $H(\Sigma)$ is significantly different from zero, we can study it using a simulation exercise. We fix $b_* = 2$ and $c = 2$, but vary $q \in (0,1]$ and compute $SR(z;cq;c)$ for different z. We perform this exercise for different eigenvalue distributions of Σ and Ψ. To this end, we choose two parameters, $a_\Psi,\ a_\Sigma$, and define $\lambda_i(\Sigma) = q_\Sigma i^{a_\Sigma}$ and $\lambda_i(\Psi) = q_\Psi i^{a_\Psi}$, $i = 1, 2, \cdots N$ or P. We then normalize all eigenvalues so that $\operatorname{tr}(\Psi) = 1$ and $\operatorname{tr}(\Sigma) = N/12$, consistent with our empirical setup (see Lemma 1). We consider $a_\Sigma,\ a_\Psi \in \{0, 10\}$, corresponding to $H(\Sigma) = 0.0020, 0.0065$. Figure 1 shows that the Virtue of Complexity is observed independently of the eigenvalue distributions, although eigenvalue concentration does affect the magnitudes and the shapes of the VoC curves. More concentrated eigenvalues of Σ (corresponding to a larger a_Σ and and

higher $H(\Sigma)$) lower the Sharpe ratio level. At the same time, more concentrated eigenvalues of Ψ (higher a_Ψ and $H(\Psi)$) amplify Sharpe ratios because learning concentrated, lower dimensional signal dependence is easier in finite samples.

3.2 The Complexity Wedge

In high complexity settings ($c >> 0$), parameterization is extensive relative to the number of observations. This can be viewed as the "machine learning regime." In this regime, the law of large numbers breaks down, and "best" estimators fail to converge to the true parameters because there is not enough data to go around. As a result, complexity creates a wedge that is critical to understanding predictive performance, and one of our main theoretical contributions is to explicitly characterize this wedge.

The failure of the law of large numbers when $c > 0$ generates a gap between the trained model and the truth. This puts a finite cap on how much the trained model can learn, even in large samples and even with a correctly specified training model. Because the training model cannot fully hone in on the truth, its expected out-of-sample performance falls short of the true model's predictability. We call this wedge "Limits to learning,"

$$\text{Limits to Learning} = \text{Expected Performance(True Model)} - \text{Expected Out-of-sample Performance}$$

This wedge is linked to the well-known concept of over-fitting:

$$\text{Over-Fitting} = \text{In-Sample Performance - Expected Out-of-sample Performance}\,. \quad (29)$$

But while over-fitting describes the concept of over-parametrization and is often invoked to describe a failure to select the best model or its appropriate penalization, limits to learning capture a more overarching fundamental concept: how far the trained model is from the

best possible model in terms of expected performance. Note that, when the true model is complex and data is limited, the best possible model can never be estimated in practice.

To quantify these wedges, we use Propositions 1 and 3. As we show in the Appendix, under the hypothesis of Proposition 4,

$$\begin{aligned}\frac{(\mathcal{E}(z;cq;q))^2}{\mathcal{L}(z;cq;q)} \;\leq\; \mathcal{E}(z_*;cq;q) \;&=\; \mathrm{tr}(\Sigma)b_*q\nu(z_*;cq;c) \\ &\underbrace{=}_{(23)} \mathrm{tr}(\Sigma)b_*q\Big(\underbrace{\psi_{*,1}(q)}_{predictability} - \underbrace{(cq)^{-1}z_*\xi(z_*;cq;q)}_{optimal\ \ bias}\Big)\,,\end{aligned} \tag{30}$$

and the inequality turns into equality for $z = z_*$. Thus, the Complexity Wedge satisfies

$$\begin{aligned}&ComplexityWedge(z;cq;c) \;=\; 1 - SR(z;cq;c)/SR(0;q) \;\geq\; 1 - SR(z_*;cq;c)/SR(0;q) \\ &= \;1 - \left(\frac{H(\Sigma) + (1 + b_*\psi_{*,1}(1)\,\mathrm{tr}(\Sigma)H(\Sigma))\,\frac{1}{\mathrm{tr}(\Sigma)b_*q\psi_{*,1}(q)}}{H(\Sigma) + (1 + b_*\psi_{*,1}(1)\,\mathrm{tr}(\Sigma)H(\Sigma))\,\frac{1}{\mathrm{tr}(\Sigma)b_*q\nu(z_*;cq;c)}}\right)^{1/2}\end{aligned} \tag{31}$$

By (30), $ComplexityWedge$ is determined by the size of the optimal bias introduced. The function $z_*\xi(z_*;cq;q)$ is strictly monotone increasing in z_* and therefore, higher effective complexity z_* leads to a larger optimal bias and a larger wedge.

How big is the $ComplexityWedge$ and how does it depend on the nature of the data? Our framework allows us to answer this question by looking at the relationship between the wedge and the key parameters: complexity c, and $H(\Sigma)$, $\mathrm{tr}(\Sigma)$, b_*, as well as the eigenvalue distribution of Ψ.[22] In our simulation exercise, we pick $N = 500$, $T = 12$ (roughly consistent with our empirical analysis for the mega-cap universe), and vary P to produce models with increasing complexity. We assume that the model is correctly specified (no missing signals), corresponding to $q = 1$. That is, $c = P/(NT)$ is the "true" complexity of the data. As above,

[22] Indeed, our theoretical results imply that those parameters define the out-of-sample Sharpe ratios.

we parametrize eigenvalue distributions of Ψ and Σ to vary their concentration (Herfindahl index): $\lambda_i(\Sigma) = q_\Sigma\, i^{a_\Sigma}$, $i = 1, \cdots, N$, and $\lambda_i(\Psi) = q_\Psi\, i^{a_\Psi}$, $i = 1, \cdots, P$, where q_Σ, q_Ψ are normalization factors ensuring that $\text{tr}(\Psi)/P = 1$ and $tr(\Sigma) = N/12$.[23] We consider a_Σ, $a_\Psi \in \{0, 5, 10\}$, corresponding to $H(\Sigma) = 0.0020, 0.0065, 0.0115$, respectively. We always evaluate the complexity wedge at the optimal level of $z_* = c/b_*$.

Figure 2 shows the behaviour of the *ComplexityWedge* under the above specification. Our theory predicts the complexity wedge is increasing in the effective complexity $z_* = c/b_*$, and we confirm this conclusion with the simulation: Higher c or lower b_* lead to a higher complexity wedge. The pattern is uniform across different $H(\Sigma)$. We also observe that more concentrated eigenvalues in Ψ (higher $H(\Psi)$) are associated with faster decay of feasible Sharpe ratio, resulting in the complexity wedge increasing at a higher rate in c. In particular, when $\Psi = I$ (corresponding to $a_\Psi = 0$), complexity wedge increases to 80% for $c = 2$, implying that the infeasible Sharpe is five times higher than the feasible one. The intuition is clear: When signals are correlated, perfectly learning the whole β vector is not necessary to achieve good out-of-sample performance: It is enough to learn the projection of β on the top few principal components of the signals. By contrast, when signals are (approximately) uncorrelated, each β_i contributes equally to return predictability, and hence errors in β estimation have a much larger effect on the out-of-sample performance.

4 Empirical Results

We design experiments that mirror our theoretical environment and allow us to evaluate the role of complexity in modeling the cross-section of returns. Our experiments are couched in the machine learning method of random features regression (RFR) introduced in an influential paper Rahimi and Recht (2007) and developed for single return prediction in

[23] This is again consistent with our empirical setup. See Lemma 1.

KMZ and Kelly et al. (2022). Below we provide details of our methodology applied to panels of stocks.

Using a fixed set of 153 raw predictors from (Jensen et al., Forthcoming), we convert these into a set of P nonlinear "random features." These are exactly analogous to the features engineered in a hidden layer of a neural network.[24] By varying the number P of random features, RFR allows us to smoothly transition from low complexity models to high complexity models while holding the set of underlying predictor information fixed, with larger P models using that information more flexibly.

Our theoretical analysis in the previous section relies crucially on the assumption that the β vector is rotationally symmetric (Assumption 4), implying that all signals are roughly "equally predictive". This assumption is mirrored in our empirical analysis thanks to an especially attractive aspect of the random feature generation scheme: By construction, predictive information from all of the raw predictors is democratically distributed across the random features. Specifically, each random feature has an identical expected predictive ability in expectation. An implication of this is that the order of the random features is irrelevant—the first random feature is on the exact same footing as the last in terms of its predictive potential—so as we vary model size P, it doesn't matter how we work our way through the list of features. In short, the key point of our RFR formulation is that we are able to evaluate the empirical tradeoff of the effect of improved approximation afforded by larger P versus the statistical cost of estimating more parameters.

4.1 Data

We focus our empirical analysis on US stocks from 1969-01 to 2019-12 at the monthly frequency. We use the 153 stock characteristics data from (Jensen et al., Forthcoming), capturing a wide range of signals that have been shown to have return predictability, including

[24] Indeed, RFR is equivalent to a wide two-layer neural network with fixed (randomly drawn, as opposed to estimated) weights in the first and optimized weights in the second layer (estimated via ridge regression).

value, momentum, low-risk, accruals, profitability, skewness, etc.[25] For each characteristic, we first rank it in the cross-section within each month and normalize it to be contained between -0.5 and 0.5. The rank-standardization is standard in the literature (e.g., (Gu et al., 2020)) to remove the influence of outliers in the characteristics.

We perform our prediction exercise separately for stocks grouped by their market capitalization: mega (largest 20% of stocks based on NYSE breakpoints at each time period), large (between 80% and 50% percentile of NYSE breakpoints), small (between 50% and 20% percentile of NYSE breakpoints), and micro (between 20% and 1% percentile). This split allows us to investigate different stock return predictability across size groups. As shown later, our empirical results align with theory in each of the sub-groups, but we do find the amount of complexity and predictability differ significantly across size groups.

In our theoretical analysis, we work with covariance-normalized returns defined in (2).[26] We avoid this step because it introduces additional complications associated with modelling the conditional covariance matrix. Instead, we simply normalize the returns of all stocks by their historical daily standard deviation estimated on a one-month rolling window.[27]

4.2 Random Features

Formally, we observe a d-dimensional vector of stock characteristics $\tilde{S}_{i,t} = (\tilde{S}_{i,k,t})_{k=1}^{d}$, with $d = 153$ in our empirical analysis, for each stock $i = 1, \cdots, N$. We assume that there exists a non-linear function $f : \mathbb{R}^d \to \mathbb{R}$ such that

$$R_{i,t+1} = f(\tilde{S}_{i,t}) + \varepsilon_{i,t+1}, \ i = 1, \cdots, N. \tag{32}$$

[25] See Appendix P for the full set of characteristics used.

[26] Alternatively, one can first hedge away the exposures of all stocks to common factors, and then work directly with the residual returns.

[27] Despite this simplifying choice, our empirical results are strikingly consistent with theoretical predictions based on the (2) specification.

Our objective is to design a statistical procedure for (imperfectly) estimating f by approximating f with a linear combination of random features, which are non-linear transformations of random linear combinations of the raw features $\tilde{S}_t$. To this end, we pick an activation function $h(x) : \mathbb{R} \to \mathbb{R}$ and a probability density $\mu(\theta)$ on $\mathbb{R}^d$.

Proposition 6 *Let h be a smooth non-linear function. Suppose we sample θ_k randomly from a probability density $\mu(\theta)$ on $\mathbb{R}^d, k = 1, \cdots, P$. Suppose also that*

$$R = f(\tilde{S}) + \varepsilon \tag{33}$$

for some function f in the reproducing kernel Hilbert space for the kernel $K(x, y) = E[h(\theta' x)h(\theta' y)]$ and let $\|\cdot\|_H$ be the corresponding norm. Let $\tilde{f}_P(\tilde{S})$ be the projection of f onto the linear span of $h(\theta_k' \tilde{S})$, $i \leq P$. Then,

$$\lim_{P \to \infty} \|\tilde{f}_P - f\|_H^2 \;=\; 0 \tag{34}$$

By Proposition 6, we can pick P so large that the difference between f and $\tilde{f}$ is negligible. Thus, we may assume that

$$f(\tilde{S}) \;\approx\; \tilde{f}_P(\tilde{S}_i) \;=\; \sum_{k=1}^{P} \beta_k h(\theta_k' \tilde{S}) \;=\; \sum_{k=1}^{P} \beta_k S_k \tag{35}$$

for some coefficients β_k, where $S_k = h(\theta_k' \tilde{S})$ are the random features.

In practice the distribution of the random weights θ_i and the activation function $h(\cdot)$ are key modeling decisions, comparable to the activation function of a neural neural network (see, e.g. Gu et al., 2020).

We draw the random weights from a normal distribution $\theta_i \sim N(0, I\gamma)$ where γ is selected from a grid $[0.5, 0.6, .7, .8, .9, 1.0]$. We split the P features equally across the gamma such that $\frac{P}{6}$ we have random features drawn from each of possible γ values.

We use a relatively complex activation function: First, slightly deviating from Proposition 6, we apply two different activation functions, $h_1(x) = \cos(x)$ and $h_2(x) = \sin(x)$, obtaining two sub-groups of random features $h_1(\theta_k' \tilde{S}_{i,t})$, $h_2(\theta_k' \tilde{S}_{i,t})$ that we stack together.[28] Then, following the common ranking procedure used in cross-sectional asset pricing, we replace each $h(\theta_k' \tilde{S}_{i,t})$ for each fixed k with its cross-sectional rank among all signals $(h(\theta_k' \tilde{S}_{i,t}))_{i=1}^N$, normalize them so that they all belong to the $[-0.5, 0.5]$ interval, using the transformation

$$y_i = rank(x_i)/N - 0.5\,. \tag{36}$$

This way we obtain our final signals $S_{i,k,t}$. Cross-sectional ranking achieves two goals in our analysis. First, it ensures that all signals are cross-sectionally demeaned: $\sum_i S_{i,k,t} = 0$ for each k, t, and hence our predictions $\pi_{i,t} = \sum_k \beta_k S_{i,k,t}$ are also cross-sectionally demeaned, independent of the choice of β.[29] Second, it allows us to better control the behavior of the empirical estimates of the Σ and Ψ matrices, as is shown by the following lemma.

Lemma 1 *Let $S_t = (S_{i,k,t}) \in \mathbb{R}^{N \times P}$ be cross-sectionally ranked according to (36). Then, we have*[30]

$$(S_t' S_t)_{j,j} \;=\; \frac{1}{12} N + \frac{1}{6N} \tag{38}$$

[28] This choice corresponds to Random Fourier Features ((Rahimi and Recht, 2007)). See (see, e.g. Kelly et al., 2021) for details.

[29] Note that, since we work with vol-normalized returns (Assumption 2), our portfolios are not necessarily market neutral. We discuss implications of this fact below. In particular, we carefully control for the potential exposure to the idiosyncratic volatility factor.

[30]

$$\begin{aligned}
(S_t' S_t)_{j,j} \;&= \sum_i S_{i,j}^2 \;=\; \sum_i (i - 0.5N)^2/N^2 \;=\; N^{-2}\Big(\sum_{i=1}^N i^2 - N \sum_i \big(i + 0.25N^2\big)\Big) \\
&= \; N^{-2}(6^{-1} N(N+1)(2N+1) - N 0.5 N(N+1) + 0.25N^3) \\
&= \frac{1}{12} N + \frac{1}{6N}
\end{aligned} \tag{37}$$

We estimate the Σ matrix using

$$\hat{\Sigma} \;=\; (PT)^{-1}\sum_{t} S_t S_t'\,. \tag{39}$$

By Lemma 1,

$$\mathrm{tr}(\hat{\Sigma}) \;=\; (PT)^{-1}\sum_{t}\mathrm{tr}(S_t S_t') \;=\frac{1}{12}N+\frac{1}{6N}\,. \tag{40}$$

4.3 Training Procedure

To alleviate the high computational costs[31] of our empirical analysis, we train a new model every quarter and use the β coefficients estimated in that model to produce stock return forecasts over the next three months. That is, our β vectors are re-estimated quarterly, but conditional expected returns are based on the most recent signals, S_t. We train each model using a rolling window of 12 months and keep all stocks in the market-size groups for which we have observations for the full 12 month period. We fill missing characteristics values with zeros for all 153 characteristics. Given that we rank-standardize characteristics in the cross-section, filling missing values with zero is equivalent to replacing missing value with the cross-sectional median. Figure 3 shows the average fraction of missing signals across the 153 characteristics over time. The full-sample average of missing values are 0.11 for Mega, 0.15 for Large, 0.22 for Small, and 0.29 for Micro.

4.4 Portfolio Construction

Given a large P and the signal matrix $S_t \in \mathbb{R}^{N\times P}$ (see Section 4.2 for details of signal construction), we pick a grid of $P_1 = 100, 1000, \cdots, P$. For each P_1, we define $S_t^{(1)}$ to be the

[31] In our study of complexity, we consider models with up to 250000 random features. For a panel of $N = 2000$ stocks and $T = 12$ periods, we thus need to run a regression using a 24000×250000 matrix.

matrix in the first P_1 columns of S_t and then we compute

$$\hat{\beta}(z;q) \;=\; \left(zI + \hat{\Psi}_{1,1}\right)^{-1} \frac{1}{NT} \sum_t S_t^{(1)\prime} R_{t+1} \;\in\; \mathbb{R}^{P_1} \tag{41}$$

where

$$\hat{\Psi}_{1,1} = (NT)^{-1} \sum_t S_t^{(1)\prime} S_t^{(1)} \;\in\; \mathbb{R}^{P_1 \times P_1}\,. \tag{42}$$

Recall that $c = P/(NT)$ and $q = P_1/P \in (0,1]$. We then construct the complex linear timing strategy using

$$R_{t+1}^{Complex}(z;q) \;=\; R_{t+1}' S_t^{(1)} \hat{\beta}(z;q)\,. \tag{43}$$

In addition to (43), we also consider other cross-sectional strategies that we use for benchmarking. Namely, we define

$$R_{t+1}^{(m)} = \sum_{i=1}^{N} S_{i,t}^{(m)} R_{i,t+1},\; m \in \{Linear,\; EW\}\,, \tag{44}$$

where $S_{i,t}^{(m)}$ is the signal for stock i at time t estimated under one of the following models:

- $m = Linear$: We compute $S_{i,t}^{(linear)} = S_{i,t}^{(linear)}(z)$ by estimating a panel linear ridge regression on the original 153 ranked signals $\tilde{S}_{i,k,t}$ (see Section 4.2). We use the same rolling window of 12 months and the same grid of ridge penalties, z.

- $m = EW$: $S_{i,t}^{(EW)}$ is the *feasible* equal-weighted combination of factors, defined as the sum of the signed original ranked 153 factors, where the sign is defined with a lagged expanding window to avoid any look-ahead bias.

  To understand our construction, consider a researcher who believes that there is a relationship between $\tilde{S}_{i,k,t}$ and $R_{i,t+1}$. A common way for testing this relationship is to

construct the linear trading strategy with returns

$$\tilde{F}_{t+1} \;=\; \sum_{i=1}^{N} \tilde{S}_{i,k,t} R_{i,t+1} \tag{45}$$

An academic researcher backtests this strategy and determines is sign:

$$\mathcal{S}_t(k) \;=\; \text{sign}(\sum_{t=1}^{\tau} \tilde{F}_{t+1}) \tag{46}$$

where τ is the number of periods available to the research at the time of publication (usually, multiple decades of data). However, this factor is infeasible because computing $\mathcal{S}$ requires knowledge of the future and creates a very large look-ahead bias. Instead, our factor is feasible and uses

$$\mathcal{S}_t(k) \;=\; \text{sign}(\sum_{\tau=1}^{t-1} \tilde{F}_{\tau+1}) \tag{47}$$

to define a feasible trading strategy $\mathcal{S}_t F_{t+1}$. The feasible equal-weighted signal is then defined as[32]

$$S_{i,t}^{(EW)} \;=\; \sum_{k=1}^{153} \mathcal{S}_t(k) \tilde{S}_{i,k,t}\,. \tag{48}$$

To show that our analysis is robust to classical portfolio construction methodologies, we

[32]The difference between feasible and infeasible equal weighted (EW) strategies is small: The feasible strategy EW for the four groups (mega, large, small, micro) yields a Sharpe ratio of 0.54, 0.64, 0.73 and 1.07 respectively. By contrast, the corresponding infeasible Sharpe ratios for those same groups are as high as 0.68, 0.55, 0.72, and 1.17. Another alternative to the feasible equal-weighted strategy would be to use the rolling window of $T = 12$ instead of the expanding window for computing the sign $\mathcal{S}_t(k)$, to be fully consistent with the algorithm used for the Complex model. We have analyzed this strategy, and its performance is much worse than that for the expanding window due to large oscillations in factor returns over a short rolling window of 12 months. Indeed, the Sharpe ratios for the four groups obtained with such a strategy are equal to 0.29, 0.33, 0.48, and 0.83.

also build long-short decile portfolios based on the the different $S_{i,t}^{(m)}$, $m \in \{Complex,\ Linear, EW\}$. To this end, we sort the cross sections of the signals $S_{i,t}^{(m)}$, group stocks into 10 deciles, and construct 10 equal-weighted portfolios within each decile. We then define the long-short portfolio by subtracting the 1^{st} deciles' portfolios to the 10^{th}, based on the signals

$$\hat{S}_{i,t}^{(m)} \ = \ \mathbf{1}_{S_{i,t}^{(m)} > Q_t} - \mathbf{1}_{S_{i,t}^{(m)} < q_t}\,, \tag{49}$$

where q_t and Q_t are the time$-t$ cross-sectional 10% and 90% quantiles of the cross-section $(S_{i,t}^{(m)})_{i=1}^{N}$ of signals, respectively.

4.5 Virtue of Complexity (VoC)

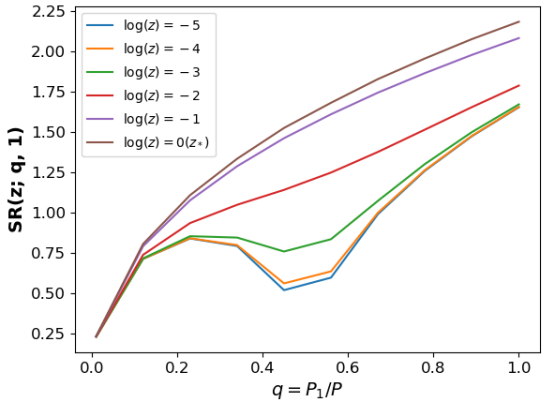
Our first empirical result in this section shows that, consistent with Theorem 5, there is a "virtue of complexity" in predicting the cross-section of stock returns: The expected out-of-sample R^2 and portfolio Sharpe ratio are generally increasing in the degree c of model complexity. Figure 4 show the out-of-sample performance of our models defined in terms of Sharpe ratio. First, we see that low values of z (and, in particular, the ridgeless limit when $z \approx 0$) lead to a non-monotonic, *double ascent* behavior (KMZ) whereby Sharpe ratio first increases in the low-complexity regime, then collapses around $cq = P_1/(NT) \approx 1$, and then increases again. The patterns of double ascent are not as pronounced as in KMZ because $c = P/(NT)$ has a time varying N. Since N is varying a lot over time, so does the model complexity, leading to a "floating" interpolation boundary. For large values of z, double ascent turns into a permanent ascent, and Sharpe ratio is monotone increasing in complexity.

Figure 4 highlights another important aspect of the dependence of model performance on complexity. Models with a moderate P do achieve comparably high Sharpe ratio, but achieving it require a very precise knowledge of the optimal shrinkage parameter z. By

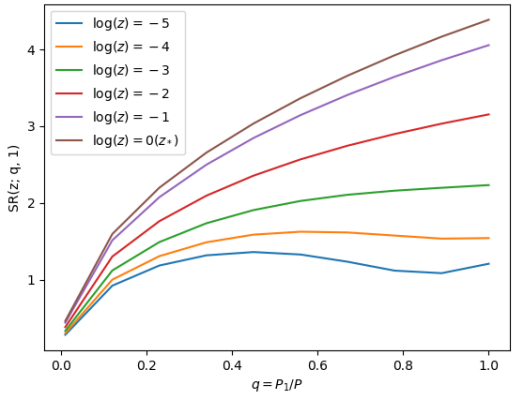
contrast, the performance of highly complex models is much less sensitive to the exact choice of z, making them more robust and easier to optimize.

Figure 5 displays the same results with performance defined in term of the out-of-sample R^2.[33] In both the R^2 and Sharpe ratio figures, we note that the performance seems to plateau around $P = 10^4$. However, for high levels of complexity, the discrepancy between the behaviors of Sharpe ratios (Figure 4) and R^2 (Figure 5) becomes significant: While high complexity models require a very large shrinkage to achieve a high out-of-sample Sharpe ratio, the same models achieve a high R^2 with a very low shrinkage. This observation highlights the important distinction between R^2 and Sharpe ratio: Highly complex models with an (inefficiently) large level of shrinkage typically have a negative R^2 and a high, positive Sharpe ratio.

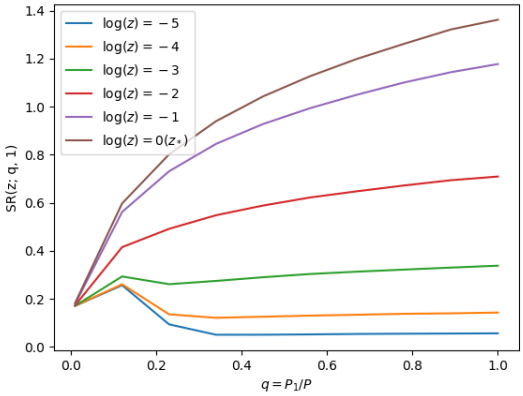
[33] We define the out-of-sample R^2 against a prediction of 0, as in (Gu et al., 2020)



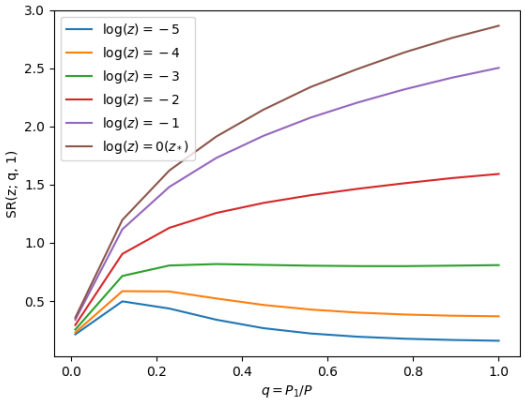
(a) $a_\Psi = 0, a_\Sigma = 0.$



(b) $a_\Psi = 10, a_\Sigma = 0.$

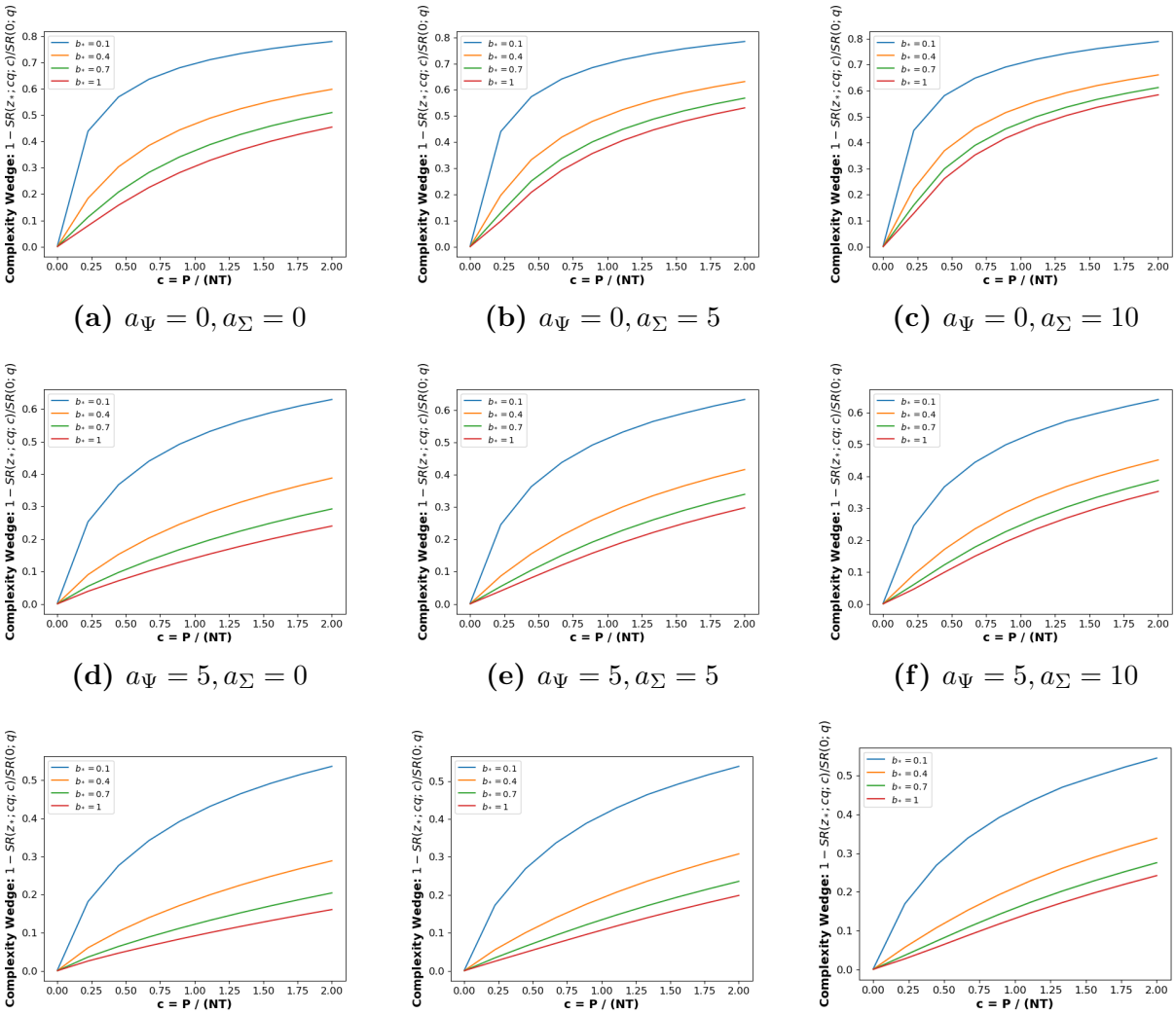


(c) $a_\Psi = 0, a_\Sigma = 10.$



(d) $a_\Psi = 10, a_\Sigma = 10.$

Figure 1: Simulation results of $SR(z; cq; c)$ under different $H(\Sigma)$ and $H(\Psi)$. a_Ψ and a_Σ are parameters governing the distribution of eigenvalues of Ψ and Σ: $\lambda_i(\Sigma) \sim i^{a_\Sigma}$, $\lambda_i(\Sigma) \sim i^{a_\Psi}$ where $i = 1, 2, 3 \cdots$. For Σ, $H(\Sigma) = tr(\Sigma^2)/tr(\Sigma)^2$ and $H(\Psi)$ is similarly defined for Ψ. We choose $N = 500, T = 12$ and we set $c = 2$ (corresponding to $P = 2NT$) and $b_* = 2$. $a_\Sigma = 0$ and 10 correspond to $H(\Sigma) = 0.002$ and 0.0065, respectively. The dip in the Sharpe ratio happens around the interpolation boundary, corresponding to $P_1 = NT = P/2$.



(a) $a_\Psi = 0, a_\Sigma = 0$ **(b)** $a_\Psi = 0, a_\Sigma = 5$ **(c)** $a_\Psi = 0, a_\Sigma = 10$

(d) $a_\Psi = 5, a_\Sigma = 0$ **(e)** $a_\Psi = 5, a_\Sigma = 5$ **(f)** $a_\Psi = 5, a_\Sigma = 10$

(g) $a_\Psi = 10, a_\Sigma = 0$ **(h)** $a_\Psi = 10, a_\Sigma = 5$ **(i)** $a_\Psi = 10, a_\Sigma = 10$

Figure 2: Simulation results of complexity wedge (31) for different b_* and c and different $H(\Sigma)$ and $H(\Psi)$, evaluated at $z_* = c/b_*$. Parameters a_Ψ and a_Σ govern the distribution of eigenvalues of Ψ and Σ: $\lambda_i(\Sigma) \sim i^{a_\Sigma}$, $\lambda_i(\Sigma) \sim i^{a_\Psi}$ where $i = 1, 2, 3 \cdots$. We choose $N = 500, T = 12$.

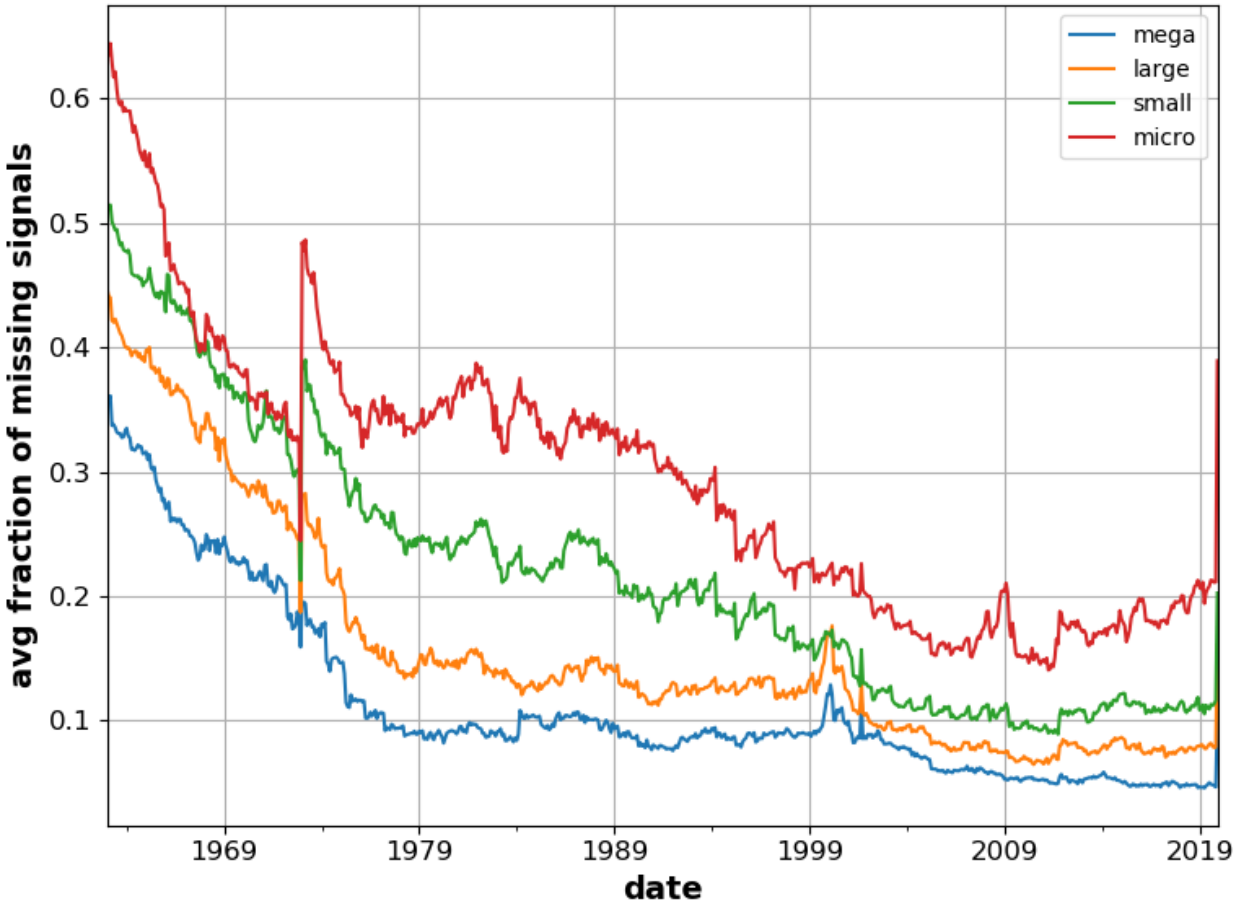
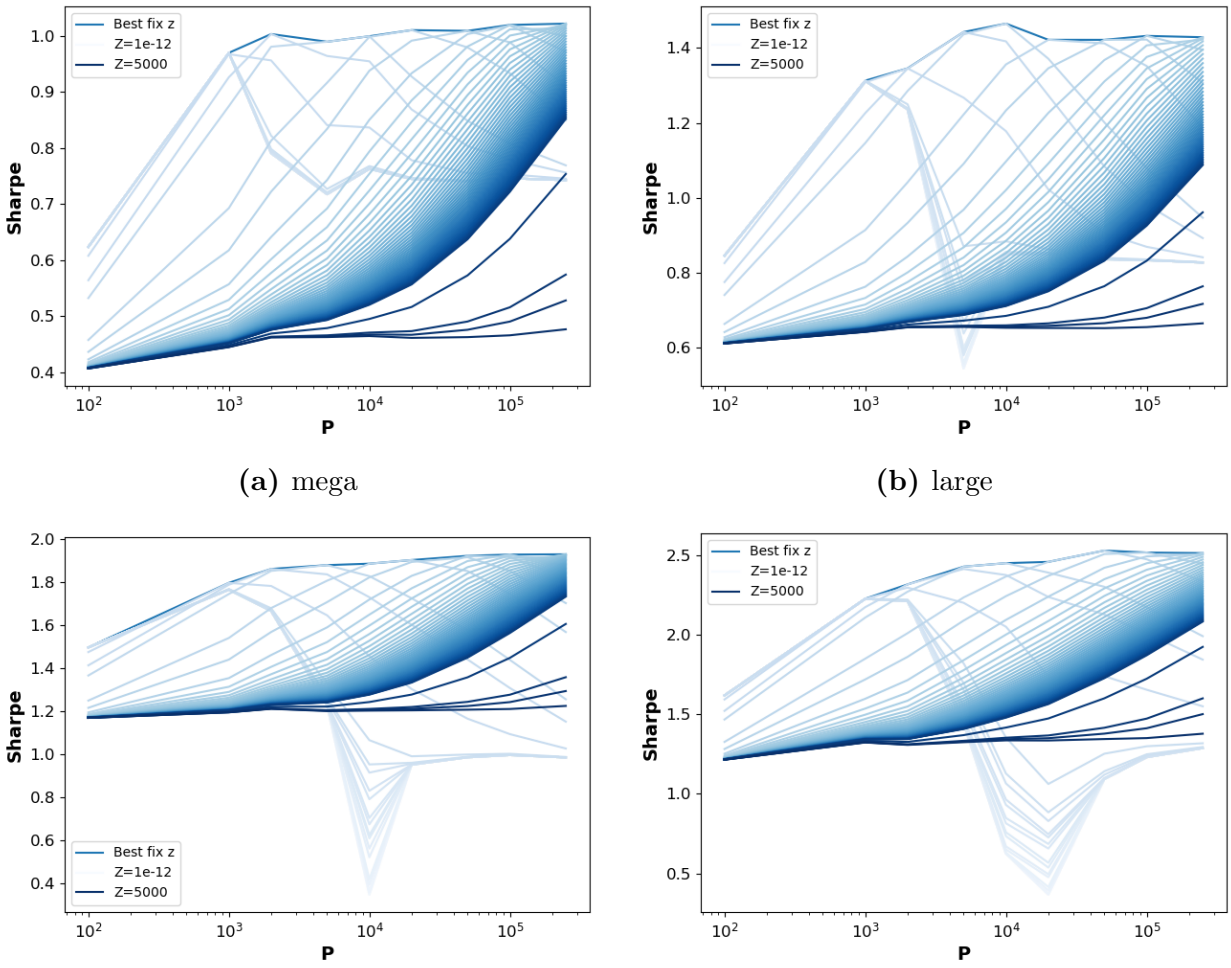


Figure 3: This figure shows the average fraction of missing signals across the 153 JKP characteristics over time. For each date, we compute for each individual signal the fraction of stocks that are missing this signal. We then average it across all 153 JKP signals.



(a) mega

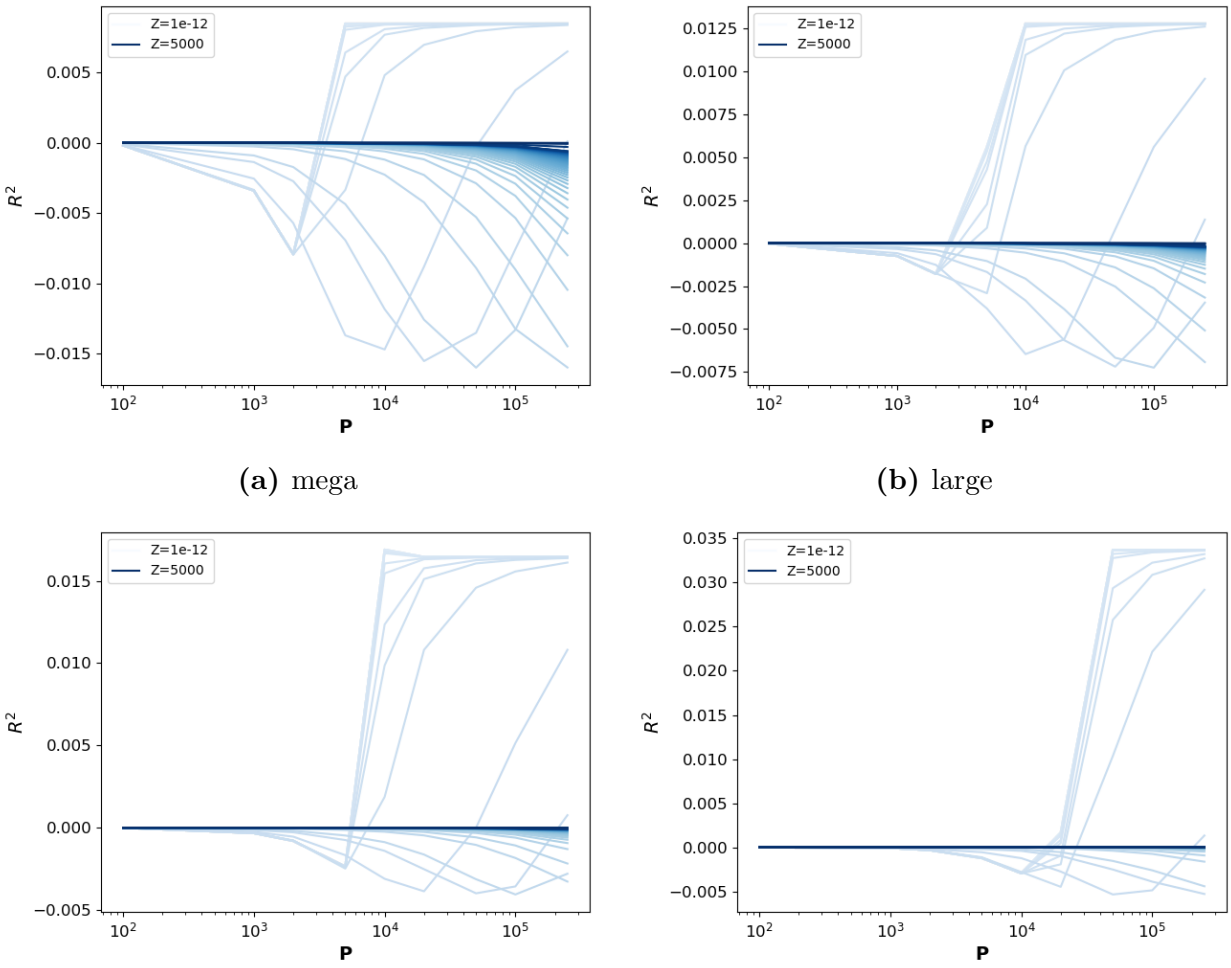
(b) large

(c) small

(d) micro

Figure 4: The figures above illustrate VOC in term of portfolio performance. On the x-axis, we show, with a log scale, the number of features P. Each colored line corresponds to the performance obtained with a different penalty term z. The darkest blue indicates our largest z (5000), and the lightest blue our smallest z (1e-12).

.



(a) mega

(b) large

(c) small

(d) micro

Figure 5: The figures above illustrate VOC in term of predictive power. On the x-axis, we show the number of features P on a log scale. On the y-axis, we show the out-of-sample R^2. Each colored line shows the performance computed for a different fixed penalty term z. The darkest blue indicates our largest z (5000), and the lightest blue our smallest z (1e-12).

4.6 Complexity Versus Simplicity

Theorem 5 (the virtue of complexity) is a claim about absolute performance. It has no implications about relative performance of the complex strategy with respect to simpler benchmarks. The goal of this section is to understand whether the highly non-linear strategies produced by the random features regression indeed capture return predictability that goes beyond simple, linear relationships.

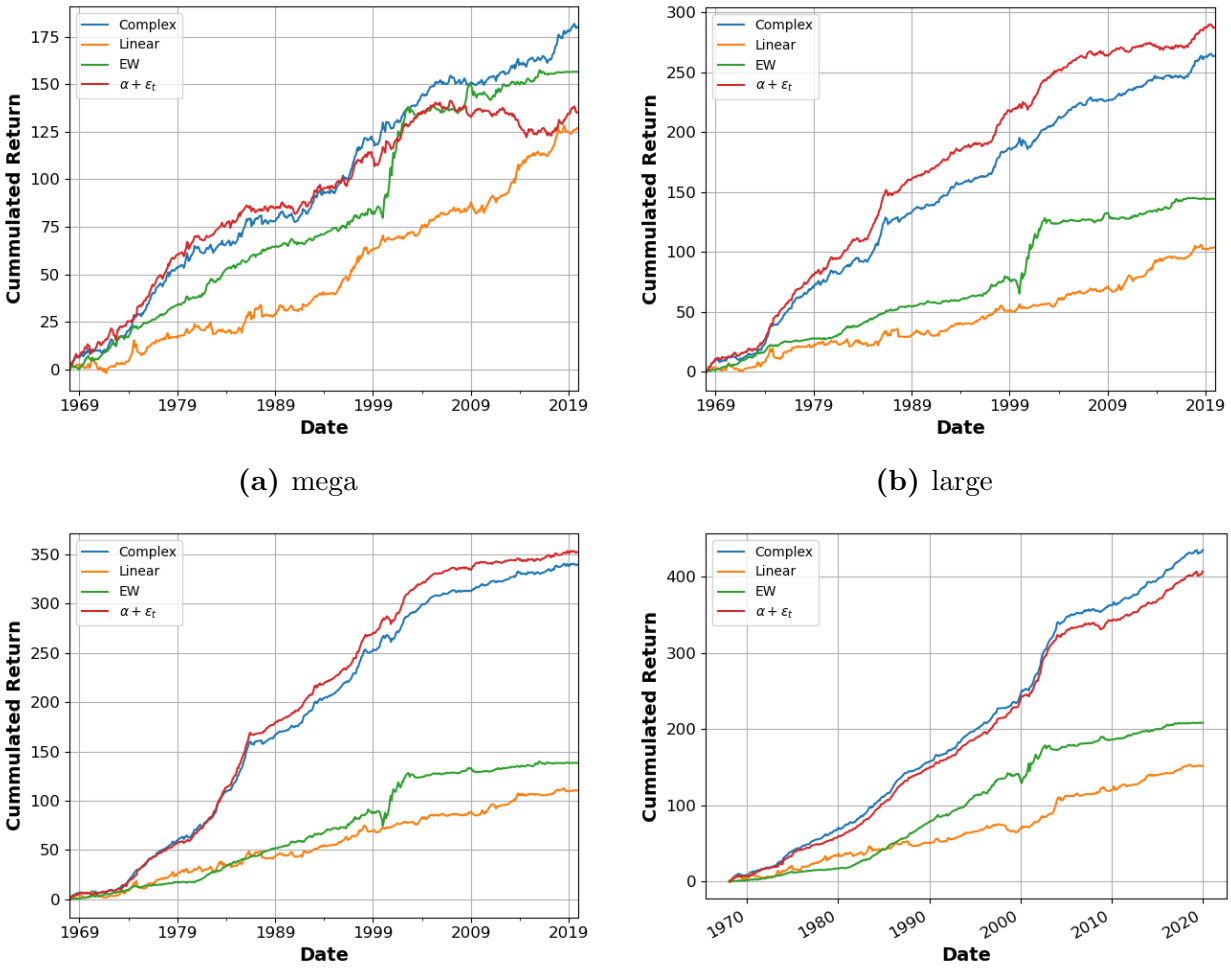
We start our analysis by investigating the relative performance of the complex strategies with respect to the two benchmark portfolios, *Linear* and *EW*, constructed in Section 4.4. To this end, we first construct the residuals from the regression

$$R_t^{(Complex)} = \alpha + \beta_1 R_t^{(Linear)} + \beta_2 R_t^{(EW)} + \epsilon_t \tag{50}$$

and investigate their dynamics. Figure 6 shows the cumulative performance of the residual $\alpha + \epsilon_t$ in (50), as well as the performance of the the complex strategy ($P = 10^4$) and the two linear benchmarks. To allow comparison across time series, we normalized each strategy by its standard deviation.

Two important conclusions can be drawn from this figure. First, while we see that the complex model always outperforms the linear counterparts, this over-performance is larger in smaller firms. We propose two non-mutually exclusive explanations: i) the higher costs associated with implementing the trading strategy with small firms prevents arbitrageurs from removing non-linear predictability; and ii) the linear factors are, in part, the product of data-mining on large firms in the economy (McLean and Pontiff, 2016). Second, we note that the residuals the complex strategy ($\alpha + \epsilon_t$) are high but tend to flatten out toward the end of the sample. As before, we note that this pattern is significantly more pronounced for large firms. This result is coherent with the idea that the amount of capital deployed in quantitative arbitrage strategies exploiting complex machine learning-based strategies has

been monotonically increasing since the beginning of the 21^{st} century, but large trading costs prevented the arbitrageurs from fully removing non-linear return predictability for smaller firms.



(a) mega

(b) large

(c) small

(d) micro

Figure 6: In the figure above, we show: a) the cumulative performance of our complex model with $P = 10^4$; b) the cumulative performance of our two linear benchmarks *EW* and *Linear*, and c) cumulative performance of the residuals $\alpha + \epsilon_t$, estimated from: $R_t^{(Complex)} = \alpha + \beta_1 R_t^{(Linear)} + \beta_2 R_t^{(EW)} + \epsilon_t$. We normalize all four time series by their full sample standard deviation.

Having established that the virtue of complexity originates from a genuinely non-linear predictive model, we now proceed to show that this performance is not driven by an exposure

to standard equity factors. To this end, we estimate of the following regressions:

$$R_t^{(m)} = \alpha + \beta_1 MKT_t + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 RMW_t + \beta_5 CMA_t + \beta_6 VOLA_t + \beta_7 MOM_t + \beta_8 R_t^{(Linear)} + \beta_9 R_t^{(EW)} + \epsilon_t \,. \quad (51)$$

Here, $MKT_t, SMB_t, HML_t, RMW_t, CMA_t$ and MOM_t are the returns of the standard Fama-French 5 factors (Fama and French, 2015) and the momentum factor (Carhart (1997)), respectively. $VOLA_t$ is the Fama and French idiosyncratic volatility factor constructed using data (Jensen et al., Forthcoming).[34] The inclusion of the $VOLA_t$ factor is crucial for our analysis because, since we work with volatility-normalized stocks returns (see

Table 1 reports the t-statistics of α from the regression (51). We can clearly see that α is significant for all size groups (mega, large, small, and micro), complexity levels, and configurations of the regression. More importantly, we see that the t-values increase in P and, once again, plateau around $P = 10^4$. Thus, more complex models achieve superior performance by introducing capturing non-linearities that are not spanned by standard factors and less complex models.

The signals $S_{i,t}^{Complex}$ are built as models for expected returns and the performance of the cross-sectional strategy $R_{t+1}^{Complex}$ discussed above shows that $S_{i,t}^{Complex}$ indeed contain information about the cross-section of expected stock returns. Another common way of testing the cross-sectional relationship between $S_{i,t}^{Complex}$ and $R_{i,t+1}$ is by building the long-short decile portfolios (see section 4.4).

For consistency, we perform the same decile transformation (49) with all three models, $S_{i,t}^{(m)}$, $m \in \{Complex,\ Linear,\ EW\}$, and repeat the analysis of Table 1 for these decile-based strategies in Tables 2 and 3. We find that α is positive and statistically significant for all but the *mega* firm group. This means that the complex model achieves its superior

[34] We obtain the factor returns from the Kenneth French website, and the Global Factor Data https://jkpfactors.com/

Table 1: In the table below, we show the performance of three different types of models: *Linear*, *EW*, and *Complex*. See Section 4.4. For each model, we show the Sharpe ratio (SR) and the t-statistics of α from (51).

		Linear	EW (OOS)	P=100	P=1000	P=2000	P=5000	P=10000	P=20000	P=50000	P=100000	P=250000
Mega	EW (OOS)	5.13	0.00	2.93	5.75	6.01	6.32	6.73	6.76	6.77	6.79	6.76
	Linear	0.00	6.31	1.04	3.36	3.02	3.12	3.25	2.80	3.11	3.18	3.28
	FF6	5.53	5.44	2.71	5.14	5.37	5.53	5.61	5.61	5.60	5.67	5.68
	FF6+VOLA	5.53	5.46	2.70	5.13	5.36	5.52	5.60	5.60	5.59	5.66	5.68
	FF6+VOLA+Linear+EW	0.00	0.00	1.27	3.80	4.19	4.31	4.62	4.43	4.65	4.66	4.68
Large	EW (OOS)	4.53	0.00	4.42	8.04	9.02	9.65	9.80	9.86	9.82	9.81	9.72
	Linear	0.00	6.06	2.53	4.59	5.45	6.17	6.24	5.75	5.60	5.68	5.59
	FF6	4.35	4.72	3.06	5.08	6.05	7.10	7.43	7.02	6.69	6.78	6.64
	FF6+VOLA	4.34	4.72	3.06	5.07	6.04	7.09	7.42	7.01	6.69	6.77	6.63
	FF6+VOLA+Linear+EW	0.00	0.00	3.17	6.32	7.35	8.27	8.49	8.51	8.43	8.43	8.31
Small	EW (OOS)	5.01	0.00	6.92	11.54	12.29	12.85	12.62	13.00	12.84	12.84	12.90
	Linear	0.00	6.03	7.49	10.65	10.99	10.57	10.91	10.91	11.02	11.02	11.05
	FF6	4.33	4.40	6.56	10.29	10.84	10.63	10.67	10.81	10.78	10.78	10.82
	FF6+VOLA	4.32	4.40	6.59	10.39	10.97	10.71	10.76	10.89	10.86	10.86	10.90
	FF6+VOLA+Linear+EW	0.00	0.00	6.24	10.57	11.32	11.88	11.59	11.95	11.77	11.77	11.81
Micro	EW (OOS)	7.04	0.00	5.56	10.83	11.85	12.37	13.13	12.11	13.34	13.21	13.07
	Linear	0.00	9.04	6.42	11.70	12.08	12.12	12.26	11.87	12.32	12.24	12.35
	FF6	5.96	7.54	5.84	11.00	11.45	11.50	11.62	11.23	11.61	11.56	11.67
	FF6+VOLA	5.95	7.53	5.84	11.01	11.45	11.51	11.63	11.24	11.63	11.59	11.69
	FF6+VOLA+Linear+EW	0.00	0.00	4.36	9.11	9.97	10.59	11.29	10.24	11.37	11.24	11.10

performance by doing something akin to ranking stocks using their linear-regression-based expected returns. This result is illustrates the power of the complex non-linear models: They are able to discover subtle non-linear relationship that would normally require human expert knowledge.

Looking at the other regression coefficients, we note that, perhaps surprisingly, while the complex strategy correlates with the decile-transformed linear model, there is no clear association between standard asset pricing factors and our complex long-short strategies.

We report more summary statistics for complex signals transformed using (49) in Tables 6-11 in Appendix O.

Table 2: In the table below, we use long-short (H-L) decile strategies defined for table 4 to 11, for the Complex portfolio, Linear, and EW strategies. The Complex strategy is defined by $P = 250,000$ and z maximizing the Sharpe ratio of the long-short strategy. We show the results of various combinations of the regression:
$R_t^{(m)} = \alpha + \beta_1 + MKT_t + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 RMW_t + \beta_5 CMA_t + \beta_6 MOM_t + \beta_7 VOLA_t + \beta_8 R_t^{(Linear)} + \beta_9 R_t^{(EW)} + \epsilon_t$
We show the results for the mega and large samples. The '*' defines the degree of statistical significance: * 10%, ** 5%, *** 1%. The values in parenthesis show the standard deviation of the coefficient.
We also show the information ratio defined as the annualized Sharpe ratio of the time-series $\alpha + \epsilon_t$

	Mega					Large				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
α	0.1634***	0.0287*	0.0272*	0.1862***	0.0230	0.1907***	0.0674***	0.0733***	0.1845***	0.0695***
	(0.0223)	(0.0151)	(0.0152)	(0.0237)	(0.0162)	(0.0165)	(0.0117)	(0.0118)	(0.0171)	(0.0123)
Linear		0.8882***	0.8834***		0.8840***		0.6481***	0.6575***		0.6571***
		(0.0296)	(0.0302)		(0.0304)		(0.0236)	(0.0236)		(0.0237)
EW	1.9823***		0.2606		0.2749	0.0643		-0.6508***		-0.6458***
	(0.4689)		(0.3097)		(0.3159)	(0.3052)		(0.2054)		(0.2100)
MKT_t				-0.2597	0.1517				0.6122	0.0210
				(0.5762)	(0.4023)				(0.4160)	(0.3039)
SMB_t				-0.0600	0.2332				0.8682	0.1503
				(0.8108)	(0.6625)				(0.5855)	(0.4993)
HML_t				0.0661	0.4096				0.5841	0.2814
				(1.1405)	(0.7828)				(0.8236)	(0.5900)
RMW_t				-1.4611	-0.7611				-0.3278	0.9180
				(1.1092)	(0.9529)				(0.8010)	(0.7231)
CMA_t				-0.2377	-0.0960				-0.4927	0.0738
				(1.7026)	(1.1276)				(1.2292)	(0.8497)
MOM_t				0.1128	0.3903				0.4846	0.4346
				(0.5491)	(0.3551)				(0.3964)	(0.2674)
$VOLA_t$					-0.4225					0.8150
					(0.6834)					(0.5162)
Info. Ratio	1.036	0.281	0.266	1.166	0.226	1.643	0.864	0.946	1.6	0.904
Observations	623	623	623	623	623	624	624	624	624	624
R^2	0.0264	0.5908	0.5906	-0.0065	0.5879	-0.0015	0.5471	0.5536	0.0042	0.5554

Table 3: The table below show the analysis presented in Table 3 on the small and micro sample.

	Small					Micro				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
α	0.1985***	0.0567***	0.0659***	0.1821***	0.0649***	0.2294***	0.0632***	0.0654***	0.2167***	0.0627***
	(0.0138)	(0.0112)	(0.0114)	(0.0141)	(0.0117)	(0.0121)	(0.0096)	(0.0100)	(0.0118)	(0.0101)
Linear		0.5773***	0.5778***		0.5693***		0.5865***	0.5867***		0.5753***
		(0.0242)	(0.0240)		(0.0241)		(0.0210)	(0.0211)		(0.0212)
EW	-0.5336**		-0.5548***		-0.5050***	-0.0539		-0.0805		-0.0466
	(0.2073)		(0.1492)		(0.1518)	(0.1606)		(0.1071)		(0.1098)
MKT_t				0.9157***	0.2003				0.8171***	0.2241
				(0.3438)	(0.2719)				(0.2871)	(0.2127)
SMB_t				0.4293	-0.2204				1.0385**	0.2873
				(0.4839)	(0.4456)				(0.4042)	(0.3491)
HML_t				0.6257	0.6361				0.0856	0.0237
				(0.6806)	(0.5274)				(0.5685)	(0.4125)
RMW_t				-1.1375*	-0.0577				-0.9097	-0.2672
				(0.6620)	(0.6423)				(0.5529)	(0.5014)
CMA_t				-0.6971	-0.4471				1.4051*	1.3219**
				(1.0159)	(0.7573)				(0.8486)	(0.5922)
MOM_t				0.8374**	0.5395**				0.4925*	0.1064
				(0.3276)	(0.2393)				(0.2737)	(0.1875)
$VOLA_t$					0.6937					0.2624
					(0.4599)					(0.3596)
Info. Ratio	2.056	0.808	0.949	1.911	0.947	2.809	1.161	1.2	2.722	1.167
Observations	624	624	624	624	624	624	624	624	624	624
R^2	0.009	0.476	0.4866	0.0262	0.4927	-0.0014	0.5546	0.5542	0.0413	0.5615

4.7 Explaining Out-of-Sample Sharpe Ratios with Random Matrix Theory

Our goal in this section is to understand: What properties of the data generating process might be responsible for the empirically observed performance of complex strategies. To answer this question, we appeal to our explicit, closed form theoretical characterizations of out-of-sample behavior of the complex strategy from Propositions 2, 3, and 12.

As we know from Figure 4, there are large differences in the dependence of Sharpe ratios on complexity across the four size groups. For example, for $P = 10000$, Sharpe ratios increase from roughly 1 for the mega group to 2.4 for the micro-caps.[35] According to Proposition 3, out-of-sample Sharpe ratio depends on the following quantities: (1) the size of the "true" β vector, $\|\beta\|^2 = b_*$; (2) eigenvalue distribution of Σ; (3) eigenvalue distribution of $\hat{\Psi}$. It is crucial for our analysis that all quantities in Proposition 3 are expressed in terms of the eigenvalue distribution of $\hat{\Psi}$ (which is observable) and not Ψ (which is not observable). At the same time, the large P assumption implies that $\hat{\Sigma}$ can be efficiently estimated by (39) because the effective sample size for estimating this $N \times N$ matrix is TP.

Not surprisingly, the quantity b_* defining the amount of predictability is an important driving factor of strategy performance, and the out-of-sample Sharpe ratio is monotone increasing in b_*. By (26), the achievable Sharpe ratio is bounded from above by its limit as $b_* \to \infty$, and we arrive at the following result.

Lemma 2 *By any $\theta > 0$ satisfying*

$$\theta \ < \ \frac{1}{(H(\Sigma)(1+\frac{\psi_{*,1}\hat{\nu}(z;cq;c)}{\nu^2(z;cq;c)}))^{1/2}}\,, \tag{52}$$

there exists a unique b_* such that $SR(z;cq,c)(b_*) = \theta$. It is given by

$$b_* \ = \ \frac{-B \ + \ \sqrt{B^2 - 4AC}}{2A}\,, \tag{53}$$

[35] Many complex strategies are only tradeable at very small capital and need to be adjusted to account for transaction costs. See Jensen et al. (2022).

where

$$\begin{aligned} A &= (SR^{-2} - H(\Sigma))(\operatorname{tr}\Sigma\nu)^2 - \psi_{*,1}(1)\operatorname{tr}(\Sigma)H(\Sigma)\operatorname{tr}(\Sigma)\hat{\nu}(z;cq;q) \\ B &= -\operatorname{tr}(\Sigma)\hat{\nu}(z;cq;q) + \psi_{*,1}(1)\operatorname{tr}(\Sigma)H(\Sigma)\operatorname{tr}(\Sigma)c\nu'(z;cq;q) \\ C &= \operatorname{tr}(\Sigma)c\nu'(z;cq;q)\,. \end{aligned} \tag{54}$$

We now use this lemma to perform a calibration exercise: Given the estimated eigenvalue distributions of $\hat{\Psi}$ and $\hat{\Sigma}$ and the observed realized Sharpe ratios of the complex strategy for each (z, P) combination, we can recover the unique b_* that can explain this Sharpe ratio using Lemma 2. Namely, for each size group *size*, and each (z, P), we observe realized returns $R^{size}_{t+1}(z, P)$, compute the full sample realized Sharpe ratio $SR^{size}(z, P)$, and then calibrate a $b_*(z, P)$ that exactly matches this Sharpe ratio. See Appendix A for details.

The calibrated quantities $b_*(z, P)$ are reported in Figure 10. Not surprisingly, all estimates are monotone increasing in P: Indeed, we know from Figure 4 that (due to the Virtue of Complexity), Sharpe ratios increase in P and, hence, so does b_*. The magnitudes of calibrated b_* for moderately complex models (with $P > 1000$) all lie between 0.05 and 0.2, implying an infeasible R^2 (see (14)) of

$$\frac{b_*\operatorname{tr}(\hat{\Sigma})}{b_*\operatorname{tr}(\hat{\Sigma}) + N} = \frac{b_*}{b_* + 12} \in [0.005, 0.015]\,, \tag{55}$$

where we have used that $\psi_{*,1}\operatorname{tr}(\Sigma) \approx \operatorname{tr}(\hat{\Sigma}) = N/12$ by Lemma 1. These extremely low levels of predictability are in the same ballpark as those obtained with modern machine learning models (see, e.g., Gu et al. (2020)).

Figure 10 clearly shows that the calibrated magnitudes of b_* are comparable across the four size groups, with the micro size group having slightly lower b_* levels. This is surprising. Indeed, conventional wisdom suggests that small stocks are much less liquid and hence exhibit much stronger predictability that simply cannot be exploited due to high transaction

costs.[36] Figure 10 suggests that this intuition is at the very least incomplete, and the superior performance might be attributed to the correlation structure of the non-linear signals.

To understand the sources of the large dispersion in Sharpe ratios across size groups, we investigate the three channels that define the theoretical values of Sharpe ratios in our model: (1) $b_*(size; z, P)$; (2) properties of the Σ matrix (captured by the two quantities, $Q1(size; P)) = (H(\hat{\Sigma}), \mathrm{tr}(\hat{\Sigma}))$; and (3) properties of the $\hat{\Psi}$ matrix, as captured by $Q2(size; z, P) = \hat{\nu}(z, P),\ \nu(z, P),\ \nu'(z, P)$. We focus on the two "most distant" size groups, micro and mega, and proceed as follows:

- By the definition of $b_*(size; z, P)$, we have

$$SR^{size}(z, P) \ = \ F(b_*(size; z, P);\ Q1(size; P);\ Q2(size; z, P))\,, \tag{56}$$

  where the function F is given by

$$F(b_*, Q1, Q2) \ = \ \frac{1}{\sqrt{H(\hat{\Sigma}) + \left(1 + b_* \,\mathrm{tr}(\hat{\Sigma}) H(\hat{\Sigma})\right) \frac{\mathrm{tr}(\hat{\Sigma})(\hat{\nu} - c\hat{\nu}')}{\left(b_* \,\mathrm{tr}(\hat{\Sigma})\nu\right)^2}}} \tag{57}$$

- Thus, replacing $b_*(mega)$ with $b_*(micro)$, we isolate the effect on differences in Sharpe ratios that are purely due to potential differences in b_*; replacing $Q1(mega; P)$ with $Q1(micro; P)$ isolates the effects originating from $\hat{\Sigma}$; and replacing $Q2(mega; z, P)$ by $Q2(micro; z, P)$ isolates the effects originating from $\hat{\Psi}$.

Figure 9 shows the results of this experiment, with $SR^{micro}(z, P)$ as well as

$$F(b_*(micro; z, P);\ Q1(mega; P);\ Q2(mega; z, P)),$$

[36] For example, many anomalies have been criticized for working only for small stocks, with value-weighted portfolios exhibiting much lower returns than equal weighted portfolios. See, for example, Asparouhova et al. (2010).

and

$$F(b_*(mega; z, P);\ Q1(micro; P);\ Q2(mega; z, P))\,,$$

and

$$F(b_*(mega; z, P);\ Q1(mega; P);\ Q2(micro; z, P))\,.$$

We also investigate separately the effects of two parts of $Q1:\ Q11 = H(\hat{\Sigma}),\ Q12 = \mathrm{tr}(\hat{\Sigma})$. As we can see, the impact of replacing $b_*(mega)$ with $b_*(micro)$ is minimal, with the level of the Sharpe ratio slightly dropping, consistent with the fact that the calibrated b_* is lower for the micro group. By contrast, the impact of the two eigenvalue distributions is large. We conclude that the differences in out-of-sample Sharpe ratios between mega- and micro-groups are not necessarily driven by differences in the degrees of predictability, b_*. Instead, these differences are consistent with a data generating process having the same amount of predictability, but different correlation structures of signals across characteristics and across stocks (see Figure 1 for a simulation illustrating a large impact of these correlation structures on the Sharpe ratios). In particular, complexity offer a novel perspective on the relationships between shocks when viewed from non-linear features' point of view. Our complex strategies (like most machine learning models in finance; see Gu et al. (2020)) attempt to leverage the idea that differences in stock characteristics translate into differences in expected returns. Our results imply that these differences ought to be studied not in the space of "raw characteristics," such as value and momentum, but in the extremely high-dimensional, complex space of non-linear features.

4.8 The Complexity Wedge

In section 3.2, we introduced the notion of "Limits to learning," defined as the difference between the realizable Sharpe ratio and the infeasible Sharpe ratio that would only be accessible to an econometrician with the knowledge of the true, ultra-high-dimensional vector β (see (31)). We now use our estimates of $b_*(z, P)$ from Section 4.7, substitute them into (31), and compute the infeasible and feasible Sharpe ratios. Figure 11 reports the outcome of this experiment. We observe several intriguing patterns. First, the limits to learning are very high.[37] Second, and somewhat surprisingly, the differences between the wedges across size groups are very small, despite the fact that the mega group has about ten times fewer stocks than the micro group. For moderate to high complexity levels ($P > 1000$), most of the calibrated wedges lie in the range between 0.8 and 0.9, implying that the infeasible Sharpe ratio is five to ten times higher than the feasible one. Third, we not that the complexity wedge is low, even for very low levels of complexity. This result might be surprising, as one would naturally expect that, for $P < NT$, the true β can be estimated with high precision by the law of large numbers. Our results imply that this intuition is incorrect. The reason is that, as we explain above, the "effective" model complexity is defined by c/b_*. Even if c is close to zero (e.g., for $P = 100$, $NT \approx 2000$ for the mega size group, implying a $c \approx 0.05$), a small degree of predictability ($b_* \approx 0.025$ in Figure 10) makes the *effective complexity*, $c/b_* \approx 2$, large.

The large size of the complexity wedge has important economic implications. The very existence of sustainable predictive power of characteristics for stock returns in a non-complex framework is a puzzle. Classical "non-complex" economics must label every high-performing strategy as either a statistical anomaly, an economic anomaly, or a factor. By contrast, in a "complex" framework, the existence of vast "limits to learning" suggests a way to reconcile

[37] The results for very low levels of z should be taken with caution because covariance matrices are close to being degenerate, leading to instabilities.

highly performing strategies and equilibrium prices. The formal proof of this intuition is, however, outside the scope of this paper.

4.9 Which Covariates Matter?

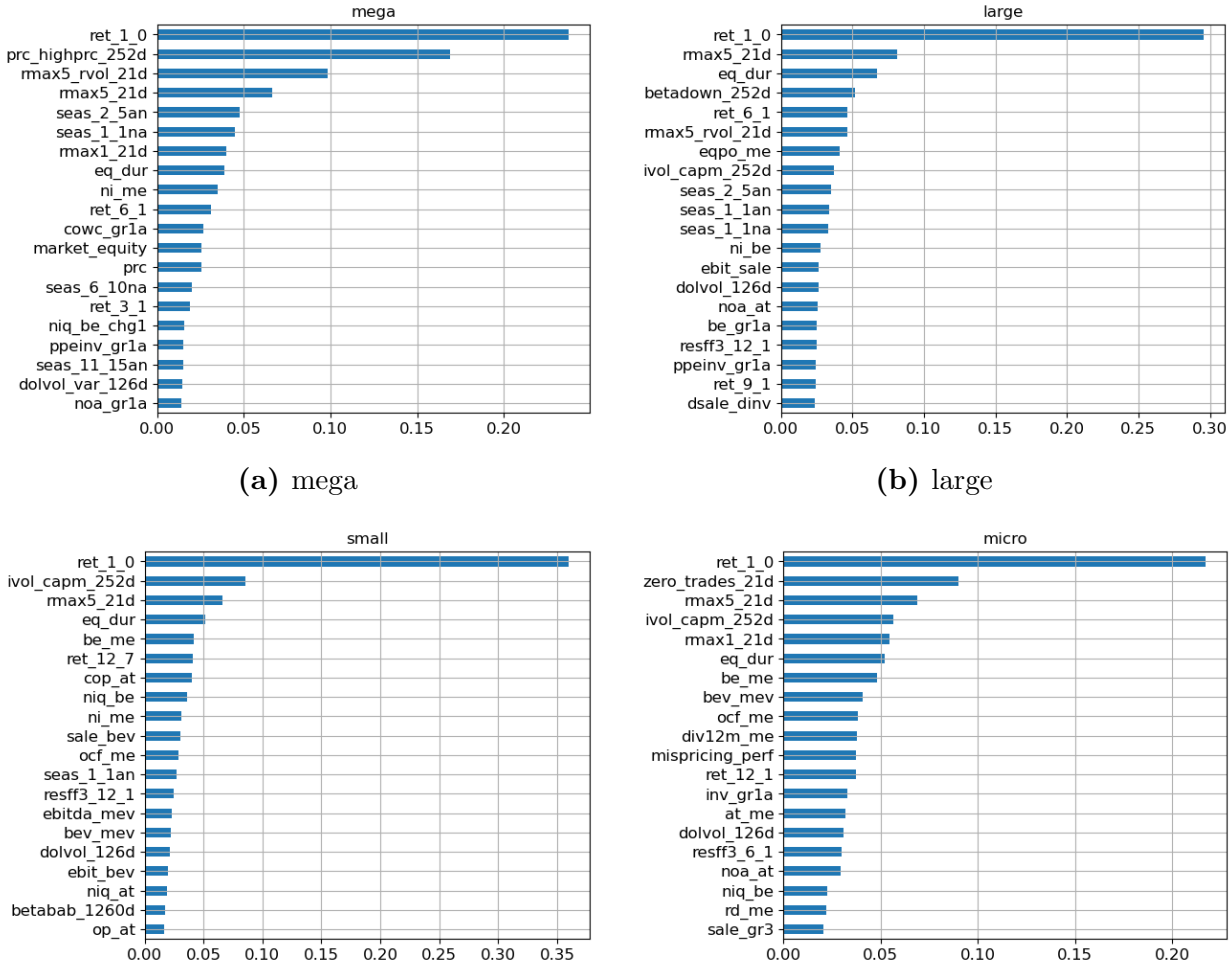
We conclude this section with an investigation of the relative importance of various stock characteristics for the performance of our complex models. We focus on $P = 10^4$ and the value of z maximizing the Sharpe ratio.

For each fixed k, $k = 1, \cdots, 153$, and each model $m \in \{Complex, Linear, EW\}$, we compute the Sharpe ratio obtained after setting all values of a given predictor $\tilde{S}_{i,k,t}$, $i = 1, \cdots, N$ to zero, re-compute the random features and their cross-sectional ranks (for the complex strategy) and investigate the impact on model performance.[38] We measure feature importance as the percentage drop in the Sharpe ratio.

Figure 7 shows the top 20 most important features (i.e., the ones causing the largest percentage drop in the Sharpe ratio) for each size group. As in (Gu et al., 2020), we normalize the feature importance into each model so that they sum to one. Figure 8 presents the importance of all covariates, sorting them from most important to least important based on their average importance across models.

Consistently with known findings ((Gu et al., 2020), (Chen et al., 2019) (Avramov et al., 2021), (Jensen et al., 2022)), we find that one-month momentum is the most important feature or all size groups. The importance of other features is highly heterogeneous across groups.

[38]Note that 0 is the natural value for the raw signals because they are also ranked based on the transformation (49) and missing values are also replaced with zeros.

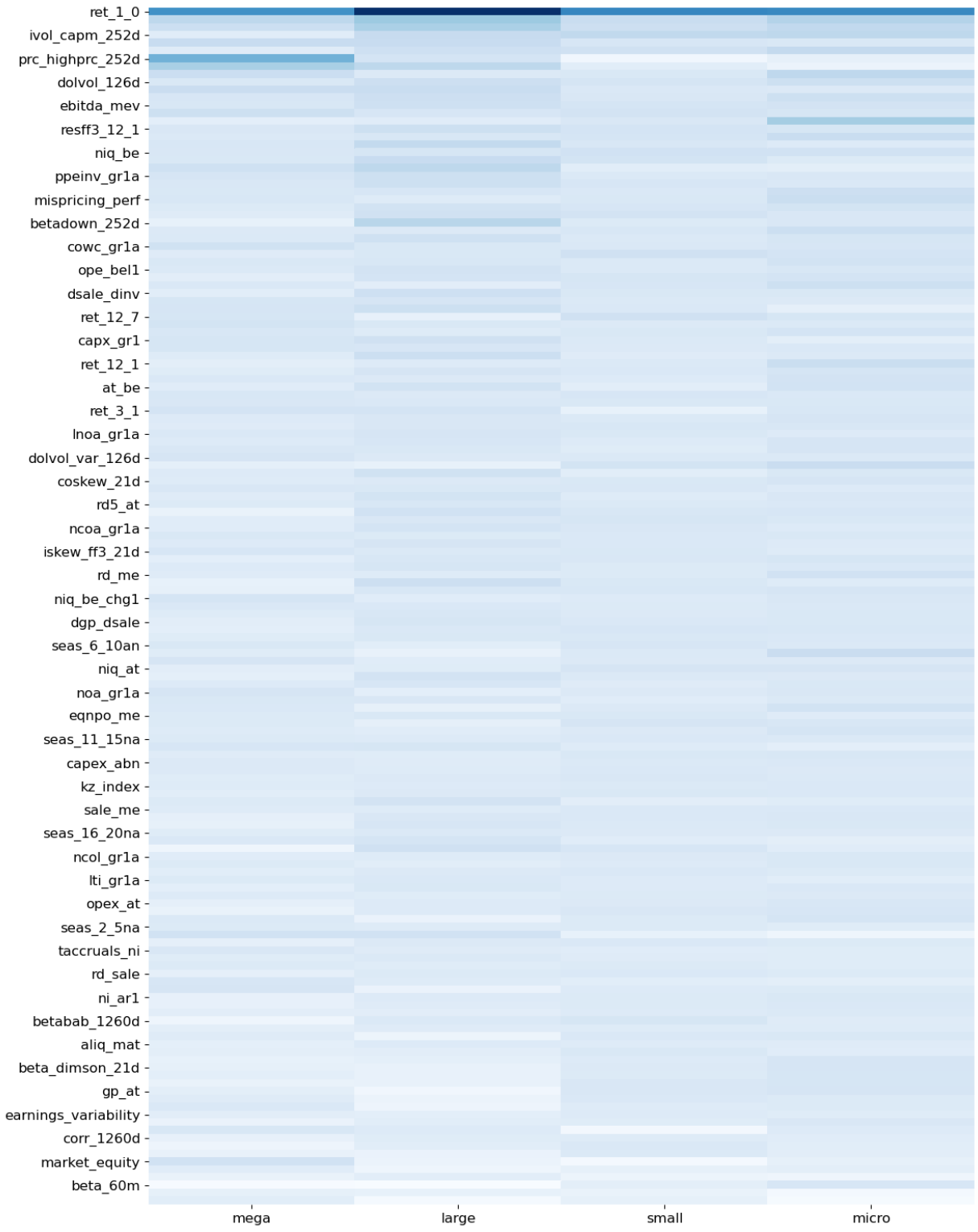


(a) mega **(b)** large

(c) small **(d)** micro

Figure 7: In the figures above, we show the variable importance for the top 20 most influential variables in each model. We compute those variable on the complex model for $P = 10^4$.

Using pre-trained β, we re-estimate the model's signals by setting one variable v in turn to 0. We then compute the out-of-sample Sharpe ratio of the model without the variable SR_v. We defined the variable importance as $\frac{SR}{SR_v} - 1$. We normalize the variable into each model to sum to one.



(a) mega

Figure 8: In the figure above, we show the ranking of 153 variables in terms of overall model contribution. We order the variable based on the sum of their ranks over all models, with the most influential variable on the top and the least influential on the bottom. The columns correspond to market size, and the color gradients within each column indicate the most influential (dark blue) to the least influential (white) variables.

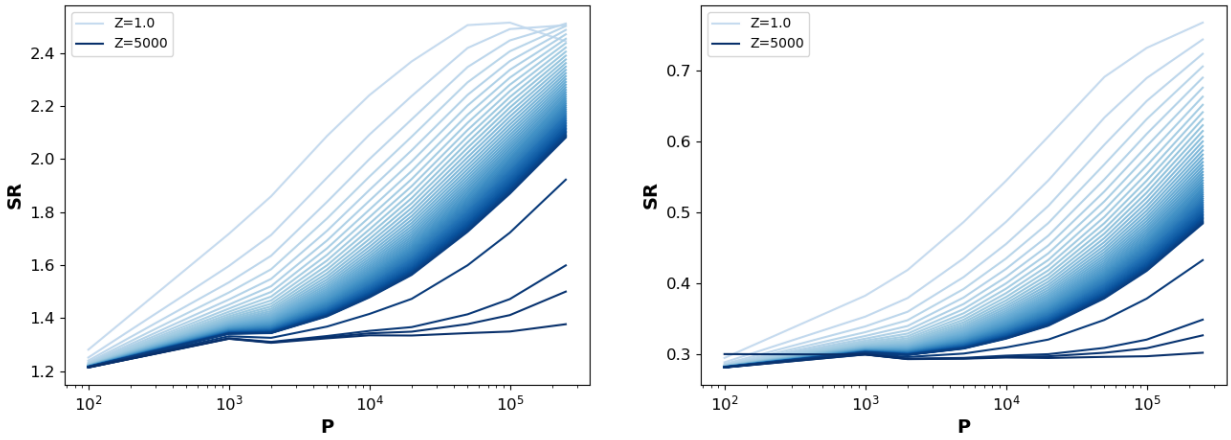
5 Conclusion

The statistical theory of large cross-sections of stocks relies heavily on standard limit theorems that assume *zero complexity,* so that the number of estimated coefficients is negligible relative to the panel size. We provide empirical evidence that the dependence of stock returns on characteristics is characterized by a complexity that is significantly different from zero. Complexity in the cross-section implies that standard limit theorems[39] do not apply and the asymptotic out-of-sample R^2 and Sharpe Ratios are significantly different from their in-sample counter-parts. Furthermore, optimal performance of complex models is achieved with a very large amount of shrinkage that is proportional to model complexity. We use random matrix theory to derive explicit expressions for the out-of-sample performance of ultra-high-dimensional predictive models for large cross-sections of stock returns. We test these expressions empirically and find strong support for our theoretical predictions. In particular, US stocks feature the *virtue of complexity*: Out-of-sample performance increases in model parametrization even for P comparable with NT.

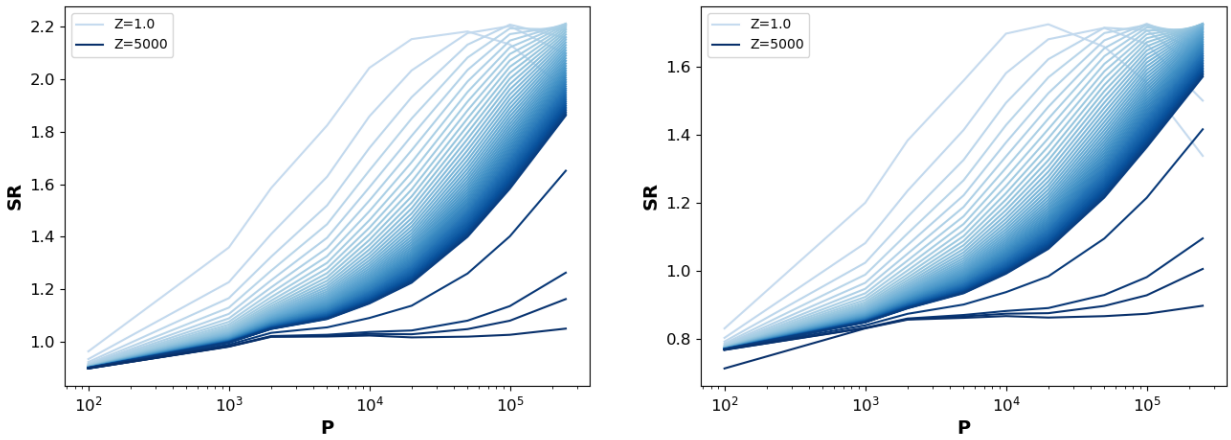
In this paper, we investigate one single machine learning algorithm: The random Fourier Features of Rahimi and Recht (2007) that is perfectly suited for our theoretical analysis of complex models based on ridge regression. This begs the question: Is there an optimal machine learning algorithm for estimating non-linearities in stock returns? In general, the answer is no. In the high complexity regime, different models all deliver *different* approximations of the true DGP, with none strictly dominating the others. Mei and Montanari (2019) show that high model complexity poses an insurmountable obstacle for any random feature regression—it is impossible to learn the "true" dependency when the model is complex. In this case, different learning algorithms recover different aspects (projections) of the truth on different subspaces of functions. As a result, we would expect that predictions derived from diverse machine learning models contain complementary information.

[39] See, e.g., Gagliardini et al. (2016) Chen et al. (2021).

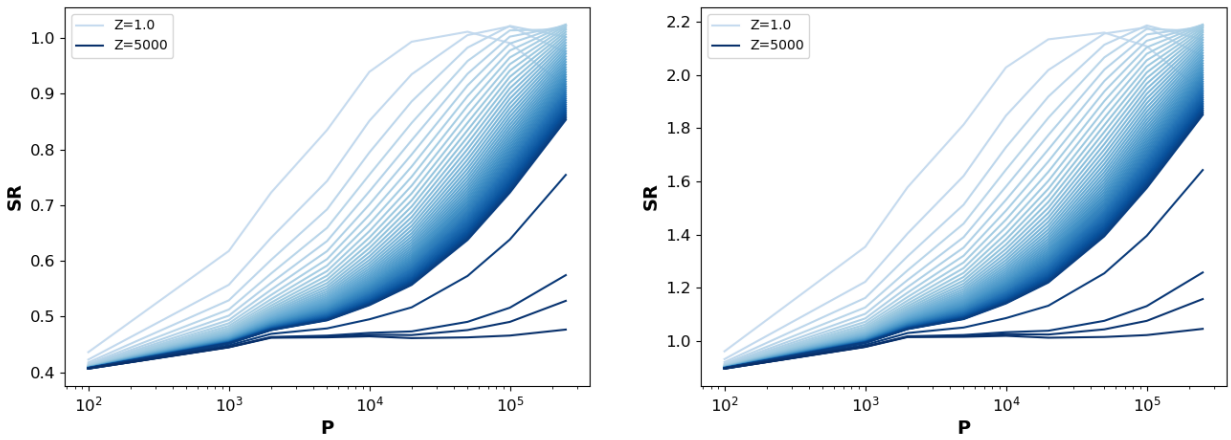
How should one optimally extract and combine this information? How does complexity move over time? How does it interact with liquidity and what part of complex models are implementable with a large notional capital without being eroded by price impact? How is complexity determined in equilibrium and what is its role for shaping the factor structure of returns? How do we measure feature importance and what is the role of different machine learning algorithms and their ability to discover the true non-linear structure of returns? We leave all these interesting questions for future research.



(a) $SR^{\text{micro}}\,(z;P)$ **(b)** $F\,(b_*(\text{ micro; });Q1(\text{ mega });Q2(\text{ mega }))$

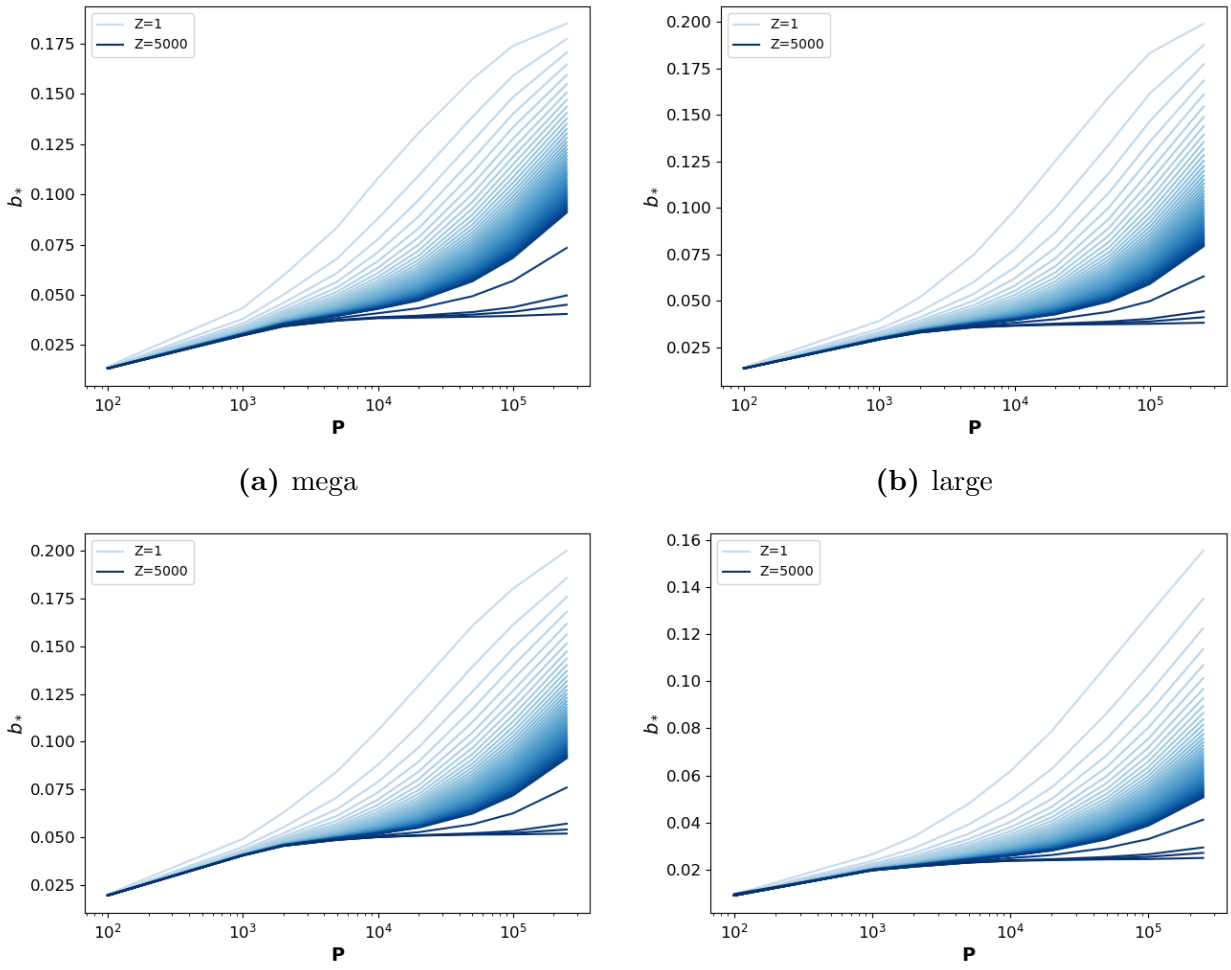


(c) $F\,(b_*(\text{ mega });Q1(\text{ micro });Q2(\text{ mega }))$ **(d)** $F\,(b_*(\text{ mega });Q1(\text{ mega });Q2(\text{ micro; }))$



(e) $F\,(Q11(\text{ micro });Q12(\text{ mega })))$ **(f)** $F\,(Q11(\text{ mega });Q12(\text{ micro }))$

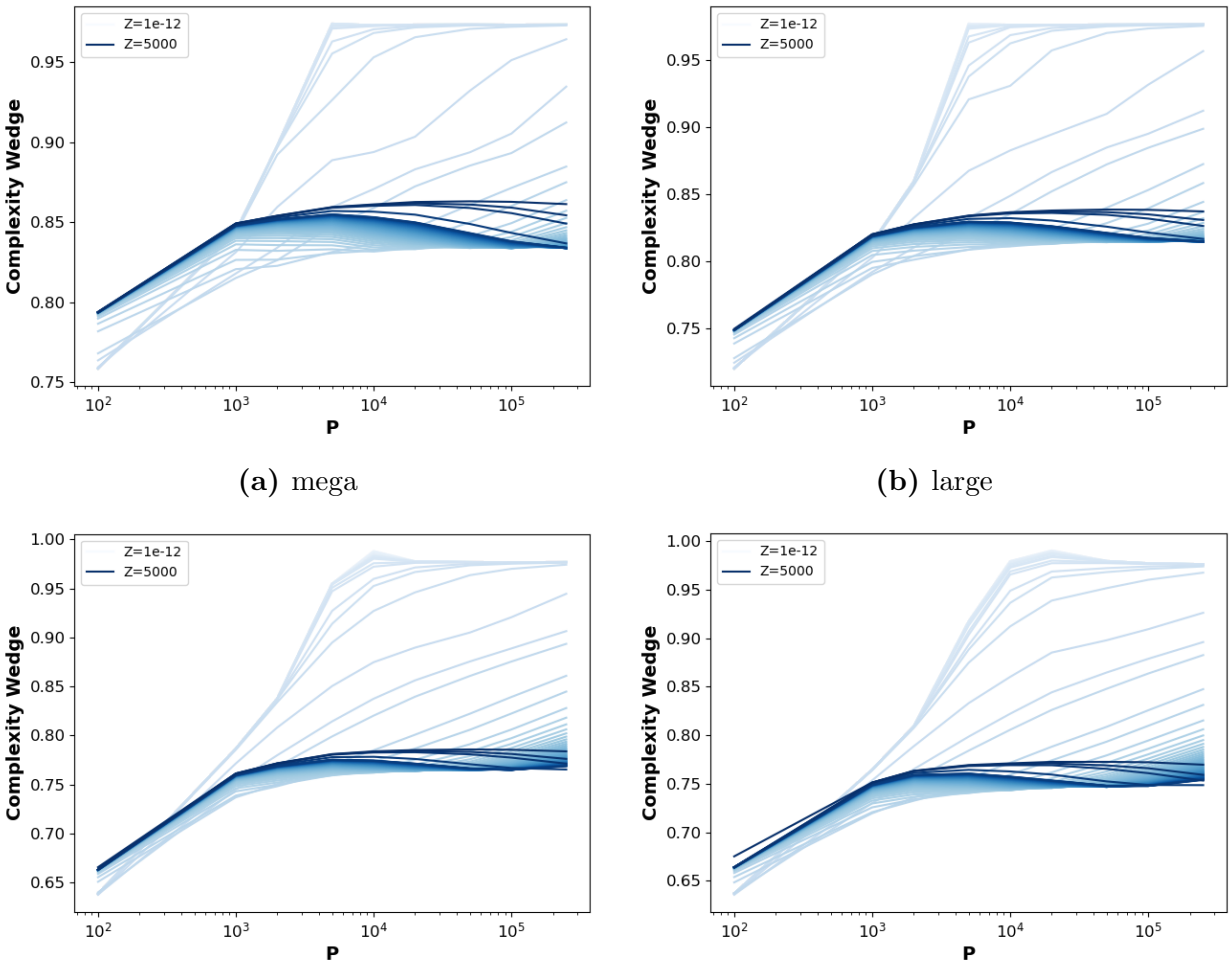
Figure 9: This figure investigates sources of differences is Sharpe ratios between the mega- and micro-groups, following the calibration experiment based on formulas (56) and (57).



(a) mega **(b)** large

(c) small **(d)** micro

Figure 10: In the figure above, we show the calibrated values of $b_*(z, P)$ based on Lemma 2.



(a) mega

(b) large

(c) small

(d) micro

Figure 11: The complexity wedge, $1 - SR(z; cq; c)/SR(0; q)$, as function of z and P (see (31)) with calibrated values of $b_*(z, P)$ based on Lemma 2. The values for low $z < 1$ are very high due to instabilities produces by close-to-degenerate signal covariance matrices. .

References

Abhyankar, Abhay, Devraj Basu, and Alexander Stremme, "The optimal use of return predictability: An empirical study," *Journal of Financial and Quantitative Analysis*, 2012, *47* (5), 973–1001.

Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song, "A convergence theory for deep learning via over-parameterization," in "International Conference on Machine Learning" PMLR 2019, pp. 242–252.

Asparouhova, Elena, Hendrik Bessembinder, and Ivalina Kalcheva, "Liquidity biases in asset pricing tests," *Journal of Financial Economics*, 2010, *96* (2), 215–237.

Avramov, Doron, Si Cheng, and Lior Metzker, "Machine learning versus economic restrictions: Evidence from stock return predictability," *Available at SSRN 3450322*, 2021.

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 biasvariance 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.

Bianchi, Daniele, Matthias Büchner, and Andrea Tamoni, "Bond risk premiums with machine learning," *The Review of Financial Studies*, 2021, *34* (2), 1046–1089.

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

Burkholder, Donald L, "Martingale transforms," *The Annals of Mathematical Statistics*, 1966, *37* (6), 1494–1504.

Carhart, Mark M, "On persistence in mutual fund performance," *The Journal of finance*, 1997, *52* (1), 57–82.

Chen, Luyang, Markus Pelger, and Jason Zhu, "Deep learning in asset pricing," *arXiv preprint arXiv:1904.00745*, 2019.

Chen, Qihui, Nikolai Roussanov, and Xiaoliang Wang, "Semiparametric Conditional Factor Models: Estimation and Inference," *arXiv preprint arXiv:2112.07121*, 2021.

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.

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

Du, Simon, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai, "Gradient descent finds global minima of deep neural networks," in "International Conference on Machine Learning" PMLR 2019, pp. 1675–1685.

Du, Simon S, Xiyu Zhai, Barnabas Poczos, and Aarti Singh, "Gradient descent provably optimizes over-parameterized neural networks," *arXiv preprint arXiv:1810.02054*, 2018.

Fama, Eugene F and Kenneth R French, "A five-factor asset pricing model," *Journal of financial economics*, 2015, *116* (1), 1–22.

Fan, Jianqing, Jianhua Guo, and Shurong Zheng, "Estimating number of factors by adjusted eigenvalues thresholding," *Journal of the American Statistical Association*, 2022, *117* (538), 852–861.

___ **, Yingying Fan, and Jinchi Lv**, "High dimensional covariance matrix estimation using a factor model," *Journal of Econometrics*, 2008, *147* (1), 186–197.

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

Feng, Guanhao, Jingyu He, and Nicholas G Polson, "Deep learning for predicting asset returns," *arXiv preprint arXiv:1804.09314*, 2018.

Ferson, Wayne E and Andrew F Siegel, "The efficient use of conditioning information in portfolios," *The Journal of Finance*, 2001, *56* (3), 967–982.

Filipović, Damir and Amir Khalilzadeh, "Machine Learning for Predicting Stock Return Volatility," *Swiss Finance Institute Research Paper*, 2021, (21-95).

___ **, Markus Pelger, and Ye Ye**, "Stripping the Discount Curve-a Robust Machine Learning Approach," *Swiss Finance Institute Research Paper*, 2022, (22-24).

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

Gagliardini, Patrick and Hao Ma, "Extracting statistical factors when betas are time-varying," *Swiss Finance Institute Research Paper*, 2019, (19-65).

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

Ghorbani, Behrooz, Song Mei, Theodor Misiakiewicz, and Andrea Montanari, "When do neural networks outperform kernel methods?," *Advances in Neural Information Processing Systems*, 2020, *33*, 14820–14830.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, "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 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.

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

Jacot, Arthur, Franck Gabriel, and Clément Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," *arXiv preprint arXiv:1806.07572*, 2018.

Jegadeesh, Narasimhan and Sheridan Titman, "Returns to buying winners and selling losers: Implications for stock market efficiency," *The Journal of finance*, 1993, *48* (1), 65–91.

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.

Kelly, Bryan, Semyon Malamud, and Kangying Zhou, "The Virtue of Complexity in Machine Learning Portfolios," *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, "Shrinking the cross-section," *Journal of Financial Economics*, 2020, *135* (2), 271–292.

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

Leippold, Markus, Qian Wang, and Wenyu Zhou, "Machine learning in the Chinese stock market," *Journal of Financial Economics*, 2022, *145* (2), 64–82.

Lewellen, Jonathan, "The cross section of expected stock returns," *Forthcoming in Critical Finance Review, Tuck School of Business Working Paper*, 2014, (2511246).

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

Ma, Hao, "Conditional Latent Factor Models Via Econometrics-Based Neural Networks," *Available at SSRN*, 2021.

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.

Mei, Song and Andrea Montanari, "The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve," *Communications on Pure and Applied Mathematics*, 2019.

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.

Richards, Dominic, Jaouad Mourtada, and Lorenzo Rosasco, "Asymptotics of ridge (less) regression under general source condition," in "International Conference on Artificial Intelligence and Statistics" PMLR 2021, pp. 3889–3897.

Rossi, Alberto G, "Predicting stock market returns with machine learning," *Georgetown University*, 2018.

Silverstein, Jack W and ZD Bai, "On the empirical distribution of eigenvalues of a class of large dimensional random matrices," *Journal of Multivariate analysis*, 1995, *54* (2), 175–192.

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.

Yaskov, Pavel, "A short proof of the Marchenko–Pastur theorem," *Comptes Rendus Mathematique*, 2016, *354* (3), 319–322.

A Details for Calibrating b_*

- For size group *size*, and each (z, P), we observe realized returns $R^{size}_{t+1}(z,P)$ for the complex strategy, as well as the time series $\hat{\nu}_t(z,P)$, $\nu_t(z,P)$, $\nu'_t(z,P)$, $\text{tr}(\hat{\Sigma}_t)$, $H(\hat{\Sigma}_t)$ and $c_t = P/(TN_t)$.[40] We then compute the full sample realized Sharpe ratio $SR^{size}(z,P)$.

- We compute full sample means of all the random matrix theory quantities,

$$\hat{\nu}_t(z,P),\ \nu_t(z,P),\quad \text{tr}(\hat{\Sigma}_t),\ H(\hat{\Sigma}_t).$$

  The only special care needs to be taken of $c\nu'(z;cq;q)$, whose mean needs to be computed as

$$\frac{1}{T}\sum_t c_t \nu'_t(z;cq;q)\,,$$

  accounting for the fact that c_t moves over time due to the changing number of stocks.

- we substitute these full-sample means into formulas (54), with SR replaced by $SR^{size}(z,P)$, and recover $b_*(z,P)$ using formula (53).

B Preliminary Calculations

A key aspect of our paper, and one rooted in Assumptions 2 and 4, is that realized out-of-sample returns are independent of the specific realization of β. This is due to a law of large numbers in the $P \to \infty$ limit, and is guaranteed by the following lemma.[41]

Lemma 3 *As $P \to \infty$ we have*

$$\beta' A_P \beta - P^{-1} b_* \,\text{tr}(A_P) \ \to 0$$

in probability for any bounded sequence of matrices A_P. In particular, $\beta'\Psi\beta \ \to\ b_*\psi_{*,1}$.

Lemma 4 *Under Assumption 2, define*

$$\tilde{\beta} \ = \ \Psi^{1/2}\beta\,. \tag{58}$$

For simplicity, we also assume that $\kappa = E[X^4_{i,k} - 1]$ is independent of i,k.[42] Then,

$$E[S'_t S_t] \ = \ \text{tr}(\Sigma)\Psi \text{ and } E[F_{t+1}] \ = \ \text{tr}(\Sigma)\Psi\beta\,,$$

[40] As we explain above, we re-compute the $\hat{\beta}$ vector estimate every three months. For this reason, the random matrix quantities are also estimated at quarterly frequency.

[41] It is possible to use the results in (Hastie et al., 2019) to extend our analysis to generic β distributions. We leave this important direction for future research.

[42] The general case is slightly more cumbersome, but all kurtosis terms vanish in the random matrix limit.

where

$$F_{t+1} \;=\; S_t' R_{t+1}\,. \tag{59}$$

Then,

$$E[F_{t+1}F_{t+1}'] \;=\; ((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))\Psi\beta\beta'\Psi + \operatorname{tr}(\Sigma^2)(\kappa-2)\Psi^{1/2}\operatorname{diag}(\tilde\beta^2)\Psi^{1/2}+\Psi\Big(\operatorname{tr}(\Sigma)+\beta'\Psi\beta\operatorname{tr}(\Sigma^2)\Big)$$

and

$$\beta' E[F_{t+1}F_{t+1}']\beta \;=\; ((\operatorname{tr}\Sigma)^2+2\operatorname{tr}(\Sigma^2))(\beta'\Psi\beta)^2+\operatorname{tr}(\Sigma^2)(\kappa-2)\|(\Psi^{1/2}\beta)^2\|^2+\beta'\Psi\beta\operatorname{tr}(\Sigma) \tag{60}$$

Proof of Lemma 4. We have

$$E[F_{t+1}F_{t+1}'] \;=\; E[S_t'(S_t\beta+\varepsilon)(S_t\beta+\varepsilon)'S_t] \;=\; E[S_t'S_t\beta\beta'S_t'S_t] \;+\; E[S_t'S_t]\,,$$

and

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

where we have used the following lemma, whose proof follows by direct calculation.

Lemma 5 *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,*

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

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

$$\begin{aligned} E[S_t'S_t\beta\beta'S_t'S_t] \;&=\; E[\Psi^{1/2}X_t'\Sigma X_t\Psi^{1/2}\beta\beta'\Psi^{1/2}X_t'\Sigma X_t\Psi^{1/2}] \;= E[\Psi^{1/2}X_t'\Sigma X_t\tilde\beta\tilde\beta'X_t'\Sigma X_t\Psi^{1/2}] \\ &=\; \Psi^{1/2}E[\tilde X_t'D\tilde X_t\tilde\beta\tilde\beta'\tilde X_t'D\tilde X_t]\Psi^{1/2} \end{aligned} \tag{61}$$

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

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

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

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

Thus,

$$\begin{aligned} & E[\tilde{X}_t' D \tilde{X}_t \tilde{\beta} \tilde{\beta}' \tilde{X}_t' D \tilde{X}_t]_{k_1,k_2} \\ &= \; ((\mathrm{tr}\,\Sigma)^2 - \mathrm{tr}(\Sigma^2)) \tilde{\beta}_{k_1} \tilde{\beta}_{k_2} \; + 2\,\mathrm{tr}(\Sigma^2) \tilde{\beta}_{k_1} \tilde{\beta}_{k_2} (1 - \delta_{k_1,k_2}) + \; \mathrm{tr}(\Sigma^2)(\kappa \tilde{\beta}_{k_1}^2 + \|\tilde{\beta}\|^2) \delta_{k_1,k_2} \\ &= \; ((\mathrm{tr}\,\Sigma)^2 + \mathrm{tr}(\Sigma^2)) \tilde{\beta}_{k_1} \tilde{\beta}_{k_2} + \mathrm{tr}(\Sigma^2)((\kappa - 2) \tilde{\beta}_{k_1}^2 + \|\tilde{\beta}\|^2) \delta_{k_1,k_2} \end{aligned} \tag{62}$$

Thus, by formula (61), we get

$$E[S_t' S_t \beta \beta' S_t' S_t] \;=\; ((\mathrm{tr}\,\Sigma)^2 + \mathrm{tr}(\Sigma^2)) \Psi \beta \beta' \Psi + \mathrm{tr}(\Sigma^2)((\kappa - 2) \Psi^{1/2}\, \mathrm{diag}(\tilde{\beta}_{k_1}^2) \Psi^{1/2} + \|\tilde{\beta}\|^2 \Psi) \tag{63}$$

and the claim follows because $\|\tilde{\beta}\|^2 = \beta' \Psi \beta$. □

The following corollary follows directly from spectral decomposition of symmetric matrices.

Corollary 7 *For any symmetric matrix A, we have*

$$\begin{aligned} E[S_t' S_t A S_t' S_t] \;&=\; ((\mathrm{tr}\,\Sigma)^2 + \mathrm{tr}(\Sigma^2)) \Psi A \Psi + \mathrm{tr}(\Sigma^2)\,\mathrm{tr}(\Psi A) \Psi \\ &+ \; \mathrm{tr}(\Sigma^2)\,(\kappa - 2) \Psi^{1/2}\, \mathrm{diag}(\Psi^{1/2} A \Psi^{1/2}) \Psi^{1/2} \end{aligned} \tag{64}$$

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

Proof. Writing

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

we can apply Lemma 4. □

C Preliminary Calculations for the Mis-Specified Model

Our objective is to maximize the unconditional Sharpe ratio

$$SR_t \;=\; \frac{E[R_{t+1}^\pi]}{\mathrm{Var}[R_{t+1}^\pi])^{1/2}}\,. \tag{65}$$

If we have access to full information about the DGP (1), the strategy maximizing (65) is given by[43],[44]

$$\pi_t(\beta) = S_t\beta \;=\; \pi_t^{\text{Cond. MV}} = (\text{Var}_t[R_{t+1}^2])^{-1}E_t[R_{t+1}]\,. \tag{67}$$

When the model is mis-specified, we only have access to a subset of signals, $S_t^{(1)}$.

We focus on linear strategies: $\pi_t \;=\; S_t^{(1)}\hat{\beta}$, where $\hat{\beta}$ is only allowed to depend on $\beta^{(1)}$. Then,

$$E[\pi_t' R_{t+1}] \;=\; E[\hat{\beta}'(S_t^{(1)})'((S_t^{(1)}\beta^{(1)} + S_t^{(2)}\beta^{(2)})] \;\to\; \hat{\beta}'\,\text{tr}(\Sigma)(\Psi_{1,1}\beta^{(1)}) \tag{68}$$

whereas, assuming for simplicity that $\beta^{(1)}$ and $\beta^{(2)}$ are independent, we get

$$\begin{aligned}
&E[\pi_t' R_{t+1}R_{t+1}'\pi_t] \\
&= E[\hat{\beta}'(S_t^{(1)})'(S_t^{(1)}\beta^{(1)} + S_t^{(2)}\beta^{(2)} + \varepsilon_{t+1})(S_t^{(1)}\beta^{(1)} + S_t^{(2)}\beta^{(2)} + \varepsilon_{t+1})'S_t^{(1)}\hat{\beta}] \\
&= \hat{\beta}'E[(S_t^{(1)})'(S_t^{(1)}\beta^{(1)}(\beta^{(1)})'(S_t^{(1)})' + S_t^{(2)}\beta^{(2)}(\beta^{(2)})'(S_t^{(2)})' + I)]S_t^{(1)}\hat{\beta}] \\
&= \hat{\beta}'E[(S_t^{(1)})'(S_t^{(1)}\beta^{(1)}(\beta^{(1)})'(S_t^{(1)})' + \Sigma P^{-1}\,\text{tr}(\Psi_{2,2}) + I)]S_t^{(1)}\hat{\beta}] \\
&= \hat{\beta}'E[((\text{tr}\,\Sigma)^2 + \text{tr}(\Sigma^2))\Psi_{1,1}\beta^{(1)}(\beta^{(1)})'\Psi_{1,1} + \text{tr}(\Sigma^2)\,\text{tr}(\Psi_{1,1}\beta^{(1)}(\beta^{(1)})')\Psi_{1,1} \\
&+ \text{tr}(\Sigma^2)\,(\kappa-2)\Psi_{1,1}^{1/2}\,\text{diag}(\Psi_{1,1}^{1/2}\beta^{(1)}(\beta^{(1)})'\Psi_{1,1}^{1/2})\Psi_{1,1}^{1/2} \;+\; P^{-1}b_*(S_t^{(1)})'S_t^{(2)}(S_t^{(2)})'S_t^{(1)} + \Psi_{1,1}]\hat{\beta}
\end{aligned} \tag{69}$$

We also assume that the second component of β corresponding to $S_t^{(2)}$ is unknown, in which case Assumption 4 implies that

$$E[R_{t+1}|S_t^{(1)}] \;=\; S_t^{(1)}\beta^{(1)}\,. \tag{70}$$

Assuming for simplicity that the trace (i.e., the total variance of $S_t^{(2)}$) of the conditional covariance matrix of missing signals, $\text{tr}(S_t^{(2)}(S_t^{(2)})')$ does not depend on the realization of $S_t^{(1)}$ (this is the case, e.g., when S_t is jointly Gaussian), we get

$$\begin{aligned}
&E_t[R_{t+1}R_{t+1}'|S_t^{(1)}] \\
&= E_t[(S_t^{(1)}\beta^{(1)} + S_t^{(2)}\beta^{(2)} + \varepsilon_{t+1})(S_t^{(1)}\beta^{(1)} + S_t^{(2)}\beta^{(2)} + \varepsilon_{t+1})'|S_t^{(1)}] \\
&= E[R_{t+1}|S_t^{(1)}]E[R_{t+1}|S_t^{(1)}]' \;+\; I \;+\; b_*P^{-1}\,\text{tr}(\Psi_{2,2})\Sigma\,,
\end{aligned} \tag{71}$$

[43] Note that

$$\pi_t' R_{t+1} \;=\; (S_t\beta)'\Sigma_{\varepsilon,t}^{-1/2}\tilde{R}_{t+1} \;=\; (\Sigma_{\varepsilon,t}^{-1/2}\tilde{S}_t\beta)'\Sigma_{\varepsilon,t}^{-1/2}\tilde{R}_{t+1} \;=\; \beta'\tilde{S}_t\Sigma_{\varepsilon,t}^{-1}\tilde{R}_{t+1}\,. \tag{66}$$

is the return on the conditionally mean-variance efficient portfolio in the original return space, $\tilde{R}_{t+1}$.

[44] While strategy π_t is conditionally mean-variance efficient, it is not the optimizer of the unconditional objective $\frac{E[R_{t+1}^\pi]}{\text{Var}[R_{t+1}^\pi]}$, which takes the form $\pi_t^{\text{Uncond. MV}} = S_t\beta/(N + \beta'S_t'S_t\beta)$. See (Hansen and Richard, 1987; Ferson and Siegel, 2001; Abhyankar et al., 2012). In the limit as $P \to \infty$, the normalization factor $N + \beta'S_t'S_t\beta$ converges to a constant. By Lemma 3 in the Appendix, $N + \beta'S_t'S_t\beta \to_{P\to\infty} N + b_*\psi_{*,1}\,\text{tr}(\Sigma)$.

implying that the conditionally mean-variance efficient portfolio is given by

$$\pi_t^{(1)} \;=\; \left(I \;+\; b_* P^{-1}\operatorname{tr}(\Psi_{2,2})\Sigma\right)^{-1} S_t^{(1)}\beta^{(1)}\,. \tag{72}$$

Proposition 8 states the behavior of timing strategy $\pi_t = S_t\beta^{(1)}$ when $T \to \infty$ and $P/T \to 0$ (i.e., when the predictive parameter β is known).

Proposition 8 (High Complexity Limit of the Infeasible Strategy) *Let*

$$H(\Sigma) \;=\; \frac{\operatorname{tr}(\Sigma^2)}{(\operatorname{tr}(\Sigma))^2} - 1 \;\in\; [N^{-1}-1, 0]\,. \tag{73}$$

In the limit as $P \to \infty$, the conditional mean and variance to the infeasible market timing strategy $\pi_t = S_t\beta^{(1)}$ converge to

$$E_t[\pi_t' R_{t+1}] \to_{P\to\infty} \operatorname{tr}(\Sigma) q b_* \psi_{*,1}(q) \tag{74}$$

$$\operatorname{Var}_t\left[(\pi_t' R_{t+1})^2\right] \to_{P\to\infty} \;(q\psi_{*,1}(q) b_* \operatorname{tr}(\Sigma))^2 \left(H(\Sigma) \;+ (q\psi_{*,1}(q) b_* \operatorname{tr}(\Sigma))^{-1}(1 + H(\Sigma) b_* \psi_{*,1}(1) \right) \tag{75}$$

For $q = 1$, the infeasible market timing conditional Sharpe ratio is given by

$$SR_t \to \frac{1}{\sqrt{2H(\Sigma) + (b_*\psi_{*,1}\operatorname{tr}(\Sigma))^{-1}}} < \left(\frac{N}{2}\right)^{1/2}. \tag{76}$$

At the same time,

$$R^2 \to \frac{\operatorname{tr}(\Sigma)\beta'\Psi\beta}{\operatorname{tr}(\Sigma)\beta'\Psi\beta + N} \to \frac{\operatorname{tr}(\Sigma) b_* \psi_{*,1}}{\operatorname{tr}(\Sigma) b_* \psi_{*,1} + N}. \tag{77}$$

Proof of Proposition 8. By (68),

$$E[\pi_t' R_{t+1}] \;=\; \beta^{(1)}\operatorname{tr}(\Sigma)\Psi_{1,1}\beta^{(1)} \;=\; P^{-1}\operatorname{tr}(\Psi_{1,1})\operatorname{tr}(\Sigma) \;=\; q\psi_{*,1}(q) b_* \operatorname{tr}(\Sigma)\,,$$

whereas

$$\begin{aligned}
&E[\pi_t' R_{t+1} R_{t+1}' \pi_t] \\
&= \;(\beta^{(1)})' E[((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2))\Psi_{1,1}\beta^{(1)}(\beta^{(1)})'\Psi_{1,1} + \operatorname{tr}(\Sigma^2)\operatorname{tr}(\Psi_{1,1}\beta^{(1)}(\beta^{(1)})')\Psi_{1,1} \\
&+ \;\operatorname{tr}(\Sigma^2)\,(\kappa-2)\Psi_{1,1}^{1/2}\operatorname{diag}(\Psi_{1,1}^{1/2}\beta^{(1)}(\beta^{(1)})'\Psi^{1/2})\Psi^{1/2} \;+\; P^{-1} b_* (S_t^{(1)})' S_t^{(2)} (S_t^{(2)})' S_t^{(1)} + \Psi_{1,1}\operatorname{tr}(\Sigma)](\beta^{(1)}) \\
&\approx \;(q\psi_{*,1}(q) b_* \operatorname{tr}(\Sigma))^2 (1 + 2H(\Sigma)) \;+ q\psi_{*,1}(q) b_* \operatorname{tr}(\Sigma) + \;Term,
\end{aligned} \tag{78}$$

where, by Lemma 18, we have

$$\begin{aligned}
Term &\approx P^{-2}b_*^2 \operatorname{tr} E[S'_{t,2}S_{t,1}S'_{t,1}S_{t,2}] \\
&\approx P^{-2}b_*^2 \operatorname{tr}\Big[((\operatorname{tr}\Sigma)^2 - \operatorname{tr}(\Sigma^2))\Psi_{2,1}\Psi_{1,2} \;+\; 2(\operatorname{tr}\Sigma^2)\Psi_{2,1}\Psi_{1,2} \\
&+ \operatorname{tr}\Sigma^2 \operatorname{tr}(\Psi_{1,1})\Psi_{2,2}\Big] \\
&\approx P^{-2}b_*^2((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2)) \operatorname{tr}(\Psi_{2,1}\Psi_{1,2}) \\
&+ \operatorname{tr}(\Sigma^2)b_*^2 q\psi_{*,1}(q)P^{-1} \operatorname{tr}(\Psi_{2,2}) \\
&\approx \operatorname{tr}(\Sigma^2)b_*^2 q\psi_{*,1}(q)P^{-1} \operatorname{tr}(\Psi_{2,2})
\end{aligned} \tag{79}$$

so that

$$\begin{aligned}
&E[\pi'_t R_{t+1}R'_{t+1}\pi_t] - (E[\pi'_t R_{t+1}])^2 \\
&\approx (q\psi_{*,1}(q)b_* \operatorname{tr}(\Sigma))^2\left(H(\Sigma) \;+ (q\psi_{*,1}(q)b_* \operatorname{tr}(\Sigma))^{-1}(1 + H(\Sigma)b_*\psi_{*,1}(1)\right)
\end{aligned} \tag{80}$$

□

D Preliminaries of Random Matrix Theory

First, to understand the central role of $\hat{\Psi}$'s eigenvalue distribution in determining the limiting behavior of the least squares estimator, suppose momentarily that we could replace $\hat{\Psi}$ with its true unobservable signal covariance,

$$\operatorname{tr}(\Sigma/N)\Psi \;=\; \frac{1}{N}E[S'_t S_t] \;=\; E[\hat{\Psi}]. \tag{81}$$

For any symmetric matrix Ψ, a convenient matrix identity states

$$\frac{1}{P}\operatorname{tr}\left((\Psi - zI)^{-1}\right) = \frac{1}{P}\sum_{i=1}^{P}(\lambda_i(\Psi) - z)^{-1}\,,$$

where $\lambda_i(\Psi)$ are the eigenvalues of Ψ. Using formula (5), we can rewrite this identity as

$$\frac{1}{P}\operatorname{tr}\left((\Psi - zI)^{-1}\right) = \int \frac{1}{x - z}dF^{\Psi}(x)\;,\; z < 0\,.$$

From this identity, we immediately see the fundamental connection between ridge regularization and the distribution of eigenvalues for Ψ. The right-side quantity is the *Stieltjes transform* of the eigenvalue distribution of Ψ, denoted F^{Ψ}. By Assumption 3, this distribution is well behaved when $P \to \infty$ and converges to a non-random distribution H. Thus, we

have

$$m_\Psi(z) \coloneqq \int \frac{1}{x-z} dH(x) = \lim_{P\to\infty} \frac{1}{P} \operatorname{tr}\left((\Psi - zI)^{-1}\right). \tag{82}$$

The function $m_\Psi(z)$ is the *limiting* Stieltjes transform of the eigenvalue distribution of Ψ. Equation (82) is a powerful step towards understanding the least squares estimator in the machine learning regime (and hence machine learning predictions and portfolios). It states that key properties of the limiting inverse of the ridge-regularized signal covariance matrix can be completely characterized if we just know Ψ's eigenvalue distribution.

The problem, of course, is that the true Ψ is unobservable. We only observe its sample counterpart, $\hat{\Psi}$, thus we only have empirical access to the Stieltjes transform of $\hat{\Psi}$'s eigenvalues. The empirical counterpart to the unobservable $m_\Psi(z)$ is

$$m(z;c) \coloneqq \lim_{P\to\infty} \frac{1}{P} \operatorname{tr}\left((\hat{\Psi} - zI)^{-1}\right).$$

In traditional finite P statistics, we would have convergence between the sample covariance $\hat{\Psi}$ and the true covariance $\operatorname{tr}(\Sigma/N)\Psi$ as $T \to \infty$. One might be tempted to think that $\lim_{P\to\infty} \frac{1}{P} \operatorname{tr}\left((\hat{\Psi} - zI)^{-1}\right)$ and $\lim_{P\to\infty} \frac{1}{P} \operatorname{tr}\left((\operatorname{tr}(\Sigma/N)\Psi - zI)^{-1}\right)$ also converge as $T \to \infty$. But this is not the case. The limiting eigenvalue distributions of $\hat{\Psi}$ and $\operatorname{tr}(\Sigma/N)\Psi$ remain divergent in the limit as $T \to \infty$ if $P/(NT) \to c > 0$. Here we see a first glimpse of the complexity of machine learning and how we can understand it with random matrix theory. In the Appendix (see Theorem 13), we show how $m(-z;c)$ can be computed from $m_{\operatorname{tr}(\Sigma/N)\Psi}(-z)$ using results of (Silverstein and Bai, 1995) and (Bai and Zhou, 2008). In particular, $m(-z;c) > m(-z;0) = m_{\operatorname{tr}(\Sigma/N)\Psi}(-z)$ for all $c > 0$.[45] The next result shows that, quite remarkably, if we constrain ourselves to linear ridge regression estimators, all asymptotic expressions depend only on $m(z;c)$ and do not require m_Ψ.[46]

Proposition 9 *Recall formulas* (18)*. We have*

$$\lim_{T\to\infty} \frac{1}{NT} \operatorname{tr}((zI + \hat{\Psi})^{-1}\Psi) \;\to\; \xi(z;c) \tag{84}$$

[45] Theorem 13 in the Appendix is a generalized version of the (Marčenko and Pastur, 1967) theorem that accommodates non-i.i.d. S_t. When signals are i.i.d. with $\Psi = I_{P\times P}$, $\Sigma = I_{N\times N}$, and $m_\Psi(z) = (1-z)^{-1}$, (Marčenko and Pastur, 1967) show that

$$m(-z;c) = \frac{2}{((1-c)+z) + \sqrt{((1-c)+z)^2 + 4cz}} \tag{83}$$

By direct calculation, (83) is indeed the unique positive solution to (93) when $m_\Psi(z) = (1-z)^{-1}$. While the eigenvalue distributions of the sample and true covariance matrices do not coincide, Theorem 13 describes the precise non-linear way they relate to each other. In particular, when $P > NT$, the matrix $\hat{\Psi}$ has $P - NT$ zero eigenvalues and therefore, $P^{-1} \operatorname{tr}\left((zI + \hat{\Psi})^{-1}\right)$ contains a singular part, $P^{-1}(P - NT)z^{-1} = (1 - c^{-1})z^{-1}$.

[46] Note that $m_{a\Psi}(-z) = P^{-1} \operatorname{tr}((a\Psi - z)^{-1}) = P^{-1}a^{-1}m_\Psi(-z/a)$. It is possible to develop *non-linear* shrinkage estimators analogous to those developed by (Ledoit and Wolf, 2020) for covariance matrices. Such estimators would require knowledge of the true eigenvalue distribution of Ψ which can be recovered from $m(z;c)$ using equation (93).

in probability, where $\xi(z)$ is the unique solution to

$$1 \;=\; zm(-z;c) + c^{-1}\left(1 - \xi(z)^{-1} m_\Sigma(-\xi(z)^{-1})\right) . \tag{85}$$

When $\Sigma = I$, we get

$$\xi(z) \;=\; \frac{1 - zm(-z;c)}{c^{-1} - 1 + zm(-z;c)} \tag{86}$$

The quantity $\operatorname{tr} E[(zI + \hat{\Psi})^{-1}\Psi]$ appears in virtually every expression we analyze to describe portfolio behavior. It depends on an interaction between the sample and true signal covariance matrix and arises in the computation of both the expected return and leverage of the timing strategy (see equation (8)). One would imagine, then, that we need to know the limiting eigenvalue distribution of both matrices (or their Stieltjes transforms, m and m_Ψ) in order to describe $\operatorname{tr} E[(zI + \hat{\Psi})^{-1}\Psi]$. Proposition 9 shows that this is not the case—we only need to know the empirical version, $m(-z;c)$ as well as $m_\Sigma(-z)$. This is a powerful result. It will allow us to quantify the expected out-of-sample behavior of machine learning portfolios based only on the eigenvalue distribution of the sample signal covariance $\hat{\Psi}$ (which is observable) without requiring us to know the eigenvalues of Ψ.[47]

We refer to the constant c as "model complexity," which (as preceding results show) plays a critical role in understanding model behavior. It describes the limiting ratio of predictors to data points: $P/(NT) \to c$. When T grows at a faster rate than the number of predictors (i.e., $c \to 0$) the limiting eigenvalue distributions of $\hat{\Psi}$ and $\operatorname{tr}(\Sigma/N)\Psi$ in fact converge: $m(-z;0) = m_{\operatorname{tr}(\Sigma/N)\Psi}(-z)$. As c becomes positive, these distributions fail to converge, and their divergence is wider for larger c. It is therefore clear that behavior of the least squares estimator in the machine learning regime will differ from the true coefficient, even when $T \to \infty$, as long as $c > 0$.

E Asymptotic Quantities

We will need the following auxiliary result.

Proposition 10 *In the limit $T, P, P_1 \to \infty$, $P/(NT) \to c$, $P_1/P \to q \in (0, 1]$,*

$$\lim_{T\to\infty} \frac{1}{NT} \operatorname{tr}((zI + \hat{\Psi}_{1,1})^{-1}\Psi_{1,1}) \;\to\; \xi(z; cq; q)$$

in probability, where $\xi(z; cq; q)$ is the unique solution to

$$1 \;=\; zm(-z; cq; q) + c^{-1}\left(1 - \xi(z; cq; q)^{-1} m_\Sigma(-\xi(z; cq; q)^{-1})\right) . \tag{87}$$

[47] Heuristically, $E[\hat{\Psi}] = \operatorname{tr}(\Sigma/N)\Psi$ and hence $\operatorname{tr}(\Sigma/N) \operatorname{tr} E[(zI + \hat{\Psi})^{-1}\Psi] \approx \operatorname{tr} E[(zI + \hat{\Psi})^{-1}\hat{\Psi}]$. However, random matrix corrections make the true relationship non-linear.

Furthermore,

$$\begin{array}{lll} \nu(z;cq;q) & = \psi_{*,1}(q) - (cq)^{-1} z\xi(z;cq;q) & = \lim P^{-1}\operatorname{tr}(\hat{\Psi}_{1,1}(zI+\hat{\Psi}_{1,1})^{-1}\Psi_{1,1}) \quad > \ 0 \\ \nu'(z;cq;q) & = -(cq)^{-1}(\xi(z;cq;q) + z\xi'(z;cq;q)) & = -\lim P^{-1}\operatorname{tr}(\hat{\Psi}_{1,1}(zI+\hat{\Psi}_{1,1})^{-2}\Psi_{1,1}) \quad < \ 0 \\ \hat{\nu}(z;cq;q) & = \nu(z;cq;q) + z\nu'(z;cq;q) & = \lim P^{-1}\operatorname{tr}(\hat{\Psi}_{1,1}^2(zI+\hat{\Psi}_{1,1})^{-2}\Psi_{1,1}) \quad > \ 0. \end{array} \tag{88}$$

In addition, we will also need

$$\begin{aligned} \xi_{2,1}(z;cq;q) &= \lim_{T\to\infty}(NT)^{-1}\operatorname{tr} E[(zI+\hat{\Psi}_{1,1})^{-1}\Psi_{1,2}\Psi_{1,2}'] \ \geq \ 0 \\ \widehat{\xi}_{2,1}(z;cq;q) &= \lim_{T\to\infty}(NT)^{-1}\operatorname{tr} E[(zI+\hat{\Psi}_{1,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{1,1})^{-1}\Psi_{1,2}\Psi_{1,2}'] \ \geq \ 0\,. \end{aligned} \tag{89}$$

While the existence of the limits in (89) cannot be guaranteed in general, the expectations are uniformly bounded for $z > 0$ (since so are the Ψ matrices). Hence, by passing to a subsequence of T, P, we can always assume the limits in (89) exist.

The following is true.

Proposition 11 *The expected return on the market timing strategy converges in probability to*

$$\begin{aligned} &\mathcal{E}(z;cq;q) \coloneqq \lim E[\hat{\pi}_t(z)'R_{t+1}|\hat{\beta}] \\ &= \operatorname{tr}(\Sigma) b_* q\Big(\nu(z;cq;q) \ + \ N^{-1}(cq)^{-1}\xi_{2,1}(z;cq;q)\operatorname{tr}(\Sigma(I+\xi(z;cq;q)\,\Sigma)^{-1})\Big) \end{aligned} \tag{90}$$

Proposition 12 *We have*

i) Expected leverage converges in probability to

$$\begin{aligned} &\mathcal{L}(z;cq;q) \coloneqq \lim \operatorname{tr} E[\hat{\pi}_t(z)\hat{\pi}_t(z)'|\hat{\beta}] \\ &= \operatorname{tr}(\Sigma) q\Big(\underbrace{b_*\hat{\nu}(z;cq;q)}_{signal\ \ variance} \ \underbrace{-\ c\nu'(z;cq;q)}_{noise}\Big) \\ &+ \ b_*\operatorname{tr}(\Sigma)\underbrace{\frac{\operatorname{tr}(\Psi_{2,2})}{P}\gamma(z;cq;q)(-cq\nu'(z;cq;q))}_{missing\ \ signal\ \ variance} \\ &+ \ b_*\operatorname{tr}(\Sigma) \quad \underbrace{\Gamma(z;cq;q)}_{covariance\ \ with\ \ missing\ \ signals} \end{aligned} \tag{91}$$

where

$$\begin{aligned}\gamma(z;cq;q) &= \frac{\operatorname{tr}(\Sigma^2(I+\xi(z;cq;q)\,\Sigma)^{-2})}{\operatorname{tr}(\Sigma(I+\xi(z;cq;q)\Sigma)^{-2})}\\ \Gamma(z;cq;q) &= (N^{-1}\operatorname{tr}((I+\xi_{1,1}(z)\,\Sigma)^{-1}\Sigma))^2 c^{-1}\widehat{\xi}_{2,1}(z)\\ &+\ 2q\nu'(z;cq;q)\xi_{2,1}(z;cq;q)\,\frac{\frac{1}{N^2}\operatorname{tr}((I+\xi_{1,1}(z)\,\Sigma)^{-1}\Sigma)\operatorname{tr}((I+\xi_{1,1}(z)\,\Sigma)^{-2}\Sigma^2)}{\operatorname{tr}((I+\xi_{1,1}(z)\,\Sigma)^{-2}\Sigma/N)}\end{aligned}$$

ii) Out-of-sample R^2 converges in probability to

$$R^2(z;cq;q) = \frac{2\mathcal{E}(z;cq;q) - \mathcal{L}(z;cq;q)}{N + b_*\psi_{*,1}(1)\operatorname{tr}(\Sigma)} \tag{92}$$

F Asymptotic Eigenvalue Distribution of $\hat{\Psi}$

The following result of (Silverstein and Bai, 1995) and (Bai and Zhou, 2008) relates the limiting eigenvalue of distribution of $\hat{\Psi}$ to that of Ψ.

Theorem 13 *Suppose that all diagonal elements of Σ are identical.*[48] *For any $c > 0$ and $z < 0$, the distribution of eigenvalues of $\hat{\Psi}$ in the limit as $P, T \to \infty$, $P/(NT) \to c$ converges to a distribution whose Stieltjes transform, $m(z;c)$, is the unique positive solution to the equation*

$$m(z;c) = \frac{1}{1\ -\ c\ -\ c\,z\,m(z;c)}\,m_\Psi\left(\frac{z}{1\ -\ c\ -\ c\,z\,m(z;c)}\right). \tag{93}$$

Furthermore, for $c > 1$, there exists functions $m_*(c) > 0 > n_*(c)$ such that $cm_*(c)$ is monotone decreasing in c and

$$m(-z;c) = (1-c^{-1})z^{-1}\ +\ m_*(c)\ +\ n_*(c)\,z\ +\ O(z^2).$$

Proof of Theorem 13. To prove he result, we note that we can write

$$\hat{\Psi}\ =\ \frac{1}{NT}\sum_i\sum_t S_t(i)S_t(i)' \tag{94}$$

where $S_t(i) \in \mathbb{R}^P$ is the signal vector for stock i at time t. to be able to apply the result of (Bai and Zhou, 2008), we need to show that

$$E[S_t(i)S_t(i)']\ =\ \Psi\ \operatorname{tr}(\Sigma) \tag{95}$$

The claim follows because $S_t\ =\ \Sigma^{1/2}X_t\Psi^{1/2}$ and, hence,

$$S_t(i)\ =\ \Psi^{1/2}X_t'(\Sigma^{1/2})_i \tag{96}$$

[48] Our results without this assumption become less elegant.

where $(\Sigma^{1/2})_i$ is the i'th column of the matrix $\Sigma^{1/2}$, and hence,

$$E[S_t(i)S_t(i)'] \;=\; \Psi\,\mathrm{tr}((\Sigma^{1/2})_i(\Sigma^{1/2})_i') \;=\; \Psi\,\Sigma_{i,i}\,. \tag{97}$$

By Assumption 2, $\Sigma_{i,i}$ are all identical and hence $\Sigma_{i,i} = N^{-1}\,\mathrm{tr}(\Sigma)$.

It remains to verify the second technical condition of (Bai and Zhou, 2008): that

$$E[(S_t(i)'BS_t(i) - \mathrm{tr}(B\Psi\,\frac{1}{N}\,\mathrm{tr}(\Sigma)))^2] \;=\; o((NT)^2).$$

for any bounded B. This can be verified by direct calculation:

$$\begin{aligned}
&E[(S_t(i)'BS_t(i) - \mathrm{tr}(B\Psi\,\frac{1}{N}\,\mathrm{tr}(\Sigma)))^2] \\
&= E[S_t(i)'BS_t(i)S_t(i)'BS_t(i) - 2S_t(i)'BS_t(i)\,\mathrm{tr}(B\Psi\,\frac{1}{N}\,\mathrm{tr}(\Sigma)) + (\mathrm{tr}(B\Psi\,\frac{1}{N}\,\mathrm{tr}(\Sigma)))^2] \\
&= \mathrm{tr}\,E[BS_t(i)S_t(i)'BS_t(i)S_t(i)'] - (\mathrm{tr}(B\Psi\,\frac{1}{N}\,\mathrm{tr}(\Sigma)))^2
\end{aligned} \tag{98}$$

and then the required technical condition follows from the results of (Yaskov, 2016) and Assumption 2 (the Lindenberg condition). The proof is complete. □

G Proof of Proposition 9

Proposition 14 *We have*

$$\frac{1}{NT}\,\mathrm{tr}(zI + \hat{\Psi}_T)^{-1}\Psi - \mathrm{tr}\,E[\frac{1}{NT}(zI + \hat{\Psi}_T)^{-1}\Psi] \;\to\; 0$$

almost surely, while

$$\lim_{T\to\infty}\frac{1}{NT}\,\mathrm{tr}\,E[(zI + \hat{\Psi}_T)^{-1}\Psi] \;\to\; \xi(z)\,, \tag{99}$$

where $\xi(z)$ is the unique solution to (85).

Proof of Proposition 14. The proof will be based on the following general result.

Lemma 6 *Let Ψ_T be any sequence of (random) matrices that is uniformly bounded and independent of $\hat{\Psi}_T$. Then,*

$$\frac{1}{NT}\,\mathrm{tr}(zI + \hat{\Psi}_T)^{-1}\Psi_T - E[\frac{1}{NT}\,\mathrm{tr}(zI + \hat{\Psi}_T)^{-1}\Psi_T] \;\to\; 0$$

almost surely.

Proof. Let $\hat{\Psi}_{T,t} \;=\; \frac{1}{NT}\sum_{\tau\neq t} S_\tau' S_\tau$. We will be using the Woodbury matrix identity

$$(zI+\hat{\Psi}_T)^{-1} \;=\; (zI+\hat{\Psi}_{T,t})^{-1} - \frac{1}{NT}(zI+\hat{\Psi}_{T,t})^{-1}S_t'(I+\frac{1}{NT}S_t(zI+\hat{\Psi}_{T,t})^{-1}S_t')^{-1}S_t(zI+\hat{\Psi}_{T,t})^{-1}$$

(100)

The proof of the first claim is based on a martingale argument from (Bai and Zhou, 2008). Let E_τ denote the conditional expectation given $S_{\tau+1}, \cdots, S_T$. Let also

$$q_T(z) = \frac{1}{P} \operatorname{tr}(zI + \hat{\Psi}_T)^{-1} \Psi_T$$

With this notation, since $\hat{\Psi}_{T,t}$ is independent of S_t, we have

$$(E_{t-1} - E_t)[\frac{1}{P} \operatorname{tr}(zI + \hat{\Psi}_{t,T})^{-1} \Psi] \;=\; 0$$

and therefore

$$\begin{aligned}
E[q_T(z)] - q_T(z) \;&=\; E_0[q_T(z)] - E_T[q_T(z)] = \sum_{t=1}^{T} (E_{t-1}[q_T(z)] - E_t[q_T(z)]) \\
&= \sum_{t=1}^{T} (E_{t-1} - E_t)[q_T(z)] \\
&= \sum_{t=1}^{T} (E_{t-1} - E_t)[q_T(z) - \frac{1}{P} \operatorname{tr}(zI + \Psi_{T,t})^{-1} \Psi_T] \qquad (101) \\
&= \frac{1}{P} \sum_{t=1}^{T} (E_{t-1} - E_t)[\operatorname{tr}(zI + \hat{\Psi}_T)^{-1} \Psi_T - \operatorname{tr}(zI + \hat{\Psi}_{T,t})^{-1} \Psi_T] \\
&= -\frac{1}{P} \sum_{\tau=1}^{T} (E_{t-1} - E_t)[\gamma_t] \,,
\end{aligned}$$

where we have used (100) and defined

$$\gamma_t \;=\; \operatorname{tr}\left(\frac{1}{NT} (zI + \hat{\Psi}_{T,t})^{-1} S_t' (I + \frac{1}{NT} S_t (zI + \hat{\Psi}_{T,t})^{-1} S_t')^{-1} S_t (zI + \hat{\Psi}_{T,t})^{-1} \Psi \right) \qquad (102)$$

We will need the following properties of the trace:

Lemma 7 *If A, B are symmetric positive semi-definite, then*

$$\operatorname{tr}(AB) \;\leq\; \operatorname{tr}(A) \|B\|$$

and

$$\operatorname{tr}(A^{1/2} B A^{1/2}) \;\leq\; \operatorname{tr}(B) \, \|A\|$$

Thus,

$$\begin{aligned}\|\gamma_t\| &\leq \|\Psi\| \operatorname{tr}\left(\frac{1}{NT}(zI+\hat{\Psi}_{T,t})^{-1}S_t'(I+\frac{1}{NT}S_t(zI+\hat{\Psi}_{T,t})^{-1}S_t')^{-1}S_t(zI+\hat{\Psi}_{T,t})^{-1}\right) \\ &\leq z^{-1}\operatorname{tr}\left(\frac{1}{NT}(zI+\hat{\Psi}_{T,t})^{-1/2}S_t'(I+\frac{1}{NT}S_t(zI+\hat{\Psi}_{T,t})^{-1}S_t')^{-1}S_t(zI+\hat{\Psi}_{T,t})^{-1/2}\right) \\ &= z^{-1}\operatorname{tr}(B(zI+B)^{-1}) \leq Nz^{-1}.\end{aligned} \tag{103}$$

where

$$B = \frac{1}{NT}S_t(zI+\hat{\Psi}_{T,t})^{-1}S_t' \in \mathbb{R}^{N\times N}.$$

Thus,

$$(E_{t-1}-E_t)[\operatorname{tr}(zI+\hat{\Psi}_T)^{-1}\Psi] = (E_{t-1}-E_t)[\gamma_t]$$

forms a bounded martingale difference sequence. Applying the Burkholder-Davis-Gundy inequality (see, e.g., (Burkholder, 1966)), we get

$$\begin{aligned}E[|q_T(z)-E[q_T(z)]|^q] &\leq K_q M^{-q} E\left(\sum_{t=1}^{T}|(E_{t-1}-E_t)[\gamma_t]|^2\right)^{q/2} \\ &\leq K_q(2N/z)^q M^{-q/2}\left(\frac{P}{T}\right)^{-q/2}.\end{aligned} \tag{104}$$

Almost sure convergence follows with $q>2$ from the following lemma.

Lemma 8 *Suppose that*

$$E[|X_T|^q] \leq T^{-\alpha}$$

for some $\alpha>1$ and some $q>0$. Then, $X_T\to 0$ almost surely.

Proof. It is known that if

$$\sum_{T=1}^{\infty} Prob(|X_T|>\varepsilon) < \infty$$

for any $\varepsilon>0$, then $X_T\to 0$ almost surely. In our case, the Chebyshev inequality implies that

$$Prob(|X_T|>\varepsilon) \leq \varepsilon^{-q}E[|X_T|^q] \leq T^{-\alpha}$$

and convergence follows because $\alpha>1$. □

The proof of Lemma 6 is complete. □

The proof of the second claim of Proposition 14 is based on the identity

$$P \;=\; \mathrm{tr}\, E[(zI+\hat{\Psi}_T)^{-1}(zI+\hat{\Psi}_T)]\,.$$

Then,

$$P \;=\; \mathrm{tr}\, E[(zI+\hat{\Psi}_T)^{-1}(zI+\hat{\Psi}_T)] \;=\; Pz\, m(-z) \;+\; \frac{1}{NT}\sum_t \mathrm{tr}\, E[(zI+\hat{\Psi}_T)^{-1}S_t'S_t]\,.$$

Since signals are i.i.d. over time, we get that all these expectations are in fact identical and we get

$$1 \;=\; z\, m(-z) + \frac{1}{PN}\,\mathrm{tr}\, E[(zI+\hat{\Psi}_T)^{-1}S_t'S_t] \;=\; z\, m(-z) + \frac{c^{-1}N^{-1}}{NT}\,\mathrm{tr}\, E[(zI+\hat{\Psi}_T)^{-1}S_t'S_t] \tag{105}$$

for any t. Hence, by (100),

$$\begin{aligned}
&\mathrm{tr}\, E[(zI+\hat{\Psi}_T)^{-1}S_t'S_t]\\
&=\; \mathrm{tr}\, E[S_t\Big((zI+\hat{\Psi}_{T,t})^{-1}\\
&-\; \frac{1}{NT}(zI+\hat{\Psi}_{T,t})^{-1}S_t'(I+\frac{1}{NT}S_t(zI+\hat{\Psi}_{T,t})^{-1}S_t')^{-1}S_t(zI+\hat{\Psi}_{T,t})^{-1}\Big)S_t']\\
&=\; \mathrm{tr}\, E[B-(NT)^{-1}B^2(I+(NT)^{-1}B)^{-1}] \;=\; \mathrm{tr}\, E[B_T(I+(NT)^{-1}B_T)^{-1}]
\end{aligned} \tag{106}$$

where

$$B_T \;=\; S_t(zI+\hat{\Psi}_{T,t})^{-1}S_t'$$

and where we have used that $\mathrm{tr}(ZS)=\mathrm{tr}(SZ)$ for any two matrices $S,\ Z$. We will first prove the following result.

Lemma 9

$$\frac{1}{NT}S_t(zI+\hat{\Psi}_{T,t})^{-1}S_t' - \xi_T(z)\,\Sigma \;\to\; 0$$

in probability.

Proof of Lemma 9. We have

$$(S_t(zI+\hat{\Psi}_{T,t})^{-1}S_t')_{i,j} \;=\; (S_t(i)'(zI+\hat{\Psi}_{T,t})^{-1}S_t(j)) \;=\; \mathrm{tr}((zI+\hat{\Psi}_{T,t})^{-1}S_t(j)S_t(i)')\,.$$

and Lemma 6 implies that

$$\frac{1}{NT}\,\mathrm{tr}((zI+\hat{\Psi}_{T,t})^{-1}S_t(j)S_t(i)') \;-\; \frac{1}{NT}\,\mathrm{tr}\left(E[((zI+\hat{\Psi}_{T,t})^{-1})]S_t(j)S_t(i)'\right) \;\to\; 0$$

almost surely.[49] Now,

$$\frac{1}{NT}\operatorname{tr}\left(E[((zI+\hat{\Psi}_{T,t})^{-1})]S_t(j)S_t(i)'\right) \;=\; \frac{1}{NT}S_t(i)'E[((zI+\hat{\Psi}_{T,t})^{-1})]S_t(j)$$

and the same argument as in the proof of Theorem 13, we can replace $S_t(j)S_t(i)'$ by its expectation and get

Lemma 10 *We have*

$$\frac{1}{NT}\operatorname{tr}\left(C_T S_t(j)S_t(i)'\right) \;-\; \frac{1}{NT}\operatorname{tr}\left(C_T E[S_t(j)S_t(i)']\right) \;\to\; 0$$

in probability for any uniformly bounded sequence of matrices C_T that are independent of S_t.

We have $S_t(i) = \Psi^{1/2}X_t'Q(i)$ where $Q(i)$ is the i-th column of $Q = \Sigma^{1/2}$, $i = 1, \cdots, N$. Hence,

$$E[S_t(j)S_t(i)'] \;=\; E[\Psi^{1/2}X_t'Q(i)Q(j)'X_t\Psi^{1/2}] \;=\; \Psi\,\operatorname{tr}(Q(i)Q(j)') \;=\; \Psi\left(Q(j)'Q(i)\right) \;=\; \Psi(Q'Q)_{i,j}$$

where $Q'Q = \Sigma$. The proof of Lemma 9 is complete.

□

Lemma 9 implies that[50]

$$\frac{1}{NT}\operatorname{tr}E[B_T(I+(NT)^{-1}B_T)^{-1}] \;-\; \xi_T(z)\operatorname{tr}\left(\Sigma(I+\xi_T(z)\Sigma)^{-1}\right) \;\to\; 0$$

and thus equation (105) implies that

$$zm_T(-z) + c^{-1}N^{-1}\xi_T(z)\operatorname{tr}\left(\Sigma(I+\xi_T(z)\Sigma)^{-1}\right) \;-1 \;\to\; 0\,.$$

The final claim follows now because $m_T(z) \to m(z)$ and the left-hand side is continuous and strictly monotonic in ξ. The proof of Proposition 14 is complete. □

H Proofs for the Misspecified Model

We will use $c_1 = P_1/(TN) = cq$ to denote the complexity of the mis-specified model, so that $q = c_1/c$. We will also use $\hat{\beta}_1$ to denote the estimated β vector for the observable part of the signals, $S_{t,1}$. We start with the observation that

$$\begin{aligned} &E[\|R_{t+1} - S_{t,1}\hat{\beta}\|^2] \\ &= \operatorname{tr}E[R_{t+1}R_{t+1}'] \;-\; 2E[\beta'S_t'S_{t,1}\hat{\beta}_1] \;+\; \operatorname{tr}E[S_{t,1}\hat{\beta}_1\hat{\beta}_1'S_{t,1}'] \qquad (107)\\ &= \operatorname{tr}E[R_{t+1}R_{t+1}'] \;-\; 2E[\beta'S_t'S_{t,1}\hat{\beta}_1] \;+\; \operatorname{tr}(\Sigma)\operatorname{tr}E[\Psi_{1,1}\hat{\beta}_1\hat{\beta}_1'] \end{aligned}$$

[49] While S_t are not uniformly bounded, the proof follows by a standard truncation argument.

[50] The claim follows from convergence in probability because $\frac{1}{NT}B_T(I+(NT)^{-1}B_T)^{-1}$ is uniformly bounded: B_T is symmetric positive semi-definite and, hence, $\|\frac{1}{NT}B_T(I+(NT)^{-1}B_T)^{-1}\| \le 1$.

H.1 Computing the First Moment of the Efficient Portfolio

Lemma 11 *We have*

$$E[\beta' S_t' S_{t,1} \hat{\beta}_1 | \hat{\beta}_1] \;\to\; b_* \operatorname{tr}(\Sigma) \frac{c_1}{c} (\psi_{*,1} - c_1^{-1} z \xi(z)) \;+\; b_* \operatorname{tr}(\Sigma) N^{-1} c^{-1} \xi_{2,1,T}(z) \operatorname{tr}(\Sigma (I + \xi_{1,1}(z)\, \Sigma)^{-1}) \,, \tag{108}$$

in probability, where

$$\xi_{2,1}(z) \;=\; \lim_{T\to\infty} \frac{1}{NT} \operatorname{tr} E[\Psi_{1,2} \Psi_{2,1} (zI + \hat{\Psi}_{T,1,t})^{-1}] \tag{109}$$

In particular, if $T^{-1} \operatorname{tr}(\Psi_{1,2}\Psi_{2,1}) \to 0$, *we get*

$$E[\beta' S_t' S_{t,1} \hat{\beta}_1 | \hat{\beta}_1] \;\to\; b_* \operatorname{tr}(\Sigma) \frac{c_1}{c} (\psi_{*,1} - c_1^{-1} z \xi(z)) \tag{110}$$

Proof of Lemma 11. We have

$$S_t \beta \;=\; S_{t,1} \beta_1 + S_{t,2} \beta_2$$

$$\frac{1}{NT} \sum_t F_t \;=\; \frac{1}{NT} \sum_{t=1}^{T} S_{t,1}' (S_t \beta + \varepsilon_{t+1}) \;=\; \hat{\Psi}_T \beta \;+\; q_T \tag{111}$$

where

$$q_T \;=\; \frac{1}{NT} \sum_{t=1}^{T} S_{t,1}' \varepsilon_{t+1} \tag{112}$$

and

$$\hat{\Psi}_T \beta \;=\; \hat{\Psi}_{T,1} \beta_1 + \hat{\Psi}_{T,2} \beta_2$$

where

$$\hat{\Psi}_{T,k} \;=\; \frac{1}{NT} \sum_{t=1}^{T} S_{t,1}' S_{t,k}, \;\; k = 1, 2 \,,$$

Therefore,

$$\hat{\beta} \;=\; (zI + \hat{\Psi}_{T,1})^{-1} (\hat{\Psi}_{T,1} \beta_1 + \hat{\Psi}_{T,2} \beta_2 \;+\; q_T) \tag{113}$$

Using this identity and Assumption 4, we have (using that ε_t are independent of S_t and have

zero means) that

$$
\begin{aligned}
& E[\beta' S_t' S_{t,1} \hat{\beta}] \\
&= \operatorname{tr}(\Sigma) E[(\beta_1' \Psi_{1,1} + \beta_2' \Psi_{2,1})(zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_{T,1}\beta_1 + \hat{\Psi}_{T,2}\beta_2)] \\
&= \operatorname{tr}(\Sigma) \operatorname{tr} E[\Psi_{1,1}\,(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}\beta_1\beta_1'] \\
&+ \operatorname{tr}(\Sigma) E[\beta_1' \Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\beta_2] \\
&+ \operatorname{tr}(\Sigma) E[\beta_2' \Psi_{2,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}\beta_1] \\
&+ \operatorname{tr}(\Sigma) E[\beta_2' \Psi_{2,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\beta_2] \\
&= \{\text{using that } E[\beta\beta'] = b_* P^{-1} I\} \\
&= \operatorname{tr}(\Sigma) b_*\, P^{-1} \operatorname{tr} E[\Psi_{1,1}\,(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}] \;+\; \operatorname{tr}(\Sigma) P^{-1} b_* \operatorname{tr} E[\Psi_{2,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}]
\end{aligned}
\tag{114}
$$

The first term satisfies

$$
P^{-1} \operatorname{tr} E[\Psi_{1,1}\,(zI + \hat{\Psi}_{T,1})^{-1}(zI + \hat{\Psi}_{T,1} - zI)] \;\to\; P^{-1} P_1(\psi_{*,1} - c_1^{-1} z\xi(z)) \tag{115}
$$

in probability by Proposition 14. To deal with the second term, we will need a technical lemma.

Lemma 12

$$
\frac{1}{P} \operatorname{tr} E[\Psi_{2,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}] \;\to\; c^{-1}\xi_{2,1}(z)\xi_{1,1}^{-1}(z)c_1(1 - zm(-z; c_1))\,, \tag{116}
$$

where

$$
\xi_{2,1}(z) \;=\; \lim_{T\to\infty} \frac{1}{NT} \operatorname{tr} E[\Psi_{1,2}\Psi_{2,1}(zI + \hat{\Psi}_{T,1,t})^{-1}]\,. \tag{117}
$$

Proof of Lemma 12. We have that, by symmetry over time, and using the Woodbury matrix identity

$$
(zI+\hat{\Psi}_{T,1})^{-1} \;=\; (zI+\hat{\Psi}_{T,1,t})^{-1} - \frac{1}{NT}(zI+\hat{\Psi}_{T,1,t})^{-1} S_{t,1}'(I+\frac{1}{NT} S_{t,1}(zI+\hat{\Psi}_{T,1,t})^{-1} S_{t,1}')^{-1} S_{t,1}(zI+\hat{\Psi}_{T,1,t})^{-1} \tag{118}
$$

implies

$$
\begin{aligned}
&\frac{1}{P}\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}] \\
&= \frac{1}{P}\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1})^{-1}\frac{1}{NT}\sum_{t=1}^{T} S'_{t,1}S_{t,2}] \\
&= \frac{1}{NP}\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1})^{-1}S'_{t,1}S_{t,2}] \\
&= \frac{1}{NP}\operatorname{tr} E[\Psi_{2,1}\Big((zI+\hat{\Psi}_{T,1,t})^{-1} \\
&- \frac{1}{NT}(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t,1}(I+\frac{1}{NT}S_{t,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t,1})^{-1}S_{t,1}(zI+\hat{\Psi}_{T,1,t})^{-1}\Big)S'_{t,1}S_{t,2}] \\
&= \frac{1}{NP}\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t,1}S_{t,2}] \\
&- \frac{1}{NP}\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t,1}(I+C_T)^{-1}C_T S_{t,2}] \\
&= \frac{1}{NP}\operatorname{tr}(\Sigma)\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t})^{-1}\Psi'_{2,1}] \\
&- \frac{1}{NP}\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t,1}(I+C_T)^{-1}C_T S_{t,2}]
\end{aligned}
\tag{119}
$$

By Lemma 9,

$$
C_T = \frac{1}{NT}S_{t,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t,1} \to \xi_{1,1}(z)\,\Sigma
$$

in probability. By a similar argument,

$$
\frac{1}{NT}S_{t,2}\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t,1} \to \xi_{2,1,T}(z)\,\Sigma \tag{120}
$$

where we have defined

$$
\xi_{2,1,T}(z) = \frac{1}{NT}\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t})^{-1}\Psi'_{2,1}]\,, \tag{121}
$$

where

$$
c_1 = P_1/(NT) = c_{P_1}/N. \tag{122}
$$

Thus,[51]

$$
\begin{aligned}
&\frac{1}{P}\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}] \\
&= \frac{1}{NP}\operatorname{tr}(\Sigma)\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t})^{-1}\Psi_{2,1}'] \\
&- \frac{1}{NP}\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S_{t,1}'(I+C_T)^{-1}C_T S_{t,2}] \\
&\approx N^{-1}c^{-1}\xi_{2,1,T}(z)\operatorname{tr}(\Sigma) \;-\; N^{-1}c^{-1}\operatorname{tr} E[\xi_{2,1,T}(z)\,\Sigma(I+\xi_{1,1}(z)\,\Sigma)^{-1}\xi_{1,1}(z)\,\Sigma] \\
&= N^{-1}c^{-1}\xi_{2,1,T}(z)\operatorname{tr}(\Sigma(I-(I+\xi_{1,1}(z)\,\Sigma)^{-1}\xi_{1,1}(z)\,\Sigma)) \\
&= N^{-1}c^{-1}\xi_{2,1,T}(z)\operatorname{tr}(\Sigma(I+\xi_{1,1}(z)\,\Sigma)^{-1}) \\
&= N^{-1}c^{-1}\xi_{2,1,T}(z)\xi_{1,1}^{-1}(z)\operatorname{tr}(\xi_{1,1}(z)\Sigma(I+\xi_{1,1}(z)\,\Sigma)^{-1}) \\
&= c^{-1}\xi_{2,1,T}(z)\xi_{1,1}^{-1}(z)N^{-1}(N-\operatorname{tr}((I+\xi_{1,1}(z)\,\Sigma)^{-1})) \\
&= c^{-1}\xi_{2,1,T}(z)\xi_{1,1}^{-1}(z)c_1(1-zm(-z;c_1))\,,
\end{aligned}
\tag{123}
$$

where we have used the defining equation for $\xi_{1,1}$:

$$
1-zm(-z;c_1) = c_1^{-1} - c_1^{-1}N^{-1}\operatorname{tr}((I+\xi_{1,1}(z)\Sigma)^{-1})\,. \tag{124}
$$

The proof of Lemma 12 is complete. □

The proof of Lemma 11 is complete. □

H.2 $\mathcal{L}$ in the correctly specified case

To prove the desired result, we first consider the correctly specified case.

Lemma 13 *In the correctly specified case,*

$$
\operatorname{tr}(\Psi\hat{\beta}\hat{\beta}') \to \;\; b_*(\psi_{*,1} - 2zc^{-1}\xi(z) - z^2c^{-1}\xi'(z)) + \xi(z) + z\xi'(z) \tag{125}
$$

in probability.

Proof of Lemma 13.

$$
\begin{aligned}
&\operatorname{tr}(\Psi E[\hat{\beta}\hat{\beta}']) \\
&= \operatorname{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}(\hat{\Psi}_T\beta \;+\; q_T)(\hat{\Psi}_T\beta \;+\; q_T)'(zI+\hat{\Psi}_T)^{-1}]) \\
&= \operatorname{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}(\hat{\Psi}_T\beta \;+\; q_T)(\beta'\hat{\Psi}_T \;+\; q_T')(zI+\hat{\Psi}_T)^{-1}]) \\
&= \operatorname{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}(\hat{\Psi}_T\beta\beta'\hat{\Psi}_T \;+\; q_Tq_T')(zI+\hat{\Psi}_T)^{-1}])
\end{aligned}
\tag{126}
$$

[51] Passage to the limit is easy justified using the boundedness of $(I+C_T)^{-1}C_T$.

where we have used that $E[q_T|S] = 0$. Now,

$$\begin{aligned}
E[q_T q_T'|S] &= \frac{1}{N^2T^2}E[\sum_t S_t'\varepsilon_{t+1}\sum_{t_1}\varepsilon_{t_1+1}S_{t_1}|S] \\
&= \frac{1}{N^2T^2}E[\sum_{t,t_1} S_t'\varepsilon_{t+1}\varepsilon_{t_1+1}'S_{t_1}|S] \\
&= \frac{1}{N^2T^2}E[\sum_t S_t'\varepsilon_{t+1}\varepsilon_{t+1}'S_t|S] \\
&= \frac{1}{N^2T^2}\sum_t S_t'E[\varepsilon_{t+1}\varepsilon_{t+1}'|S]S_t \\
&= \frac{1}{N^2T^2}\sum_t S_t'S_t \\
&= \frac{1}{N^2T^2}\sum_t S_t'S_t = \frac{1}{NT}\hat{\Psi}_T
\end{aligned} \tag{127}$$

and therefore (126) takes the form

$$\begin{aligned}
&\mathrm{tr}(\Psi E[\hat{\beta}\hat{\beta}']) \\
&= \mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}(\hat{\Psi}_T\beta\beta'\hat{\Psi}_T + q_Tq_T')(zI+\hat{\Psi}_T)^{-1}]) \\
&= \mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}(\hat{\Psi}_T\beta\beta'\hat{\Psi}_T + \frac{1}{NT}\hat{\Psi}_T)(zI+\hat{\Psi}_T)^{-1}]) \\
&= \mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T\beta\beta'\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}]) \\
&+ \mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}\frac{1}{NT}\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}]) \\
&= \mathrm{tr}\, E[\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}\Psi(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T\beta\beta'] \\
&+ \mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}\frac{1}{NT}(zI+\hat{\Psi}_T - zI)(zI+\hat{\Psi}_T)^{-1}]) \\
&= \{\text{using that } E[\beta\beta'] = b_*P^{-1}I\} \\
&= b_*P^{-1}\,\mathrm{tr}\, E[\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}\Psi(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T] \\
&+ \frac{1}{NT}\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}(zI+\hat{\Psi}_T)(zI+\hat{\Psi}_T)^{-1}]) \\
&- z\frac{1}{NT}\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}(zI+\hat{\Psi}_T)^{-1}]) \\
&= b_*P^{-1}\,\mathrm{tr}\, E[\Psi(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}] \\
&+ \frac{1}{NT}\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}]) \\
&- z\frac{1}{NT}\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-2}]) \\
&= Term1 + Term2 + Term3
\end{aligned} \tag{128}$$

We now proceed with each term:

$$\begin{aligned}
&(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1} \;=\{\text{all matrices commute}\}\; = \;(zI+\hat{\Psi}_T)^{-2}\hat{\Psi}_T^2\\
&= \;(zI+\hat{\Psi}_T)^{-2}(\hat{\Psi}_T^2+2z\hat{\Psi}_T+z^2I-2z\hat{\Psi}_T-z^2I)\\
&= \;(zI+\hat{\Psi}_T)^{-2}(\hat{\Psi}_T^2+2z\hat{\Psi}_T+z^2I-2z(\hat{\Psi}_T+zI)+z^2I)\\
&= \;(zI+\hat{\Psi}_T)^{-2}\Big((zI+\hat{\Psi}_T)^2-2z(\hat{\Psi}_T+zI)+z^2I\Big)\\
&= \;I\;-\;2z(zI+\hat{\Psi}_T)^{-1}+z^2(zI+\hat{\Psi}_T)^{-2}
\end{aligned} \tag{129}$$

Therefore

$$\begin{aligned}
Term1 \;&=\; b_*P^{-1}\,\mathrm{tr}\,E[\Psi(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}]\\
&=\; b_*P^{-1}\,\mathrm{tr}\,E[\Psi(I\;-\;2z(zI+\hat{\Psi}_T)^{-1}+z^2(zI+\hat{\Psi}_T)^{-2})]
\end{aligned} \tag{130}$$

and

$$Term2 \;=\; \frac{1}{NT}\,\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}]) \;\to\; \xi(z) \tag{131}$$

and hence

$$\frac{d}{dz}\frac{1}{NT}\,\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}]) \;\to\; \xi'(z) \tag{132}$$

However,

$$\frac{d}{dz}\,\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}]) \;=\; -\,\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-2}]) \tag{133}$$

and hence

$$\frac{1}{NT}\,\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-2}]) \;\to\; -\xi'(z)\,. \tag{134}$$

Summarizing, we get

$$Term3 \;=\; z\xi'(z) \tag{135}$$

whereas

$$Term1 \;=\; b_*P^{-1}\,\mathrm{tr}\,E[\Psi(I-2z(zI+\hat{\Psi}_T)^{-1}+z^2(zI+\hat{\Psi}_T)^{-2})] \;=\; b_*(\psi_{*,1}-2zc^{-1}\xi(z)-z^2c^{-1}\xi'(z)) \tag{136}$$

and hence

$$\begin{aligned}
&\mathrm{tr}(\Psi E[\hat{\beta}\hat{\beta}'])= \;Term1\;+\;Term2\;+\;Term3\\
&=\; b_*(\psi_{*,1}-2zc^{-1}\xi(z)-z^2c^{-1}\xi'(z))+\xi(z)+z\xi'(z)
\end{aligned} \tag{137}$$

The proof of Lemma 13 is complete. □

H.3 $\mathcal{L}$ in the Misspecified Case

Note that differentiating

$$1 - zm(-z;c) \;=\; c^{-1}N^{-1}(N - \mathrm{tr}((I+\xi(z)\Sigma)^{-1})) \tag{138}$$

we get

$$\xi'(z) \;=\; \frac{-m(-z,c) + zm'(-z;c)}{c^{-1}N^{-1}\,\mathrm{tr}(\Sigma(I+\xi(z)\Sigma)^{-2})} \tag{139}$$

Recall also that

$$\begin{aligned}
&\xi_{2,1}(z;cq;q) = \lim_{T\to\infty}(NT)^{-1}\,\mathrm{tr}\,E[(zI+\hat\Psi_{1,1})^{-1}\Psi_{1,2}\Psi_{1,2}'] \;\geq\; 0 \\
&\widehat{\xi}_{2,1}(z;cq;q) = \lim_{T\to\infty}(NT)^{-1}\,\mathrm{tr}\,E[(zI+\hat\Psi_{1,1})^{-1}\Psi_{1,1}(zI+\hat\Psi_{1,1})^{-1}\Psi_{1,2}\Psi_{1,2}'] \;\geq\; 0\,.
\end{aligned} \tag{140}$$

Lemma 14 *Let*

$$\widehat{\xi}_{1,1}(z) \;=\; \frac{(\xi_{1,1}(z)+z\xi_{1,1}'(z))c_1^{-1}\xi_{1,1}'(z)}{-m(-z;c_1)+zm'(-z;c_1)} \;=\; \frac{(\xi_{1,1}(z)+z\xi_{1,1}'(z))}{N^{-1}\,\mathrm{tr}(\Sigma(I+\xi(z)\Sigma)^{-2})}\,. \tag{141}$$

Then,

$$\begin{aligned}
&(\mathrm{tr}(\Sigma))^{-1}\mathcal{L}(c;cq;c) \;=\; \mathrm{tr}(\Psi_{1,1}\hat\beta\hat\beta') \;\to \\
&b_*\frac{c_1}{c}(\psi_{*,1} - 2zc_1^{-1}\xi_{1,1}(z) - z^2c_1^{-1}\xi_{1,1}'(z)) + \xi_{1,1}(z) + z\xi_{1,1}'(z) \;+\; \Delta(z)
\end{aligned} \tag{142}$$

in probability, where

$$\begin{aligned}
\Delta(z) \;&=\; \frac{\mathrm{tr}(\Psi_{2,2})}{P}\,b_*\widehat{\xi}_{1,1}(z)\frac{1}{N}\,\mathrm{tr}[\Sigma^2(I+\xi_{1,1}(z)\,\Sigma)^{-2}] \\
&+\; b_*\frac{1}{N^2}(\mathrm{tr}[(I+\xi_{1,1}(z)\,\Sigma)^{-1}\Sigma])^2c^{-1}\widehat{\xi}_{2,1}(z) \\
&-2c^{-1}b_*\frac{1}{N^2}\,\mathrm{tr}\,E[(I+\xi_{1,1}(z)\,\Sigma)^{-1}\Sigma]\widehat{\xi}_{1,1}(z)\xi_{2,1}(z)\,\mathrm{tr}[(I+\xi_{1,1}(z)\,\Sigma)^{-2}\Sigma^2]
\end{aligned} \tag{143}$$

In particular, if $T^{-1}\,\mathrm{tr}(\Psi_{1,2}\Psi_{2,1}) \to 0$, *we get by* (139) *that*

$$\begin{aligned}
\Delta(z) \;&=\; \frac{\mathrm{tr}(\Psi_{2,2})}{P}\,b_*\widehat{\xi}_{1,1}(z)\frac{1}{N}\,\mathrm{tr}[\Sigma^2(I+\xi_{1,1}(z)\,\Sigma)^{-2}] \\
&=\; \frac{\mathrm{tr}(\Psi_{2,2})}{P}\,b_*(\xi_{1,1}(z)+z\xi_{1,1}'(z))\frac{\mathrm{tr}(\Sigma^2(I+\xi_{1,1}(z)\,\Sigma)^{-2})}{\mathrm{tr}(\Sigma(I+\xi_{1,1}(z)\Sigma)^{-2})} \\
&=\; -\frac{\mathrm{tr}(\Psi_{2,2})}{P}\,b_*c_1\nu'(z;c_1,q)\frac{\mathrm{tr}(\Sigma^2(I+\xi_{1,1}(z)\,\Sigma)^{-2})}{\mathrm{tr}(\Sigma(I+\xi_{1,1}(z)\Sigma)^{-2})}
\end{aligned} \tag{144}$$

Proof of Lemma 14. We will be using the notation

$$\hat{\Psi}_T \;=\; \frac{1}{NT}\sum_{t=1}^{T} S'_{t,1}S_t \;\in \mathbb{R}^{P_1\times P}\,. \tag{145}$$

and we decompose

$$\hat{\Psi}_T \;=\; \hat{\Psi}_{T,1} \;+\; \hat{\Psi}_{T,2}\,. \tag{146}$$

By the same argument as in the proof of Proposition 14, we can work directly with expectation because $\mathrm{tr}(\Psi_{1,1}E[\hat{\beta}\hat{\beta}'])-\mathrm{tr}(\Psi_{1,1}\hat{\beta}\hat{\beta}')$ converges to zero almost surely. We thus have

$$\begin{aligned}
&\mathcal{L}/\,\mathrm{tr}(\Sigma) \;=\; \mathrm{tr}(\Psi_{1,1}E[\hat{\beta}\hat{\beta}'])\\
&=\; \mathrm{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T\beta \;+\; q_T)(\hat{\Psi}_T\beta \;+\; q_T)'(zI+\hat{\Psi}_{T,1})^{-1}])\\
&=\; \mathrm{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T})^{-1}(\hat{\Psi}_T\beta \;+\; q_T)(\beta'\hat{\Psi}_T \;+\; q'_T)(zI+\hat{\Psi}_{T,1})^{-1}])\\
&=\; \mathrm{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T\beta\beta'\hat{\Psi}_T \;+\; q_Tq'_T)(zI+\hat{\Psi}_{T,1})^{-1}])\\
&=\; \mathrm{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T\beta\beta'\hat{\Psi}_T \;+\; q_Tq'_T)(zI+\hat{\Psi}_{T,1})^{-1}])
\end{aligned} \tag{147}$$

By (127),

$$E[q_Tq'_T|S] \;=\; \frac{1}{NT}\hat{\Psi}_{T,1} \tag{148}$$

and therefore

$$\mathrm{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T,1})^{-1}q_Tq'_T(zI+\hat{\Psi}_{T,1})^{-1}]) \;\to\; \xi_{1,1}(z)+z\xi'_{1,1}(z) \tag{149}$$

in probability by (131) and (135). Thus, compared to (128), we only need to take care of the $\hat{\Psi}_T\beta\beta'\hat{\Psi}_T$ term. We have

$$\begin{aligned}
&\mathrm{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T\beta\beta'\hat{\Psi}_T)(zI+\hat{\Psi}_{T,1})^{-1}])\\
&=\; \mathrm{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_{T,1}\beta_1\beta'_1\hat{\Psi}_{T,1})(zI+\hat{\Psi}_{T,1})^{-1}])\\
&+\; \mathrm{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_{T,2}\beta_2\beta'_2\hat{\Psi}'_{T,2})(zI+\hat{\Psi}_{T,1})^{-1}])\\
&=\; P^{-1}b_*\,\mathrm{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}^2_{T,1}(zI+\hat{\Psi}_{T,1})^{-1}])\\
&+\; P^{-1}b_*\,\mathrm{tr}\,E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\hat{\Psi}'_{T,2}(zI+\hat{\Psi}_{T,1})^{-1}]
\end{aligned} \tag{150}$$

By (130) and (136), we have

$$\begin{aligned}
&P^{-1}P_1b_*P_1^{-1}\,\mathrm{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}^2_{T,1}(zI+\hat{\Psi}_{T,1})^{-1}])\\
&\to\; c^{-1}c_1b_*(\psi_{*,1}(q)-2zc_1^{-1}\xi_{1,1}(z)-z^2c_1^{-1}\xi'_{1,1}(z))
\end{aligned} \tag{151}$$

Thus, it remains to deal with the second term im of the second term in (150).

Recall that

$$\xi_{2,1}(z;cq;q) = \lim_{T\to\infty} (NT)^{-1} \operatorname{tr} E[(zI + \hat{\Psi}_{1,1})^{-1}\Psi_{1,2}\Psi'_{1,2}] \;\geq\; 0 \tag{152}$$
$$\widehat{\xi}_{2,1}(z;cq;q) = \lim_{T\to\infty} (NT)^{-1} \operatorname{tr} E[(zI + \hat{\Psi}_{1,1})^{-1}\Psi_{1,1}(zI + \hat{\Psi}_{1,1})^{-1}\Psi_{1,2}\Psi'_{1,2}] \;\geq\; 0\,.$$

Lemma 15

$$\begin{aligned}
&P^{-1}b_* \operatorname{tr} E[\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\hat{\Psi}'_{T,2}(zI + \hat{\Psi}_{T,1})^{-1}] \;\to \\
&\frac{\operatorname{tr}(\Psi_{2,2})}{P}\, b_* \frac{1}{N}\widehat{\xi}_{1,1}(z) \operatorname{tr}[\Sigma^2(I + \xi_{1,1}(z)\,\Sigma)^{-2}] \\
&+ b_* \frac{1}{N^2}(\operatorname{tr}[(I + \xi_{1,1}(z)\,\Sigma)^{-1}\Sigma])^2 c^{-1}\widehat{\xi}_{2,1}(z) \\
&- 2b_* \frac{1}{N^2}\widehat{\xi}_{1,1}(z) \operatorname{tr}((I + \xi_{1,1}(z)\,\Sigma)^{-1}\Sigma) c^{-1}\xi_{2,1}(z) \operatorname{tr}[(I + \xi_{1,1}(z)\,\Sigma)^{-2}\Sigma^2]
\end{aligned} \tag{153}$$

Proof of Lemma 15. We have

$$\begin{aligned}
&P^{-1}b_* \operatorname{tr} E[\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\hat{\Psi}'_{T,2}(zI + \hat{\Psi}_{T,1})^{-1}] \\
&=\; P^{-1}b_* \operatorname{tr} E[\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\hat{\Psi}'_{T,2}(zI + \hat{\Psi}_{T,1})^{-1}] \\
&=\; P^{-1}b_* \frac{1}{N^2T^2} \operatorname{tr} E[\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1} \sum_{t_1,t_2} S'_{t_1,1}S_{t_1,2}S'_{t_2,2}S_{t_2,1}(zI + \hat{\Psi}_{T,1})^{-1}] \\
&=\; P^{-1}b_* \frac{1}{N^2T^2} \operatorname{tr} E[\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1} \Big(T S'_{t_1,1}S_{t_1,2}S'_{t_1,2}S_{t_1,1} \\
&+ T(T-1)\, S'_{t_1,1}S_{t_1,2}S'_{t_2,2}S_{t_2,1} \Big)(zI + \hat{\Psi}_{T,1})^{-1}] \\
&=\; Term1 \;+\; Term2\,.
\end{aligned} \tag{154}$$

Here,

$$Term1 \;=\; P^{-1}b_* \frac{1}{N^2T} \operatorname{tr} E[\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1}S'_{t_1,1}S_{t_1,2}S'_{t_1,2}S_{t_1,1}(zI + \hat{\Psi}_{T,1})^{-1}] \tag{155}$$

and

$$Term2 \;\approx\; P^{-1}b_* \frac{1}{N^2} \operatorname{tr} E[\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1}S'_{t_1,1}S_{t_1,2}S'_{t_2,2}S_{t_2,1}(zI + \hat{\Psi}_{T,1})^{-1}] \tag{156}$$

Dealing with Term1 in (155) Using the Woodbury matrix identity (118), we get

$$\begin{aligned}
&(zI+\hat{\Psi}_{T,1})^{-1}S'_{t_1,1} \;=\; (zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t_1,1}\\
&-\;\frac{1}{NT}(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t,1}(I+\frac{1}{NT}S_{t,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t,1})^{-1}S_{t,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t_1,1}\\
&=\;(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t_1,1}(I+\frac{1}{NT}S_{t,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t,1})^{-1}\\
&\to\;(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}
\end{aligned} \tag{157}$$

and therefore

$$\begin{aligned}
&Term1\\
&=\;P^{-1}b_*\frac{1}{N^2T}\,\mathrm{tr}\,E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}\\
&\times S_{t_1,2}S'_{t_1,2}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}]\\
&=\;P^{-1}b_*\frac{1}{N^2T}\,\mathrm{tr}\,E[S_{t_1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t_1,1}\\
&\times(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2}S'_{t_1,2}(I+\xi_{1,1}(z)\,\Sigma)^{-1}]
\end{aligned} \tag{158}$$

Now, by the same argument as in the proof of Lemma 9 and Lemma 10, we have

$$\frac{1}{NT}S_{t_1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t_1,1}\;\to\;\frac{1}{NT}E[\mathrm{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1})]\Sigma \tag{159}$$

and Lemma 16 below implies that

$$\begin{aligned}
&Term1\\
&\to P^{-1}b_*\frac{1}{N}\widehat{\xi}_{1,1}(z)\,\mathrm{tr}\,E[\Sigma(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2}S'_{t_1,2}(I+\xi_{1,1}(z)\,\Sigma)^{-1}]\\
&=\;\frac{\mathrm{tr}(\Psi_{2,2})}{P}\,b_*\frac{1}{N}\widehat{\xi}_{1,1}(z)\,\mathrm{tr}\,E[\Sigma(I+\xi_{1,1}(z)\,\Sigma)^{-1}\Sigma(I+\xi_{1,1}(z)\,\Sigma)^{-1}]
\end{aligned} \tag{160}$$

where we have used that

$$E[S_{t_1,2}S'_{t_1,2}]\;=\;\mathrm{tr}(\Psi_{2,2})\Sigma\,. \tag{161}$$

Lemma 16 *We have*

$$\frac{1}{NT}E[\mathrm{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1})]\;\to\;\widehat{\xi}_{1,1}(z)\;=\;\frac{(\xi_{1,1}(z)+z\xi'_{1,1}(z))c_1^{-1}\xi'_{1,1}(z)}{-m(-z;c_1)+zm'(-z;c_1)} \tag{162}$$

Proof of Lemma 16. We have

$$\frac{1}{NT}\,\mathrm{tr}\,E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}]\;\to\;\xi_{1,1}(z)$$

and therefore

$$\frac{1}{NT}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_{T,1,t})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}] \;=\; \frac{1}{NT}\,\mathrm{tr}\,E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-2}] \;\to\; -\xi'_{1,1}(z)$$

By the same argument as above,

$$\begin{aligned}
&\frac{1}{NT}S_{t_1,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t_1,1}\\
&\approx\; \frac{1}{NT}\,\mathrm{tr}\,E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}]\Sigma\\
&\approx\; \widehat{\xi}_{1,1}(z)\Sigma
\end{aligned} \tag{163}$$

Hence, by (157) and (163),

$$\begin{aligned}
\xi_{1,1}(z) \;&\approx\; \frac{1}{NT}\,\mathrm{tr}\,E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}]\\
&=\; \frac{1}{NT}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_{T,1})(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}]\\
&\approx\; -z\xi'_{1,1}(z) \;+\; \frac{1}{NT}\,\mathrm{tr}\,E[\hat{\Psi}_{T,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}]\\
&=\; -z\xi'_{1,1}(z) \;+\; \frac{1}{N^2T^2}\sum_t \mathrm{tr}\,E[S'_{t,1}S_{t,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}]\\
&=\; -z\xi'_{1,1}(z) \;+\; \frac{1}{N^2T}\,\mathrm{tr}\,E[S'_{t,1}S_{t,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}]\\
&=\; -z\xi'_{1,1}(z) \;+\; \frac{1}{N^2T}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_{T,1})^{-1}S'_{t,1}S_{t,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}]\\
&=\; -z\xi'_{1,1}(z)\\
&+\; \frac{1}{N^2T}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}]\\
&=\; -z\xi'_{1,1}(z)\\
&+\; \frac{1}{N^2T}\,\mathrm{tr}\,E[(I+\xi_{1,1}(z)\,\Sigma)^{-2}S_{t_1,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t_1,1}]\\
&\approx\; -z\xi'_{1,1}(z) \;+\; \frac{1}{N}\,\mathrm{tr}\,E[(I+\xi_{1,1}(z)\,\Sigma)^{-2}\Sigma]\widehat{\xi}_{1,1}(z)
\end{aligned} \tag{164}$$

By (124),

$$-m(-z;c_1)+zm'(-z;c_1) \;=\; c_1^{-1}N^{-1}\,\mathrm{tr}((I+\xi_{1,1}(z)\Sigma)^{-2}\Sigma)\xi'_{1,1}(z) \tag{165}$$

so that

$$\xi_{1,1}(z) \;=\; -z\xi'_{1,1}(z) \;+\; \frac{-m(-z;c_1)+zm'(-z;c_1)}{c_1^{-1}\xi'_{1,1}(z)}\widehat{\xi}_{1,1}(z) \tag{166}$$

and the claim follows. The proof of Lemma 16 is complete. □

Dealing with Term2 in (156) Now, by (157),

$$
\begin{aligned}
Term2 \;&\approx\; P^{-1}b_*\frac{1}{N^2}\operatorname{tr} E[\Psi_{1,1}(zI+\hat\Psi_{T,1})^{-1}S'_{t_1,1}S_{t_1,2}S'_{t_2,2}S_{t_2,1}(zI+\hat\Psi_{T,1})^{-1}]\\
&=\; P^{-1}b_*\frac{1}{N^2}\operatorname{tr} E[S_{t_2,1}(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,1}(zI+\hat\Psi_{T,1})^{-1}S'_{t_1,1}S_{t_1,2}S'_{t_2,2}]\\
&\approx\; P^{-1}b_*\frac{1}{N^2}\operatorname{tr} E[(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1}(zI+\hat\Psi_{T,1,t_2})^{-1}\\
&\times\Psi_{1,1}(zI+\hat\Psi_{T,1,t_1})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2}S'_{t_2,2}]\\
&\approx\; P^{-1}b_*\frac{1}{N^2}\operatorname{tr} E[(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1}\\
&\times\left((zI+\hat\Psi_{T,1,t_1,t_2})^{-1}-\frac{1}{NT}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\right)\\
&\times\Psi_{1,1}\left((zI+\hat\Psi_{T,1,t_1,t_2})^{-1}-\frac{1}{NT}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S'_{t_2,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\right)\\
&S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2}S'_{t_2,2}]\\
&=\; P^{-1}b_*\frac{1}{N^2}\operatorname{tr} E[(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1}\\
&\times\left((zI+\hat\Psi_{T,1,t_1,t_2})^{-1}-\frac{1}{NT}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\right)\\
&\times\Psi_{1,1}\left((zI+\hat\Psi_{T,1,t_1,t_2})^{-1}-\frac{1}{NT}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S'_{t_2,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\right)\\
&S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2}S'_{t_2,2}]\\
&=\; P^{-1}b_*\frac{1}{N^2}\operatorname{tr} E[(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1}\\
&\times\Bigg((zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\\
&-\frac{1}{NT}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\\
&-(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\frac{1}{NT}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S'_{t_2,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\\
&+\frac{1}{NT}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\\
&\times\frac{1}{NT}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S'_{t_2,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\Bigg)\\
&S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2}S'_{t_2,2}]\\
&=\; Term21+Term22+Term23+Term24\,.
\end{aligned}
$$

(167)

We have

$$E[P^{-1}S_{t_2,1}QS'_{t_2,2}] \;=\; P^{-1}E[\mathrm{tr}[Q\Psi_{2,1}]]\Sigma \tag{168}$$

and therefore

$$\begin{aligned}
Term21 \;&=\; P^{-1}b_*\frac{1}{N^2}\,\mathrm{tr}\,E[(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1} \\
&\times\left((zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2}S'_{t_2,2}] \\
&=\; P^{-1}b_*\frac{1}{N^2}\,\mathrm{tr}\,E[(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1}QS'_{t_2,2}] \\
&=\; b_*\frac{1}{N^2}\,\mathrm{tr}[(I+\xi_{1,1}(z)\,\Sigma)^{-1}\Sigma]P^{-1}E[\mathrm{tr}(Q\Psi_{2,1})]
\end{aligned} \tag{169}$$

where

$$Q \;=\; \left((zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2}$$

and hence

$$\begin{aligned}
&P^{-1}E[\mathrm{tr}\,Q\Psi_{2,1}] \;=\; P^{-1}\,\mathrm{tr}\,E[\left((zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2}\Psi_{2,1}] \\
&=\; P^{-1}\,\mathrm{tr}\,E[S_{t_1,2}\Psi_{2,1}\left((zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}] \\
&\to\; P^{-1}\,\mathrm{tr}\,E[\Psi_{1,2}\Psi_{2,1}\left((zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)]\,\mathrm{tr}(\Sigma(I+\xi_{1,1}(z)\,\Sigma)^{-1})
\end{aligned} \tag{170}$$

Thus,

$$\begin{aligned}
&Term21 \\
&\to b_*\frac{1}{N^2}(\mathrm{tr}[(I+\xi_{1,1}(z)\,\Sigma)^{-1}\Sigma])^2 \\
&\times P^{-1}\,\mathrm{tr}\,E[\Psi_{1,2}\Psi_{2,1}\left((zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)]\,.
\end{aligned} \tag{171}$$

Proceeding to the next term, we get

$$
\begin{aligned}
&Term22 \\
&= P^{-1}b_*\frac{1}{N^2}\operatorname{tr} E[(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1} \\
&\times\left(-\frac{1}{NT}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right) \\
&S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2}S'_{t_2,2}]
\end{aligned}
\tag{172}
$$

We have

$$
\begin{aligned}
&\frac{1}{NT}S_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S'_{t_1,1} \\
&\to\ \frac{1}{NT}\operatorname{tr} E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}]\Sigma \\
&=\ \widehat{\xi}_{1,1}(z)\,\Sigma
\end{aligned}
\tag{173}
$$

Therefore,

$$
\begin{aligned}
&Term22 \\
&= P^{-1}b_*\frac{1}{N^2}\operatorname{tr} E[(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1} \\
&\times\left(-(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}\widehat{\xi}_{1,1}(z)\right)\Sigma(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2}S'_{t_2,2}] \\
&= -P^{-1}b_*\frac{1}{N^2}\operatorname{tr} E[(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1}Z_TS'_{t_2,2}]
\end{aligned}
\tag{174}
$$

where we have defined

$$
Z_T\ =\ (zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}\widehat{\xi}_{1,1}(z)\Sigma(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2} \tag{175}
$$

Using that

$$
\frac{1}{NT}S_{t_2,1}Z_TS'_{t_2,1}\to\ \frac{1}{NT}\operatorname{tr} E[\Psi_{2,1}Z_T]\Sigma \tag{176}
$$

for any bounded Z_T,[52] we get

$$\begin{aligned}
&Term22\\
&= -P^{-1}b_*\frac{1}{N^2}\operatorname{tr}((I+\xi_{1,1}(z)\,\Sigma)^{-1}\Sigma)\\
&\times \operatorname{tr} E\left[\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}\widehat{\xi}_{1,1}(z)\Sigma(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2}\right] \qquad (177)\\
&= -P^{-1}b_*\frac{1}{N^2}\operatorname{tr}((I+\xi_{1,1}(z)\,\Sigma)^{-1}\Sigma)\\
&\times \operatorname{tr} E\left[S_{t_1,2}\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}\widehat{\xi}_{1,1}(z)\Sigma(I+\xi_{1,1}(z)\,\Sigma)^{-1}\right]
\end{aligned}$$

Using that

$$\frac{1}{NT}S_{t_1,2}Z_TS'_{t_1,1}\to\ \frac{1}{NT}\operatorname{tr} E[\Psi_{1,2}Z_T]\Sigma \qquad (178)$$

with

$$Z_T\ =\ \Psi_{2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\,, \qquad (179)$$

we get

$$\begin{aligned}
&Term22\\
&= -P^{-1}b_*\frac{1}{N^2}\operatorname{tr}((I+\xi_{1,1}(z)\,\Sigma)^{-1}\Sigma)\\
&\times \operatorname{tr} E\left[\operatorname{tr}(\Psi_{1,2}\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1})(I+\xi_{1,1}(z)\,\Sigma)^{-1}\widehat{\xi}_{1,1}(z)\Sigma(I+\xi_{1,1}(z)\,\Sigma)^{-1}\Sigma\right] \qquad (180)\\
&= -b_*\frac{1}{N^2}\widehat{\xi}_{1,1}(z)\operatorname{tr}((I+\xi_{1,1}(z)\,\Sigma)^{-1}\Sigma)c^{-1}\xi_{2,1}(z)\operatorname{tr}[(I+\xi_{1,1}(z)\,\Sigma)^{-2}\Sigma^2]
\end{aligned}$$

[52] While this argument formally works for bounded Z_T, a standard truncation argument allows one to deal with $Z_T\in L_2$.

Proceeding to the next term, we get[53]

$$
\begin{aligned}
&Term23 \;\to \\
&-P^{-1}b_*\frac{1}{N^2}\operatorname{tr}E[(I+\xi_{1,1}(z)\,\Sigma)^{-1}\underbrace{S_{t_2,1}\left((zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\frac{1}{NT}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)S'_{t_2,1}}_{\to\widehat{\xi}_{1,1}(z)\Sigma} \\
&\times(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2}S'_{t_2,2}] \\
&\to -P^{-1}b_*\frac{1}{N^2}\widehat{\xi}_{1,1}(z)\operatorname{tr}E[(I+\xi_{1,1}(z)\,\Sigma)^{-1}\Sigma(I+\xi_{1,1}(z)\,\Sigma)^{-1} \\
&\underbrace{S_{t_2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2}S'_{t_2,2}}_{\to\ \operatorname{tr}(\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2})\Sigma}] \\
&\to\ -P^{-1}b_*\frac{1}{N^2}\widehat{\xi}_{1,1}(z)\operatorname{tr}E[(I+\xi_{1,1}(z)\,\Sigma)^{-1}\Sigma(I+\xi_{1,1}(z)\,\Sigma)^{-1}\Sigma\operatorname{tr}(\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1} \\
&S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2})] \\
&=\ -P^{-1}b_*\frac{1}{N^2}\widehat{\xi}_{1,1}(z)\operatorname{tr}(\Sigma^2(I+\xi_{1,1}(z)\,\Sigma)^{-2})\operatorname{tr}E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1} \\
&S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2}] \\
&=\ -P^{-1}b_*\frac{1}{N^2}\widehat{\xi}_{1,1}(z)\operatorname{tr}(\Sigma^2(I+\xi_{1,1}(z)\,\Sigma)^{-2})\operatorname{tr}E[\underbrace{S_{t_1,2}\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S'_{t_1,1}}_{\to\operatorname{tr}(\Psi_{1,2}\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1})\Sigma}(I+\xi_{1,1}(z)\,\Sigma)^{-1}] \\
&=\ -P^{-1}b_*\frac{1}{N^2}\widehat{\xi}_{1,1}(z)\operatorname{tr}(\Sigma^2(I+\xi_{1,1}(z)\,\Sigma)^{-2})\operatorname{tr}(\Sigma(I+\xi_{1,1}(z)\,\Sigma)^{-1})\operatorname{tr}E[\Psi_{1,2}\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}] \\
&=\ Term22\,.
\end{aligned}
\tag{181}
$$

[53] Convergence we use is in L_2 and passing to the limit is justified because $E[X_TY_T]\to E[XY]$ when $X_T\to X,\ Y_T\to Y$ in L_2.

Finally,

$$
\begin{aligned}
Term24 \;&=\; P^{-1}b_*\frac{1}{N^2}\operatorname{tr} E[(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1}\\
&\times\left(\frac{1}{NT}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\right.\\
&\left.\times\frac{1}{NT}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S'_{t_2,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\right)\\
&S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2}S'_{t_2,2}]\\
&=\; P^{-1}b_*\frac{1}{N^2}\operatorname{tr} E[S'_{t_2,2}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1}\\
&\times\left(\frac{1}{NT}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\right.\\
&\left.\times\frac{1}{NT}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S'_{t_2,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_2,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\right)\\
&S'_{t_1,1}(I+\xi_{1,1}(z)\,\Sigma)^{-1}S_{t_1,2}]
\end{aligned}
\tag{182}
$$

We will need the following

Lemma 17 *Suppose that $S_t=\Sigma^{1/2}X_t\Psi^{1/2}$. Then, under the ecomposition $\Psi^{1/2}=(Q_1,Q_2)$, we have*

$$
\begin{aligned}
&E[S'_{t,2}S_{t,1}ZS'_{t,1}S_{t,1}]\\
&=\; ((\operatorname{tr}\Sigma)^2-\operatorname{tr}(\Sigma^2))\Psi_{2,1}Z\Psi_{1,1}\;+\;(\operatorname{tr}\Sigma^2)\Psi_{2,1}(Z+Z')\Psi_{1,1}\\
&+\operatorname{tr}(\Sigma^2)((\kappa-2)Q'_2\operatorname{diag}(Q_1ZQ'_1)Q_1+\operatorname{tr}(Z\Psi_{1,1})\Psi_{2,1})
\end{aligned}
\tag{183}
$$

for any matrix Z.

Proof of Lemma 17. By linearity, it suffices to prove the formula for a rank-one matrix $A=\beta\gamma'$.[54] Then, $S_t=\Sigma^{1/2}X_t\Psi^{1/2}$ and we will decompose $\Psi^{1/2}$ into (Q_1,Q_2), so that $S_{t,k}=\Sigma^{1/2}X_tQ_k$. Then,

$$
E[S'_{t,2}S_{t,1}\beta\gamma'S'_{t,1}S_{t,1}]\;=\;E[Q'_2\,X'_t\Sigma X_tQ_1\beta\gamma'Q'_1X'_t\Sigma X_tQ_1]
\tag{184}
$$

Define $\tilde\beta=Q_1\beta$ and let us rotate X so that Σ is a diagonal matrix, D. The rotated X is denoted by $\tilde X$. Then,

$$
E[\tilde X'_tD\tilde X_t\tilde\beta\tilde\gamma'\tilde X'_tD\tilde X_t]_{k_1,k_2}\;=\;E[\sum_{i_1,i_2,l_1,l_2}D_{i_1}D_{i_2}\,\tilde X_{i_1,k_1}\tilde X_{i_1,l_1}\tilde\beta_{l_1}\tilde\gamma_{l_2}\tilde X_{i_2,l_2}\tilde X_{i_2,k_2}]\,.
$$

[54] Note that Corollary 7 is not applicable here.

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

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

At the same time,

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

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

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

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

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

Thus,

$$\begin{aligned}
&E[\tilde{X}_t' D \tilde{X}_t \tilde{\beta} \tilde{\gamma}' \tilde{X}_t' D \tilde{X}_t]_{k_1,k_2} \\
&= \; ((\operatorname{tr} \Sigma)^2 - \operatorname{tr}(\Sigma^2)) \tilde{\beta}_{k_1} \tilde{\gamma}_{k_2} \; + \operatorname{tr}(\Sigma^2)(\tilde{\beta}_{k_1} \tilde{\gamma}_{k_2} + \tilde{\beta}_{k_2} \tilde{\gamma}_{k_1})(1 - \delta_{k_1,k_2}) + \; \operatorname{tr}(\Sigma^2)(\kappa \tilde{\beta}_{k_1} \tilde{\gamma}_{k_1} + \tilde{\beta}' \tilde{\gamma}) \delta_{k_1,k_2} \\
&= \; ((\operatorname{tr} \Sigma)^2 - \operatorname{tr}(\Sigma^2)) \tilde{\beta}_{k_1} \tilde{\gamma}_{k_2} \; + \operatorname{tr}(\Sigma^2)(\tilde{\beta}_{k_1} \tilde{\gamma}_{k_2} + \tilde{\beta}_{k_2} \tilde{\gamma}_{k_1}) \\
&+ \operatorname{tr}(\Sigma^2)((\kappa - 2) \tilde{\beta}_{k_1} \tilde{\gamma}_{k_1} + \tilde{\beta}' \tilde{\gamma}) \delta_{k_1,k_2}
\end{aligned} \tag{185}$$

Thus, by formula (184), we get

$$\begin{aligned}
&E[S_{t,2}' S_{t,1} \beta \gamma' S_{t,1}' S_{t,1}] \; = \; ((\operatorname{tr} \Sigma)^2 - \operatorname{tr}(\Sigma^2)) Q_2' Q_1 \beta \gamma' Q_1' Q_1 \\
&+ (\operatorname{tr} \Sigma^2)(Q_2' Q_1 \beta \gamma' Q_1' Q_1 + Q_1' Q_1 \gamma \beta' Q_1' Q_2) \\
&+ \operatorname{tr}(\Sigma^2)((\kappa - 2) Q_2' \operatorname{diag}(\tilde{\beta}_{k_1} \tilde{\gamma}_{k_1}) Q_1 + (\tilde{\beta}' \tilde{\gamma}) Q_2' Q_1)
\end{aligned} \tag{186}$$

whereas $\tilde{\beta}' \tilde{\gamma} = \beta' Q_1' Q_1 \gamma$. Now,

$$Q_1 \; = \; \begin{pmatrix} Q_{1,1} \\ Q_{2,1} \end{pmatrix}, \; Q_2 \; = \; \begin{pmatrix} Q_{1,2} \\ Q_{2,2} \end{pmatrix}, \; \Psi^{1/2} \; = \; \begin{pmatrix} Q_{1,1} & Q_{1,2} \\ Q_{2,1} & Q_{2,2} \end{pmatrix}$$

Thus,

$$\Psi \; = \; \begin{pmatrix} Q_1' Q_1 & Q_1' Q_2 \\ Q_2' Q_1 & Q_2' Q_2 \end{pmatrix} \; = \; \begin{pmatrix} \Psi_{1,1} & \Psi_{1,2} \\ \Psi_{2,1} & \Psi_{2,2} \end{pmatrix} \tag{187}$$

and hence we get the required.

□

We can thus apply Lemma 17 and integrate over S_{t_2}[55]. We first need to re-define $\hat{S}_t = (I + \xi_{1,1}(z)\,\Sigma)^{-1/2} S_t$ and then apply Lemma 17 to these signals. This gives

$$\begin{aligned}
&E[S'_{t_2,2}(I + \xi_{1,1}(z)\,\Sigma)^{-1} S_{t_2,1} \\
&\times \left(\frac{1}{NT}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S'_{t_1,1}(I + \xi_{1,1}(z)\,\Sigma)^{-1} S_{t_1,1}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1} \right. \\
&\left. \times \frac{1}{NT}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S'_{t_2,1}(I + \xi_{1,1}(z)\,\Sigma)^{-1} S_{t_2,1} | S_{t_1}] \right. \\
&= E[\hat{S}'_{t_2,2}\hat{S}_{t_2,1} \\
&\left(\frac{1}{NT}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \hat{S}'_{t_1,1}\hat{S}_{t_1,1}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1} \right. \\
&\left. \times \frac{1}{NT}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \right) \hat{S}'_{t_2,1}\hat{S}_{t_2,1} | S_{t_1}] \\
&= ((\mathrm{tr}[(I + \xi_{1,1}(z)\,\Sigma)^{-1}\Sigma])^2 + \mathrm{tr}((I + \xi_{1,1}(z)\,\Sigma)^{-2}\Sigma^2))\Psi_{2,1} Z \Psi_{1,1} \\
&+ \mathrm{tr}((I + \xi_{1,1}(z)\,\Sigma)^{-2}\Sigma^2))\,\mathrm{tr}(Z\Psi_{1,1})\Psi_{2,1} \;+\; O(|\kappa - 2|)\, Q'_2\,\mathrm{diag}(Q_1 Z Q'_1) Q_1\,,
\end{aligned} \tag{188}$$

where

$$Z \;=\; \left(\frac{1}{NT}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \hat{S}'_{t_1,1}\hat{S}_{t_1,1}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1} \frac{1}{NT}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \right)$$

Now, everything reduces to computing two expectations:

$$Expectation1 \;=\; P^{-1}\,\mathrm{tr}\,E[\Psi_{2,1} Z \Psi_{1,1}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \hat{S}'_{t_1,1}\hat{S}_{t_1,2}] \tag{189}$$

and

$$Expectation2 \;=\; P^{-1}\,\mathrm{tr}\,E[\Psi_{2,1}\,\mathrm{tr}(Z\Psi_{1,1})(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \hat{S}'_{t_1,1}\hat{S}_{t_1,2}] \tag{190}$$

and

$$Expectation3 \;=\; P^{-1}\,\mathrm{tr}\,E[Q'_2\,\mathrm{diag}(Q_1 Z Q'_1) Q_1 (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \hat{S}'_{t_1,1}\hat{S}_{t_1,2}] \tag{191}$$

For $Expectation2$, we have

$$\begin{aligned}
&Expectation2 \;=\; P^{-1}\,\mathrm{tr}\,E[\Psi_{2,1}\,\mathrm{tr}(Z\Psi_{1,1})(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \hat{S}'_{t_1,1}\hat{S}_{t_1,2}] \\
&= P^{-1}\,\mathrm{tr}\,E[\hat{S}_{t_1,2}\Psi_{2,1}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \hat{S}'_{t_1,1}\,\mathrm{tr}(Z\Psi_{1,1})]
\end{aligned} \tag{192}$$

We will now show that $Expectation2$ converges to zero. Since the proof can be done element-

[55] Using the fact that S_{t_2} and S_{t_1} are independent.

by-element, it suffices to consider the case $N = 1$. In this case, we know that the quantities

$$\frac{1}{NT}\hat{S}_{t_1,2}\Psi_{2,1}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\hat{S}'_{t_1,1}$$

and

$$\begin{aligned}&\frac{1}{NT}\operatorname{tr}\left((zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\hat{S}'_{t_1,1}\hat{S}_{t_1,1}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)\\ &= \frac{1}{NT}\operatorname{tr}\left(\hat{S}_{t_1,1}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\hat{S}'_{t_1,1}\right)\end{aligned} \tag{193}$$

both converge to finite numbers in L_2. Thus, when multiplied by P^{-1}, the expectation of the product of these two quantities converges to zero. Thus, *Expectation*2 converges to zero. To compute *Expectation*1, we use

$$\begin{aligned}&Expectation1 = P^{-1}\operatorname{tr}E[\Psi_{2,1}Z\Psi_{1,1}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\hat{S}'_{t_1,1}\hat{S}_{t_1,2}]\\ &= P^{-1}\operatorname{tr}E[\hat{S}'_{t_1,2}\hat{S}_{t_1,1}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}Z'\Psi_{1,2}]\\ &= P^{-1}\operatorname{tr}E[\hat{S}'_{t_1,2}\hat{S}_{t_1,1}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\\ &\times\left((zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\frac{1}{NT}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\hat{S}'_{t_1,1}\hat{S}_{t_1,1}\frac{1}{NT}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)\Psi_{1,2}]\end{aligned} \tag{194}$$

By Lemma 17 and get

$$\begin{aligned}&E[\hat{S}'_{t_1,2}\hat{S}_{t_1,1}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\\ &\times(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\frac{1}{NT}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\hat{S}'_{t_1,1}\hat{S}_{t_1,1}]\\ &= ((\operatorname{tr}[(I + \xi_{1,1}(z)\,\Sigma)^{-1}\Sigma])^2 + \operatorname{tr}((I + \xi_{1,1}(z)\,\Sigma)^{-2}\Sigma^2))\Psi_{2,1}\hat{Z}\Psi_{1,1}\\ &+ \operatorname{tr}((I + \xi_{1,1}(z)\,\Sigma)^{-2}\Sigma^2))\operatorname{tr}(\hat{Z}\Psi_{1,1})\Psi_{2,1}\end{aligned} \tag{195}$$

where

$$\hat{Z} = (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\frac{1}{NT}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \tag{196}$$

Therefore,

$$\begin{aligned}&Expectation1 = P^{-1}\operatorname{tr}E[\Bigg(((\operatorname{tr}[(I + \xi_{1,1}(z)\,\Sigma)^{-1}\Sigma])^2 + \operatorname{tr}((I + \xi_{1,1}(z)\,\Sigma)^{-2}\Sigma^2))\Psi_{2,1}\hat{Z}\Psi_{1,1}\\ &+ \operatorname{tr}((I + \xi_{1,1}(z)\,\Sigma)^{-2}\Sigma^2))\operatorname{tr}(\hat{Z}\Psi_{1,1})\Psi_{2,1}\Bigg)\frac{1}{NT}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,2}]\end{aligned} \tag{197}$$

First, $\mathrm{tr}(\hat{Z}\Psi_{1,1})$ converges to a finite, non-random number, and hence the second term in this expression converges to zero. Second, the first term also converges to zero by a similar argument, due to the $P^{-1}(NT)^{-2}$ factor.

A slightly more tedious argument implies an analogous result for *Expectation*3.

Thus, *Term*24 converges to zero. The proof of Lemma 15 is complete.

□

The proof of Lemma 14 is complete.

□

I Computing Second Moment $\mathcal{V}$ of the Efficient Portfolio

We will need the following

Lemma 18 *Suppose that* $S_t = \Sigma^{1/2} X_t \Psi^{1/2}$. *Then, under the decomposition* $\Psi^{1/2} = (Q_1, Q_2)$,

$$\begin{aligned} &E[S'_{t,2} S_{t,1} Z S'_{t,1} S_{t,2}] \\ &= \ ((\mathrm{tr}\,\Sigma)^2 - \mathrm{tr}(\Sigma^2))\Psi_{2,1} Z \Psi_{1,2} \ + \ (\mathrm{tr}\,\Sigma^2)\Psi_{2,1}(Z + Z')\Psi_{1,2} \\ &+ \mathrm{tr}(\Sigma^2)((\kappa - 2)Q'_2\,\mathrm{diag}(Q_1 Z Q'_1)Q_2 + \mathrm{tr}(Z\Psi_{1,1})\Psi_{2,2}) \end{aligned} \tag{198}$$

for any matrix Z.

Proof of Lemma 18. By linearity, it suffices to prove the formula for a rank-one matrix $A = \beta\gamma'$.[56] Then, $S_t = \Sigma^{1/2} X_t \Psi^{1/2}$ and we will decompose $\Psi^{1/2}$ into (Q_1, Q_2), so that $S_{t,k} = \Sigma^{1/2} X_t Q_k$. Then,

$$E[S'_{t,2} S_{t,1} \beta\gamma' S'_{t,1} S_{t,2}] \ = \ E[Q'_2\, X'_t \Sigma X_t Q_1 \beta\gamma' Q'_1 X'_t \Sigma X_t Q_2] \tag{199}$$

Define $\tilde{\beta} = Q_1\beta$ and let us rotate X so that Σ is a diagonal matrix, D. The rotated X is denoted by $\tilde{X}$. Then, using formula (185), we get

$$\begin{aligned} E[S'_{t,2} S_{t,1} \beta\gamma' S'_{t,1} S_{t,2}] \ &= \ ((\mathrm{tr}\,\Sigma)^2 - \mathrm{tr}(\Sigma^2))Q'_2 Q_1 \beta\gamma' Q'_1 Q_2 \\ &+ (\mathrm{tr}\,\Sigma)^2 (Q'_2 Q_1 \beta\gamma' Q'_1 Q_2 + Q'_2 Q_1 \gamma\beta' Q'_1 Q_2) \\ &+ \mathrm{tr}(\Sigma^2)((\kappa - 2)Q'_2\,\mathrm{diag}(\tilde{\beta}_{k_1}\tilde{\gamma}_{k_1})Q_2 + (\tilde{\beta}'\tilde{\gamma})Q'_2 Q_2) \end{aligned} \tag{200}$$

whereas $\tilde{\beta}'\tilde{\gamma} = \beta' Q'_1 Q_1 \gamma$. Now,

$$Q_1 \ = \ \begin{pmatrix} Q_{1,1} \\ Q_{2,1} \end{pmatrix}, \ Q_2 \ = \ \begin{pmatrix} Q_{1,2} \\ Q_{2,2} \end{pmatrix}, \ \Psi^{1/2} \ = \ \begin{pmatrix} Q_{1,1} & Q_{1,2} \\ Q_{2,1} & Q_{2,2} \end{pmatrix}$$

Thus,

$$\Psi \ = \ \begin{pmatrix} Q'_1 Q_1 & Q'_1 Q_2 \\ Q'_2 Q_1 & Q'_2 Q_2 \end{pmatrix} \ = \ \begin{pmatrix} \Psi_{1,1} & \Psi_{1,2} \\ \Psi_{2,1} & \Psi_{2,2} \end{pmatrix} \tag{201}$$

[56] Note that Corollary 7 is not applicable here.

and hence we get the required.

□

As above, all expectations in this section are conditional on $\hat{\beta}$. We have since $\beta_1\beta_2' \to 0$ in probability that

$$\begin{aligned}
&E[(R^\pi_{t+1})^2] \;=\; E[\hat{\beta}_1' S_{t,1}' R_{t+1} R_{t+1}' S_{t,1}\hat{\beta}_1] \;=\; E[\hat{\beta}_1' S_{t,1}'((S_{t,1}\beta_1 + S_{t,2}\beta_2)(S_{t,1}\beta_1 + S_{t,2}\beta_2)' + I)S_{t,1}\hat{\beta}_1] \\
&\to\; E[\hat{\beta}_1' S_{t,1}' R_{t+1} R_{t+1}' S_{t,1}\hat{\beta}_1] \;=\; E[\hat{\beta}_1' S_{t,1}'(S_{t,1}\beta_1\beta_1' S_{t,1}' + S_{t,2}\beta_2\beta_2' S_{t,2}' + I)S_{t,1}\hat{\beta}_1] \\
&\to\; \operatorname{tr}(\Sigma)\, E[\hat{\beta}_1'\Psi_{1,1}\hat{\beta}_1] \;+\; E[\hat{\beta}_1' S_{t,1}' S_{t,1}\beta_1\beta_1' S_{t,1}' S_{t,1}\hat{\beta}_1] + \; P^{-1} b_* E[\hat{\beta}_1' S_{t,1}' S_{t,2} S_{t,2}' S_{t,1}\hat{\beta}_1] \\
&=\; \mathcal{L} \;+\; Term2 \;+\; Term3\,.
\end{aligned} \tag{202}$$

Recall that we are using the notation

$$\hat{\Psi}_T \;=\; \frac{1}{NT}\sum_{t=1}^{T} S_{t,1}' S_t \;\in \mathbb{R}^{P_1\times P}\,. \tag{203}$$

and we decompose

$$\hat{\Psi}_T \;=\; \hat{\Psi}_{T,1} \;+\; \hat{\Psi}_{T,2} \tag{204}$$

I.0.1 Term3

We have

$$\begin{aligned}
Term3 \;&=\; P^{-1} b_* E[\hat{\beta}_1' S_{t,1}' S_{t,2} S_{t,2}' S_{t,1}\hat{\beta}_1] \\
&=\; P^{-1} b_* \operatorname{tr} E[(\hat{\Psi}_T\beta \;+\; q_T)'(zI + \hat{\Psi}_{T,1})^{-1} S_{t,1}' S_{t,2} S_{t,2}' S_{t,1}(zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T\beta \;+\; q_T)] \\
&=\; P^{-1} b_* \operatorname{tr} E[S_{t,2}' S_{t,1}(zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T\beta \;+\; q_T)(\hat{\Psi}_T\beta \;+\; q_T)'(zI + \hat{\Psi}_{T,1})^{-1} S_{t,1}' S_{t,2}] \\
&\to\; P^{-1} b_* \operatorname{tr} E[S_{t,2}' S_{t,1}(zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T\beta\beta'\hat{\Psi}_T + (NT)^{-1}\hat{\Psi}_{T,1})(zI + \hat{\Psi}_{T,1})^{-1} S_{t,1}' S_{t,2}]
\end{aligned} \tag{205}$$

in probability by (148). By Lemma 18, we get

$$\begin{aligned}
&Term3 \;\to\; P^{-1}((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2))\operatorname{tr}(Z_T\Psi_{1,2}\Psi_{2,1}) \\
&+ P^{-1}\operatorname{tr}(\Sigma^2)\operatorname{tr}((\kappa-2)Q_2'\operatorname{diag}(Q_1 Z_T Q_1')Q_2) + P^{-1}\operatorname{tr}(\Sigma^2)\operatorname{tr}(Z_T\Psi_{1,1})\operatorname{tr}(\Psi_{2,2})
\end{aligned} \tag{206}$$

with

$$Z_T \;=\; (zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T\beta\beta'\hat{\Psi}_T' + (NT)^{-1}\hat{\Psi}_{T,1})(zI + \hat{\Psi}_{T,1})^{-1}\,. \tag{207}$$

Thus,

$$\begin{aligned}
&\operatorname{tr}(Z_T\Psi_{1,1}) \;=\; \operatorname{tr}((zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T\beta\beta'\hat{\Psi}_T' + (NT)^{-1}\hat{\Psi}_{T,1})(zI + \hat{\Psi}_{T,1})^{-1}\Psi_{1,1}) \\
&=\; \mathcal{L}/\operatorname{tr}(\Sigma)
\end{aligned} \tag{208}$$

by (147). At the same time,

$$\begin{aligned} & P^{-1}\operatorname{tr}(Z_T\Psi_{1,2}\Psi_{2,1}) \;\to\; b_* P^{-2}\operatorname{tr}(\hat\Psi_T'(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,2}\Psi_{2,1}(zI+\hat\Psi_{T,1})^{-1}\hat\Psi_T) \\ & +\; P^{-1}(NT)^{-1}\operatorname{tr}((zI+\hat\Psi_{T,1})^{-1}(\hat\Psi_{T,1}+zI-zI)(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,2}\Psi_{2,1}) \\ & \to\; P^{-1}b_* c^{-1}\widetilde\xi_{2,1}(z) \;+\; P^{-1}(\xi_{2,1}(z)-z\xi_{2,1}'(z)) \;\to\; 0\,. \end{aligned} \tag{209}$$

where

$$\widetilde\xi_{2,1}(z) \;=\; \lim (NT)^{-1}\operatorname{tr}(\hat\Psi_T'(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,2}\Psi_{2,1}(zI+\hat\Psi_{T,1})^{-1}\hat\Psi_T)$$

Similarly,

$$\begin{aligned} & P^{-1}\operatorname{tr}(\Sigma^2)\operatorname{tr}((\kappa-2)Q_2'\operatorname{diag}(Q_1((zI+\hat\Psi_{T,1})^{-1}(P^{-1}\hat\Psi_T\hat\Psi_T'+(NT)^{-1}\hat\Psi_{T,1})(zI+\hat\Psi_{T,1})^{-1})Q_1')Q_2) \\ & \le\; K\,P^{-1}\operatorname{tr}(Z_T) \;\to\; 0 \end{aligned} \tag{210}$$

Thus, we get

$$Term3 \;\to\; b_*\operatorname{tr}(\Sigma^2)\operatorname{tr}(\Psi_{2,2})P^{-1}\mathcal{L}/\operatorname{tr}(\Sigma) \tag{211}$$

I.0.2 Term2

By a slight modification of Lemma 18,

$$\begin{aligned} & E[S_{t,1}'S_{t,1}\beta_1\beta_1'S_{t,1}'S_{t,1}] \\ & =\; ((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))\Psi_{1,1}\beta_1\beta_1'\Psi_{1,1} \;+\operatorname{tr}(\Sigma^2)((\kappa-2)\Psi_{1,1}^{1/2}\operatorname{diag}(\Psi_{1,1}^{1/2}\beta_1\beta_1'\Psi_{1,1}^{1/2})\Psi_{1,1}^{1/2}+\operatorname{tr}(\beta_1\beta_1'\Psi_{1,1})\Psi_{1,1}) \\ & \approx\; ((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))\Psi_{1,1}\beta_1\beta_1'\Psi_{1,1} \;+\operatorname{tr}(\Sigma^2)((\kappa-2)\Psi_{1,1}^{1/2}\operatorname{diag}(\Psi_{1,1}^{1/2}\beta_1\beta_1'\Psi_{1,1}^{1/2})\Psi_{1,1}^{1/2}+b_* c^{-1}c_1\psi_{*,1}(q)\Psi_{1,1}) \end{aligned} \tag{212}$$

and therefore

$$
\begin{aligned}
&E[\hat\beta_1' S_{t,1}' S_{t,1}\beta_1\beta_1' S_{t,1}' S_{t,1}\hat\beta_1] \\
&= \operatorname{tr} E[((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))\Psi_{1,1}\beta_1\beta_1'\Psi_{1,1} \\
&+ \operatorname{tr}(\Sigma^2)((\kappa-2)\Psi_{1,1}^{1/2}\operatorname{diag}(\Psi_{1,1}^{1/2}\beta_1\beta_1'\Psi_{1,1}^{1/2})\Psi_{1,1}^{1/2}+b_*c^{-1}c_1\psi_{*,1}(q)\Psi_{1,1})\hat\beta_1\hat\beta_1'] \\
&= \operatorname{tr} E[\Bigg(((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))\Psi_{1,1}\beta_1\beta_1'\Psi_{1,1} \\
&+ \operatorname{tr}(\Sigma^2)((\kappa-2)\Psi_{1,1}^{1/2}\operatorname{diag}(\Psi_{1,1}^{1/2}\beta_1\beta_1'\Psi_{1,1}^{1/2})\Psi_{1,1}^{1/2}+b_*c^{-1}c_1\psi_{*,1}(q)\Psi_{1,1})\Bigg) \\
&\times (zI+\hat\Psi_{T,1})^{-1}(\hat\Psi_T\beta + q_T)(\hat\Psi_T\beta + q_T)'(zI+\hat\Psi_{T,1})^{-1}] \\
&\to \operatorname{tr} E[\Bigg(((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))\Psi_{1,1}\beta_1\beta_1'\Psi_{1,1} + \operatorname{tr}(\Sigma^2)b_*c^{-1}c_1\psi_{*,1}(q)\Psi_{1,1}\Bigg) \\
&\times (zI+\hat\Psi_{T,1})^{-1}(\hat\Psi_{T,1}\beta_1\beta_1'\hat\Psi_{T,1}+\hat\Psi_{T,2}\beta_2\beta_2'\hat\Psi_{T,2}' + (NT)^{-1}\hat\Psi_{T,1})(zI+\hat\Psi_{T,1})^{-1}] \\
&= ((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))T1 + ((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))T2 + ((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))T3 \\
&+ \operatorname{tr}(\Sigma^2)b_*c^{-1}c_1\psi_{*,1}(q)T4 + \operatorname{tr}(\Sigma^2)b_*c^{-1}c_1\psi_{*,1}(q)T5 + \operatorname{tr}(\Sigma^2)b_*c^{-1}c_1\psi_{*,1}(q)T6
\end{aligned}
\tag{213}
$$

where we have used the fact that the cross-terms involving $\beta_1\beta_2'$ converge to zero and that the kurtosis terms (those involving $\kappa-2$) also converge to zero.[57]

We now analyze each term separately. The first term in (213) gives

$$
\begin{aligned}
T1 &= \operatorname{tr} E[(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,1}\beta_1\beta_1'\Psi_{1,1}(zI+\hat\Psi_{T,1})^{-1}\hat\Psi_{T,1}\beta\beta'\hat\Psi_{T,1}] \\
&= E[\beta_1'\hat\Psi_{T,1}(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,1}\beta_1\beta_1'\Psi(zI+\hat\Psi_{T,1})^{-1}\hat\Psi_{T,1}\beta_1] \\
&= E[(\beta_1'\hat\Psi_{T,1}(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,1}i\beta_1)^2] \;\to\; (b_*c^{-1}c_1(\psi_{*,1}(q)-zc_1^{-1}\xi_{1,1}(z)))^2
\end{aligned}
\tag{215}
$$

in probability by Proposition 14 because all variables are uniformly bounded because $\|\beta\|$ stays bounded almost surely. Namely, first we notice that

$$
\beta'\hat\Psi_{T,1}(zI+\hat\Psi_{T,1})^{-1}\Psi\beta - b_*\frac{1}{M}\operatorname{tr}(\hat\Psi_{T,1}(zI+\hat\Psi_{T,1})^{-1}\Psi) \;\to\; 0
$$

[57] For example, defining $\tilde\beta_1=\Psi_{11}^{1/2}\beta_1$ and $A=\Psi_{1,1}^{-1/2}\hat\Psi_{T,1}(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,1}^{1/2}$ and assuming for simplicity that $\Psi_{1,1}^{-1/2}$ is bounded, we get

$$
\begin{aligned}
&\operatorname{tr} E[\beta_1'\hat\Psi_{T,1}(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,1}^{1/2}\operatorname{diag}(\Psi_{1,1}^{1/2}\beta_1\beta_1'\Psi_{1,1}^{1/2})\Psi_{1,1}^{1/2}\Psi(zI+\hat\Psi_{T,1})^{-1}\hat\Psi_{T,1}\beta_1] \\
&= E[\tilde\beta_1' A\operatorname{diag}(\tilde\beta_1\tilde\beta_1')A'\tilde\beta_1] \\
&= E[\sum_{i,j,k}\tilde\beta_i A_{i,j}\tilde\beta_j^2 A_{k,j}\tilde\beta_k] = E[\sum_{i,j}\tilde\beta_i^2 A_{i,j}^2\tilde\beta_j^2] \approx P^{-2}\operatorname{tr} E[AA'] \;\to\; 0
\end{aligned}
\tag{214}
$$

because A is bounded.

is probability. And second,

$$\frac{1}{P}\operatorname{tr}(\hat{\Psi}_{T,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi) \;=\; \frac{1}{P}\operatorname{tr}((\hat{\Psi}_{T,1}+zI-zI)(zI+\hat{\Psi}_{T,1})^{-1}\Psi) \;\to\; c^{-1}c_1(\psi_{*,1}(q)-zc_1^{-1}\xi_{1,1}(z))$$

almost surely by Proposition 14.

Then,

$$\begin{aligned} T2 &= \operatorname{tr} E[(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}\beta_1\beta_1'\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\beta_2\beta_2'\hat{\Psi}_{T,2}'] \\ &= P^{-1}b_*E[\beta_1'\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\hat{\Psi}_{T,2}'(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}\beta_1] \\ &\to P^{-2}b_*^2\operatorname{tr} E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\hat{\Psi}_{T,2}'(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}] \\ &\le \|\Psi_{1,1}\|P^{-2}b_*^2\operatorname{tr} E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\hat{\Psi}_{T,2}'(zI+\hat{\Psi}_{T,1})^{-1}] \;\to\; 0 \end{aligned} \tag{216}$$

by Lemma 15. Then,

$$\begin{aligned} T3 &= \operatorname{tr} E[(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}\beta_1\beta_1'\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\frac{1}{NT}\hat{\Psi}_{T,1}] \\ &= E[\beta_1'\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\frac{1}{NT}\hat{\Psi}_{T,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}\beta_1] \\ &\le b_*K/(NT) \;\to\; 0 \end{aligned} \tag{217}$$

for some constant K because $(zI+\hat{\Psi}_{T,1})^{-1}$ and $\Psi_{1,1}^2$ are uniformly bounded and where we have used that

$$(zI+\hat{\Psi}_{T,1})^{-1}\frac{1}{NT}\hat{\Psi}_{T,1}(zI+\hat{\Psi}_{T,1})^{-1} \;\le\; \frac{1}{NT}(zI+\hat{\Psi}_{T,1})^{-1}$$

is the sense of positive semi-definite order.

Then,

$$\begin{aligned} T4 &= \operatorname{tr} E[(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}\beta_1\beta_1'\hat{\Psi}_{T,1}] \\ &= E[\beta'\hat{\Psi}_{T,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}\beta] \end{aligned} \tag{218}$$

where

$$\beta_1'\hat{\Psi}_{T,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}\beta_1 - b_*\frac{1}{P}\operatorname{tr}(\hat{\Psi}_{T,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}) \;\to\; 0$$

in probability. Now, since the matrices $\hat{\Psi}_{T,1}$ and $(zI+\hat{\Psi}_{T,1})^{-1}$ commute, we get

$$\operatorname{tr}(\hat{\Psi}_{T,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}) \;=\; \operatorname{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-2}\hat{\Psi}_{T,1}^2)\,.$$

Using the identity

$$\hat{\Psi}_{T,1}^2 \;=\; (\hat{\Psi}_{T,1}+zI)^2 - 2z(\hat{\Psi}_{T,1}+zI)+z^2\,,$$

we get

$$\begin{aligned}&\frac{1}{P}\operatorname{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-2}((\hat{\Psi}_{T,1}+zI)^2-2z(\hat{\Psi}_{T,1}+zI)+z^2))\\ &=\ \frac{1}{P}\operatorname{tr}(\Psi_{1,1}\ -2z\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}+z^2\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-2})\,.\end{aligned} \tag{219}$$

By Proposition 14,

$$\frac{1}{P_1}\operatorname{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1})\ \to\ c_1^{-1}\,\xi_{1,1}(z)\,.$$

Then, standard arguments for analytic functions ($\xi_{1,1}(z)$ is analytic for z for $\Re z>0$)[58] imply that

$$\frac{1}{P_1}\operatorname{tr}((zI+\hat{\Psi}_{T,1})^{-2}\Psi_{1,1})\ \to\ -c^{-1}\xi'_{1,1}(z)\,. \tag{220}$$

Thus,

$$\begin{aligned}&T4=\operatorname{tr}E[(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}\beta_1\beta_1'\hat{\Psi}_{T,1}]\\ &=\ E[\beta_1'\hat{\Psi}_{T,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}\beta_1]\\ &\sim\ b_*c^{-1}c_1P_1^{-1}\operatorname{tr}(\Psi_{1,1}\ -2z\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}+z^2\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-2})\\ &=\ b_*c^{-1}c_1\psi_{*,1}(q)\ -\ b_*2zc^{-1}c_1P_1^{-1}\operatorname{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1})\ +\ b_*c^{-1}c_1P_1^{-1}z^2\operatorname{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-2})\\ &\sim\ b_*c^{-1}c_1(\psi_{*,1}(q)-2zc_1^{-1}\xi_{1,1}(z)-z^2c_1^{-1}\xi'_{1,1}(z))\end{aligned} \tag{221}$$

because

$$\frac{1}{P_1}\operatorname{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1})\ \to\ c_1^{-1}\xi_{1,1}(z) \tag{222}$$

implies

$$\frac{1}{P_1}\operatorname{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-2})\ \to\ -c_1^{-1}\xi'_{1,1}(z)\,. \tag{223}$$

[58]For analytic functions, uniform boundedness plus convergence on an open set implies converges of derivatives by the Cauchy integral formula.

Now, since $\beta_2\beta_2' \sim P^{-1} I_{P_2\times P_2}$, we get by Lemma 15 that

$$\begin{aligned}
T5 &= \operatorname{tr} E[(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,1}(zI+\hat\Psi_{T,1})^{-1}\hat\Psi_{T,2}\beta_2\beta_2'\hat\Psi_{T,2}'] \\
&\to\ P^{-1}\operatorname{tr} E[(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,1}(zI+\hat\Psi_{T,1})^{-1}\hat\Psi_{T,2}\hat\Psi_{T,2}'] \\
&\to\ \frac{\operatorname{tr}(\Psi_{2,2})}{P}\, b_* \frac{1}{N}\widehat{\xi}_{1,1}(z)\operatorname{tr}[\Sigma^2(I+\xi_{1,1}(z)\,\Sigma)^{-2}] \\
&+ b_* \frac{1}{N^2}(\operatorname{tr}[(I+\xi_{1,1}(z)\,\Sigma)^{-1}\Sigma])^2 c^{-1}\widehat{\xi}_{2,1}(z) \\
&- 2b_* \frac{1}{N^2}\widehat{\xi}_{1,1}(z)\operatorname{tr}((I+\xi_{1,1}(z)\,\Sigma)^{-1}\Sigma)c^{-1}\xi_{2,1}(z)\operatorname{tr}[(I+\xi_{1,1}(z)\,\Sigma)^{-2}\Sigma^2]
\end{aligned} \tag{224}$$

Finally,

$$\begin{aligned}
T6 &= \frac{1}{NT}\operatorname{tr} E[(zI+\hat\Psi_{T,1})^{-1}\Psi(zI+\hat\Psi_{T,1})^{-1}\hat\Psi_{T,1}] \\
&= \frac{1}{NT}\operatorname{tr} E[(zI+\hat\Psi_{T,1})^{-1}\Psi(zI+\hat\Psi_{T,1})^{-1}(zI+\hat\Psi_{T,1}-zI)] \\
&\sim\ (\xi_{1,1}(z)+z\xi_{1,1}'(z))
\end{aligned} \tag{225}$$

where we have used that

$$\operatorname{tr} E[(zI+\hat\Psi_{T,1})^{-2}\Psi] \ =\ \operatorname{tr} E[(zI+\hat\Psi_{T,1})^{-1}\Psi(zI+\hat\Psi_{T,1})^{-1}]$$

Putting all the terms together, we finally get from (202) that

$$\begin{aligned}
E[(R_{t+1}^{\pi})^2] &= \mathcal{L} + Term2 + Term3 \\
&= \mathcal{L} + \underbrace{\operatorname{tr}(\Sigma^2)\operatorname{tr}(\Psi_{2,2})P^{-1}\mathcal{L}/\operatorname{tr}(\Sigma)}_{Term3\ by\ (211)} \\
&+ ((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))T1 + ((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))T2 + ((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))T3 \\
&+ \underbrace{\operatorname{tr}(\Sigma^2)b_*c^{-1}c_1\psi_{*,1}(q)T4 + \operatorname{tr}(\Sigma^2)b_*c^{-1}c_1\psi_{*,1}(q)T5 + \operatorname{tr}(\Sigma^2)b_*c^{-1}c_1\psi_{*,1}(q)T6}_{by\ (213)} \\
&= \mathcal{L} + \operatorname{tr}(\Sigma^2)\operatorname{tr}(\Psi_{2,2})P^{-1}\mathcal{L}/\operatorname{tr}(\Sigma) \\
&+ ((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))\Big((b_*c^{-1}c_1(\psi_{*,1}(q)-zc_1^{-1}\xi_{1,1}(z)))^2\Big) \\
&+ \operatorname{tr}(\Sigma^2)b_*c^{-1}c_1\psi_{*,1}(q)\Big(b_*c^{-1}c_1(\psi_{*,1}(q)-2zc_1^{-1}\xi_{1,1}(z)-z^2c_1^{-1}\xi_{1,1}'(z))\Big) \\
&+ \operatorname{tr}(\Sigma^2)b_*c^{-1}c_1\psi_{*,1}(q)\Delta(z) \\
&+ \operatorname{tr}(\Sigma^2)b_*c^{-1}c_1\psi_{*,1}(q)\Big(\xi_{1,1}(z)+z\xi_{1,1}'(z)\Big) \\
&= \mathcal{L}(z) + \operatorname{tr}(\Sigma^2)\operatorname{tr}(\Psi_{2,2})P^{-1}\mathcal{L}(z)/\operatorname{tr}(\Sigma) + ((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))\mathcal{E}(z)^2 \\
&+ \operatorname{tr}(\Sigma^2)b_*c^{-1}c_1\psi_{*,1}(q)\mathcal{L}(z)/\operatorname{tr}(\Sigma) \\
&= \mathcal{L}(z)\Big(1+\frac{\operatorname{tr}(\Sigma^2)}{\operatorname{tr}(\Sigma)}b_*\frac{\operatorname{tr}(\Psi)}{P}\Big) + ((\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2))\mathcal{E}(z)^2
\end{aligned} \tag{226}$$

J Proofs of Propositions 11, 12, 3, and 4

Proof. Proposition 11 follows directly from Lemma 11.
Proposition 12 follows directly from Lemma 14.
Proposition 3 follows directly from formula (226).
To prove Proposition 4, we notice that, under the made assumption, we have

$$\gamma \approx \frac{\mathrm{tr}(\Sigma^2)}{\mathrm{tr}(\Sigma)} \tag{227}$$

For simplicity, we normalize $\mathrm{tr}(\Sigma) = 1$. Then, defining $k = 1 + b_* \,\mathrm{tr}(\Sigma^2)\frac{\mathrm{tr}(\Psi_{2,2})}{P}$, we get that, by (88),

$$\begin{aligned} 2\mathcal{E} - \mathcal{L} &= 2b_* q\nu - qb_*\hat{\nu} + cq\nu' k \\ &= \lim P^{-1}\,\mathrm{tr}\left((2b_* q\hat{\Psi}(zI+\hat{\Psi})^{-1} - cq(1+k)\hat{\Psi}(zI+\hat{\Psi})^{-2} - qb_*\hat{\Psi}^2(zI+\hat{\Psi})^{-2})\Psi\right) \end{aligned} \tag{228}$$

Consider the function

$$f(z) = 2b_* x(z+x)^{-1} - c(1+k)x(z+x)^{-2} - x^2(z+x)^{-2}\,. \tag{229}$$

for any $x > 0$. Then,

$$f'(z) = -2b_* x(z+x)^{-2} + 2c(1+k)x(z+x)^{-3} + 2b_* x^2(z+x)^{-3}\,. \tag{230}$$

Then, $f'(z) = 0$ is equivalent to $-2b_* x(z+x) + 2c(1+k)x + 2b_* x^2 = 0$ implying that

$$z_* = c(1+k)/b_* \tag{231}$$

Furthermore,

$$f(z_*) = \frac{b_* x^2 + 2b_* x z_* - c(1+k)x}{(z_*+x)^2} = \frac{b_* x^2 + b_* x z_*}{(z+x)^2} = b_* \frac{x}{x+z} \tag{232}$$

implying that

$$2\mathcal{E}(z_*) - \mathcal{L}(z_*) = \mathcal{E}(z_*) = b_*\nu(z_*)$$

Furthermore, we have that

$$\frac{\mathcal{E}^2}{\mathcal{L}} \tag{233}$$

Similarly,

$$\frac{\mathcal{E}^2(z)}{\mathcal{L}(z)} = \lim \frac{(b_* q)^2(\mathrm{tr}(\hat{\Psi}(zI+\hat{\Psi})^{-1}\Psi))^2}{q\,\mathrm{tr}((c(1+k)\hat{\Psi} + b_*\hat{\Psi}^2)(zI+\hat{\Psi})^{-2}\Psi)} \tag{234}$$

Define

$$f(z) \;=\; \frac{(\mathrm{tr}(\hat{\Psi}(zI+\hat{\Psi})^{-1}\Psi))^2}{\mathrm{tr}((c(1+k)\hat{\Psi}+b_*\hat{\Psi}^2)(zI+\hat{\Psi})^{-2}\Psi)}$$

Diagonalizing $\hat{\Psi}$ and defining a measure weighted by the eigenvalues of Ψ, we can rewrite it as

$$f(z) \;=\; \frac{E[X(z+X)^{-1}]^2}{E[(aX+bX^2)(z+X)^{-2}]}\,.$$

Then,

$$\begin{aligned} f'(z) \;=\; \Big(& -2E[X(z+X)^{-1}]E[X(z+X)^{-2}]E[(aX+bX^2)(z+X)^{-2}] \\ & + 2E[X(z+X)^{-1}]^2E[(aX+bX^2)(z+X)^{-3}]\Big)/E[(aX+bX^2)(z+X)^{-2}]^2\,. \end{aligned} \tag{235}$$

with $a=c(1+k), b=b_*$. Thus, $f'(z)\geq 0$ if and only if

$$E[X(z+X)^{-2}]E[(aX+bX^2)(z+X)^{-2}] \;\leq\; E[X(z+X)^{-1}]E[(aX+bX^2)(z+X)^{-3}]\,. \tag{236}$$

Changing the measure to $X(z+X)^{-1}/E[X(z+X)^{-1}]$, we can rewrite it as

$$E[(z+X)^{-1}]E[(a+bX)(z+X)^{-1}] \;\leq\; E[(a+bX)(z+X)^{-2}]\,. \tag{237}$$

The function $(z+X)^{-1}$ is decreasing in X, while $(a+bX)(z+X)^{-1}$ is decreasing in X if and only if $z<a/b$. Thus, $f(z)$ is increasing for $z<z_*$ and decreasing otherwise. The proof is complete.

To prove the virtue of complexity, we assume that the dispersion of eigenvalues of Σ is small and it remains to consider

$$\nu(z_*) \;=\; q\,\psi_{*,1} - z_* c^{-1}\xi(z_*;cq) \tag{238}$$

where

$$\xi(z,cq) \;=\; \frac{1-zm(-z;cq)}{(cq)^{-1}-1+zm(-z;cq)} \;=\; -1 \;+\; \frac{(cq)^{-1}}{(cq)^{-1}-1+zm(-z;cq)}\,. \tag{239}$$

Theorem 13 implies

$$zm(-z) = \int \frac{zdH(x)}{x(1-c+czm)+z}\,,$$

and, hence,

$$\tilde{m}(-z;c) = (1-c)z^{-1} \; + \; cm(-z;c)\,, \tag{240}$$

is the unique positive solution to

$$z = \int \frac{(1-(c-1)\tilde{m}x)\,dH(x)}{\tilde{m}(1+\tilde{m}\,x)} \tag{241}$$

Furthermore,

$$\nu(z_*) \;=\; q\psi_{*,1} - b_*^{-1}\xi(c/b_*, cq) \;=\; c^{-1}(cq\psi_{*,1} - z_*\xi(z_*;cq))$$

Thus, our goal is to show that

$$c\psi_{*,1} - z\xi(z;c)$$

is monotone increasing in c for any $z > 0$. We have

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

and hence we need

$$f(c) \;=\; c\psi_{*,1} - \frac{1}{\tilde{m}}$$

to be monotone increasing in c. That is, we need the inequality

$$f'(c) \;=\; \psi_{*,1} + \frac{\tilde{m}'(c)}{\tilde{m}^2} \;\geq\; 0\,. \tag{242}$$

We have

$$0 \;=\; \int \frac{(-\tilde{m}x-(c-1)\tilde{m}'(c)x)\tilde{m}(1+\tilde{m}x) - (1-(c-1)\tilde{m}x)(\tilde{m}'(1+\tilde{m}x)+\tilde{m}\tilde{m}'x)}{\tilde{m}^2(1+\tilde{m}x)^2} dH(x) \tag{243}$$

so that

$$\tilde{m}'(c) \;=\; \frac{-\int \frac{xdH(x)}{1+\tilde{m}x}}{\int \frac{(c-1)x\tilde{m}(1+\tilde{m}x)+(1-(c-1)\tilde{m}x)(1+2\tilde{m}x)}{\tilde{m}^2(1+\tilde{m}x)^2} dH(x)}\,, \tag{244}$$

We start with the observation that the denominator in this fraction is always non-negative. Indeed, (241) implies that

$$(c-1)\int \frac{\tilde{m}xdH(x)}{\tilde{m}(1+\tilde{m}x)} \;=\; \int \frac{dH(x)}{\tilde{m}(1+\tilde{m}x)} - z\,. \tag{245}$$

Multiplying by $\tilde{m} > 0$, we get that

$$(c-1)\int \frac{\tilde{m}xdH(x)}{(1+\tilde{m}x)} \;= \int \frac{dH(x)}{(1+\tilde{m}x)} \;- z\tilde{m} \tag{246}$$

which implies

$$\begin{aligned}
&(c-1)\int \frac{(\tilde{m}x)^2 dH(x)}{(1+\tilde{m}x)^2} \;= (c-1)\int \frac{\tilde{m}x(-1+1+\tilde{m}x)dH(x)}{(1+\tilde{m}x)^2} \\
&= \int \frac{dH(x)}{(1+\tilde{m}x)} \;- z\tilde{m} - (c-1)\int \frac{\tilde{m}xdH(x)}{(1+\tilde{m}x)^2} \\
&= -z\tilde{m} \;+\; \int \frac{(1+(2-c)\tilde{m}x)dH(x)}{(1+\tilde{m}x)^2}
\end{aligned} \tag{247}$$

Thus,

$$\begin{aligned}
&\int \frac{(c-1)x\tilde{m}(1+\tilde{m}x)+(1-(c-1)\tilde{m}x)(1+2\tilde{m}x)}{\tilde{m}^2(1+\tilde{m}x)^2} dH(x) \\
&= \int \frac{-(c-1)(x\tilde{m})^2+1+2\tilde{m}x}{(1+\tilde{m}x)^2} dH(x) \\
&= z\tilde{m} \;+\; \int \frac{c\tilde{m}xdH(x)}{(1+\tilde{m}x)^2}
\end{aligned} \tag{248}$$

so that (242) is equivalent to

$$\int xdH(x)\left(z\tilde{m} \;+\; \int \frac{c\tilde{m}x}{(1+\tilde{m}x)^2}\right) \;\geq\; \int \frac{xdH(x)}{1+\tilde{m}x} \tag{249}$$

It is straightforward to show that this inequality holds as identify in the limit as $c \to \infty$.

We can rewrite

$$(c-1)\int \frac{\tilde{m}xdH(x)}{(1+\tilde{m}x)} \;=\; \int \frac{dH(x)}{(1+\tilde{m}x)} - z\tilde{m}\,. \tag{250}$$

as

$$c-1 \;+\; z\tilde{m} - c\int \frac{dH(x)}{(1+\tilde{m}x)} \;=\; 0\,. \tag{251}$$

Lemma 19 *The function $1/\tilde{m}(c)$ is monotone increasing and*

When $c \to \infty$, $\tilde{m} \to 0$ and we get

$$1-c^{-1} \;+\; c^{-1}z\tilde{m} - \int (1-\tilde{m}x+\tilde{m}^2x^2)dH(x) \;\approx\; 0\,. \tag{252}$$

Furthermore, optimal $z = c\,y$ for some $y > 0$. Substituting $m = ac^{-1} + bc^{-2}$ gives

$$-c^{-1} \;+\; c^{-1}ycac^{-1} + E[x](ac^{-1} + bc^{-2}) - E[x^2]a^2c^{-2} \;=\; 0 \tag{253}$$

implying

$$a = (E[x] + y)^{-1}\,,$$

and

$$E[x]b - E[x^2](E[x] + y)^{-2} \;=\; 0$$

so that

$$b \;=\; E[x^2]E[x]^{-1}(E[x] + y)^{-2}\,.$$

Thus,

$$\begin{aligned}
&\int xdH(x)\left(z\tilde{m} \;+\; \int \frac{c\tilde{m}x}{(1+\tilde{m}x)^2}\right) \;-\int \frac{xdH(x)}{1+\tilde{m}x} \\
&=\; E[x]\left(z(ac^{-1} + bc^{-2}) \;+\; \int c(ac^{-1} + bc^{-2})x(1 - 2\tilde{m}x - (\tilde{m}x)^2 + 4(\tilde{m}x)^2)\right) \\
&- \int xdH(x)(1 - \tilde{m}x + (\tilde{m}x)^2) \;+\; O(c^{-3}) \\
&=\; E[x]\left(cy(ac^{-1} + bc^{-2}) \;+\; \int (a + bc^{-1})(x - 2(ac^{-1} + bc^{-2})x^2 + 3(ac^{-1})^2x^3)\right) \\
&- \int dH(x)(x - (ac^{-1} + bc^{-2})x^2 + (ac^{-1})^2x^3) \;+\; O(c^{-3}) \\
&=\; E[x]byc^{-1} + bE[x]^2c^{-1} - 2E[x]a^2c^{-1}E[x^2] + ac^{-1}E[x^2] \;+\; O(c^{-2}) \\
&=\; E[x]E[x^2]E[x]^{-1}(E[x] + y)^{-2}yc^{-1} + E[x^2]E[x]^{-1}(E[x] + y)^{-2}E[x]^2c^{-1} \\
&- 2E[x](E[x] + y)^{-2}c^{-1}E[x^2] + (E[x] + y)^{-1}c^{-1}E[x^2] \;+\; O(c^{-2}) \\
&=\; \frac{2E[x^2]y}{(E[x] + y)^2} \;>\; 0\,.
\end{aligned} \tag{254}$$

□

K Recovering b_*

Propositions 10 and 4 make concrete predictions about the behavior of out of sample Sharpe ratios and R^2. Testing these predictions would require estimating three objects: $m(-z; c)$ (the eigenvalue distribution of $\hat{\Psi}$), $m_\Sigma(-z)$ (the eigenvalue distribution of Σ) and, most importantly, the parameter b_* capturing the amount of predictability. In this Section, we use random matrix theory to develop estimators for this important parameter and then study its dynamics over time. For simplicity, we only consider the case of a correctly specified model, assuming that $q = 1$.

We start with the following result.

Proposition 15 *Consider the expected return on the in-sample timing strategy: Estimate $\hat{\beta}$ using $(S_t, R_{t+1})_{t\leq T}$, and then compute $\sum_{t=1}^{T} \hat{\beta}' S_t' R_{t+1}$. Then,*

$$\begin{aligned}
&\lim_{P,T\to\infty} (NT)^{-1} E[\sum_{t=1}^{T} \hat{\beta}' S_t' R_{t+1}] \\
&= \underbrace{\xi(z;c) N^{-1} \operatorname{tr}(\Sigma(I+\xi(z;c)\Sigma)^{-1})}_{overfitting\ noise} \\
&+\ b_* \Bigg(\underbrace{\psi_{*,1}\xi(z;c) N^{-1} \operatorname{tr}(\Sigma^2(I+\xi(z;c)\Sigma)^{-1})}_{overfit\ to\ in-sample\ signals} \\
&+\ \operatorname{tr}(\Sigma/N)(\psi_{*,1} - c^{-1} z\xi(z;c)) \Bigg)
\end{aligned} \tag{255}$$

Formula (255) implies that b_* can be recovered from the in-sample performance of the timing strategy. To proceed further, we need to discuss how we estimate Σ and Ψ.

Our first observation is that

$$\hat{\Sigma} = \frac{1}{TP} \sum_t S_t S_t' \ \to_{T,P\to\infty}\ \Sigma\psi_{*,1} \tag{256}$$

in probability. Our second observation is that, by definition, $\hat{m}(-z;c) = P^{-1} \operatorname{tr}((zI + \hat{\Psi})^{-1})$ converges to $m(-z;c)$ in probability, where

$$\hat{\Psi} \ = \ \frac{1}{TN} \sum_t S_t' S_t \, . \tag{257}$$

Our next observation is that, by (256), the unique solution $\hat{\xi}(z)$ to

$$1 - zm(-z;c) \ = \ c^{-1} N^{-1} (N - \operatorname{tr}((I + \hat{\xi}(z)\hat{\Sigma})^{-1}) \tag{258}$$

converges to $\xi(z)/\psi_{*,1}$ in probability. By formula (90) (assuming a correctly specified model, so that $q = 1$), we have

$$\mathcal{E}(z;c) \ = \ b_* \operatorname{tr}(\Sigma)\nu(z;c) \ = \ b_* \operatorname{tr}(\Sigma)(\psi_{*,1} - z\xi(z)) \ \approx \ b_* \operatorname{tr}(\Sigma)\psi_{*,1}(1 - z\hat{\xi}(z)) \ \approx \ b_* \operatorname{tr}(\Sigma)(1 - z\hat{\xi}(z)) \tag{259}$$

Similarly, formula (255) the in-sample expected return is given by converges to

$$
\begin{aligned}
& T^{-1}E[\sum_{t=1}^{T}\hat{\beta}'S_t'R_{t+1}] \\
& \approx \xi(z)\,\mathrm{tr}(\Sigma(I+\xi(z)\Sigma)^{-1}) \;+\; b_*\psi_{*,1}\xi(z)\,\mathrm{tr}(\Sigma^2(I+\xi(z)\Sigma)^{-1}) \\
& +\; \mathrm{tr}(\Sigma)b_*(\psi_{*,1}-c^{-1}z\xi(z)) \\
& \approx \underbrace{\hat{\xi}(z)\,\mathrm{tr}(\Sigma(I+\hat{\xi}(z)\Sigma)^{-1})}_{intrinsic\ overfit} \\
& +\; b_*\left(\hat{\xi}(z)\,\mathrm{tr}(\Sigma^2(I+\hat{\xi}(z)\Sigma)^{-1})+\mathrm{tr}(\Sigma)(1-zc^{-1}\hat{\xi}(z))\right),
\end{aligned}
\tag{260}
$$

and we arrive at the following result.

Proposition 16 *Given any grid of z parameters, both*

$$
\hat{b}_*(grid;inSample) \;=\; \frac{1}{len(grid)}\sum_{z\in grid}\frac{T^{-1}\sum_{t=1}^{T}\hat{\beta}(z)'S_t'R_{t+1}-\hat{\xi}(z)\,\mathrm{tr}(\Sigma(I+\hat{\xi}(z)\Sigma)^{-1})}{\hat{\xi}(z)\,\mathrm{tr}(\Sigma^2(I+\hat{\xi}(z)\Sigma)^{-1})+\mathrm{tr}(\Sigma)(1-zc^{-1}\hat{\xi}(z))} \tag{261}
$$

and

$$
\hat{b}_*(grid;OOS) \;=\; \frac{1}{len(grid)}\sum_{z\in grid}\frac{\frac{1}{T_{oos}}\sum_{t\in OOS}\hat{\pi}_t(z)'R_{t+1}}{\mathrm{tr}(\Sigma)(1-zc^{-1}\hat{\xi}(z))} \tag{262}
$$

are unbiased estimators of b_*.

Proof of Proposition 255. We have

$$
\hat{\beta} \;=\; (zI+\hat{\Psi})^{-1}(NT)^{-1}\sum_{t=1}^{T}S_t'R_{t+1}\,.
$$

We have by symmetry over t and the fact that $E_T[R_{T+1}R_{T+1}]=S_T\beta\beta'S_T'+I$ that

$$
\begin{aligned}
& (NT)^{-1}E[\sum_{t=1}^{T}\hat{\beta}'S_t'R_{t+1}] \\
& =\; N^{-1}E[\hat{\beta}'S_T'R_{T+1}] \\
& =\; N^{-1}E[R_{T+1}'S_T(zI+\hat{\Psi})^{-1}(NT)^{-1}\sum_{t=1}^{T}S_t'R_{t+1}] \\
& =\; N^{-1}(NT)^{-1}\,\mathrm{tr}\,E[R_{T+1}R_{T+1}'S_T(zI+\hat{\Psi})^{-1}S_T'] \;+\; N^{-2}(1-T^{-1})E[R_{T+1}'S_T(zI+\hat{\Psi})^{-1}S_t'R_{t+1}] \\
& =\; (N^2T)^{-1}\,\mathrm{tr}\,E[(S_T\beta\beta'S_T'+I)S_T(zI+\hat{\Psi})^{-1}S_T'] \;+\; N^{-2}(1-T^{-1})E[\beta'S_T'S_T(zI+\hat{\Psi})^{-1}S_t'S_t\beta]
\end{aligned}
$$

(263)

for any $t \neq T$. We have by the Woodbury matrix identity

$$(zI+\hat{\Psi}_T)^{-1} \;=\; (zI+\hat{\Psi}_{T,t})^{-1} - \frac{1}{NT}(zI+\hat{\Psi}_{T,t})^{-1}S_t'(I+\frac{1}{NT}S_t(zI+\hat{\Psi}_{T,t})^{-1}S_t')^{-1}S_t(zI+\hat{\Psi}_{T,t})^{-1} \tag{264}$$

Furthermore, by Lemma 9,

$$C_T \;=\; \frac{1}{NT}S_t(zI+\hat{\Psi}_{T,t})^{-1}S_t' \;\to\; \xi_T(z)\,\Sigma\,.$$

in probability. Thus, using the fact that

$$E[S_T X S_T'] \;=\; \Sigma\,\mathrm{tr}(\Psi X)$$

for any X independent of S_T, we get

$$\begin{aligned}
&(NT)^{-1}\,\mathrm{tr}\,E[S_T(zI+\hat{\Psi}_T)^{-1}S_T'] \\
&\approx\; (NT)^{-1}\,\mathrm{tr}\,E[S_T((zI+\hat{\Psi}_{T,T})^{-1} \\
&-\; \frac{1}{NT}(zI+\hat{\Psi}_{T,T})^{-1}S_T'(I+\frac{1}{NT}S_T(zI+\hat{\Psi}_{T,T})^{-1}S_T')^{-1}S_T(zI+\hat{\Psi}_{T,T})^{-1})S_T'] \\
&\approx\; \xi(z)\,\mathrm{tr}(\Sigma)\;\; - \mathrm{tr}\,E[S_T\frac{1}{NT}(zI+\hat{\Psi}_{T,T})^{-1}S_T'(I+C_T)^{-1}C_T] \\
&\approx\; \xi(z)\,\mathrm{tr}(\Sigma) - \xi(z)\,\mathrm{tr}(\Sigma(I+C_T)^{-1}C_T) \\
&\approx\; \xi(z)\,\mathrm{tr}(\Sigma(I+\xi(z)\Sigma)^{-1})\,.
\end{aligned} \tag{265}$$

Then, we have

$$\begin{aligned}
&(NT)^{-1}\,\mathrm{tr}\,E[S_T\beta\beta'S_T'S_T(zI+\hat{\Psi}_T)^{-1}S_T'] \\
&\approx\; (NT)^{-1}\,\mathrm{tr}\,E[S_T\beta\beta'S_T'S_T((zI+\hat{\Psi}_{T,T})^{-1} \\
&-\; \frac{1}{NT}(zI+\hat{\Psi}_{T,T})^{-1}S_T'(I+\frac{1}{NT}S_T(zI+\hat{\Psi}_{T,T})^{-1}S_T')^{-1}S_T(zI+\hat{\Psi}_{T,T})^{-1})S_T'] \\
&=\; (NT)^{-1}\,\mathrm{tr}\,E[S_T\beta\beta'S_T'S_T(zI+\hat{\Psi}_{T,T})^{-1}S_T'] \\
&-\; (NT)^{-1}\,\mathrm{tr}\,E[S_T\beta\beta'S_T'S_T(zI+\hat{\Psi}_{T,T})^{-1}S_T'(I+C_T)^{-1}C_T]
\end{aligned} \tag{266}$$

By Lemma 9, $(NT)^{-1}S_T(zI+\hat{\Psi}_{T,T})^{-1}S_T' \approx \xi(z)\Sigma$[59] and hence

$$\begin{aligned}
&(NT)^{-1}\operatorname{tr} E[S_T\beta\beta' S_T' S_T(zI+\hat{\Psi}_T)^{-1}S_T'] \\
&\approx \operatorname{tr} E[S_T\beta\beta' S_T'\Sigma]\xi(z) \\
&- \operatorname{tr} E[S_T\beta\beta' S_T' C_T(I+C_T)^{-1}C_T] \\
&= \operatorname{tr} E[\beta' S_T'\Sigma S_T\beta]\xi(z) \\
&- \operatorname{tr} E[\beta' S_T' C_T(I+C_T)^{-1}C_T S_T\beta] \\
&\approx b_*\psi_{*,1}\xi(z)(\operatorname{tr}(\Sigma^2) - \operatorname{tr}(\Sigma^3\xi(z)(I+\xi(z)\Sigma)^{-1})) \\
&= b_*\psi_{*,1}\xi(z)\operatorname{tr}(\Sigma^2(I+\xi(z)\Sigma)^{-1})
\end{aligned} \tag{267}$$

Similarly, since $(I+C_T)^{-1}C_T \to (I+\xi(z)\Sigma)^{-1}\xi(z)\Sigma$ in probability and is uniformly bounded, the last term in (263) satisfies

$$\begin{aligned}
&N^{-2}E[\beta' S_T' S_T(zI+\hat{\Psi})^{-1}S_t' S_t\beta] \\
&= N^{-2}E[\beta' S_T' S_T((zI+\hat{\Psi}_{T,t})^{-1} \\
&- \frac{1}{NT}(zI+\hat{\Psi}_{T,t})^{-1}S_t'(I+\frac{1}{NT}S_t(zI+\hat{\Psi}_{T,t})^{-1}S_t')^{-1}S_t(zI+\hat{\Psi}_{T,t})^{-1})S_t' S_t\beta] \\
&\approx N^{-2}E[\beta' S_T' S_T(zI+\hat{\Psi}_{T,t})^{-1}S_t' S_t\beta] \\
&- N^{-2}E[\beta' S_T' S_T(zI+\hat{\Psi}_{T,t})^{-1}S_t'(I+C_T)^{-1}C_T S_t\beta] \\
&\approx N^{-2}E[\beta' S_T' S_T(zI+\hat{\Psi}_{T,t})^{-1}\Psi\operatorname{tr}(\Sigma)\beta] \\
&- N^{-2}E[\beta' S_T' S_T(zI+\hat{\Psi}_{T,t})^{-1}\Psi\beta]\operatorname{tr}(\Sigma(I+\xi(z)\Sigma)^{-1}\xi(z)\Sigma)\,.
\end{aligned} \tag{268}$$

Furthermore, by Woodbury matrix identity and the same argument as above,

$$\begin{aligned}
&N^{-2}E[\beta' S_T' S_T(zI+\hat{\Psi}_{T,t})^{-1}\Psi\operatorname{tr}(\Sigma)\beta] \\
&= N^{-2}E[\beta' S_T' S_T((zI+\hat{\Psi}_{T,t,T})^{-1} \\
&- \frac{1}{NT}(zI+\hat{\Psi}_{T,t,T})^{-1}S_T'(I+\frac{1}{NT}S_T(zI+\hat{\Psi}_{T,t,T})^{-1}S_T')^{-1}S_T(zI+\hat{\Psi}_{T,t,T})^{-1})\Psi\operatorname{tr}(\Sigma)\beta] \\
&= N^{-2}E[\beta' S_T' S_T(zI+\hat{\Psi}_{T,t,T})^{-1}\Psi\operatorname{tr}(\Sigma)\beta] \\
&- N^{-2}E[\beta' S_T' C_T(I+C_T)^{-1}S_T(zI+\hat{\Psi}_{T,t,T})^{-1}\Psi\operatorname{tr}(\Sigma)\beta] \\
&\approx (\operatorname{tr}(\Sigma/N))^2 E[\beta'\Psi(zI+\hat{\Psi}_{T,t,T})^{-1}\Psi\beta] \\
&- E[\beta'\Psi(zI+\hat{\Psi}_{T,t,T})^{-1}\Psi\operatorname{tr}(\Sigma)\beta]N^{-2}\operatorname{tr}(\Sigma)\operatorname{tr}(\Sigma(I+\xi(z)\Sigma)^{-1}\xi(z)\Sigma)
\end{aligned} \tag{269}$$

[59] Formally, since we only impose conditions on four moments of X, we need to integrate $E[S_T\beta\beta' S_T' S_T(zI+\hat{\Psi}_{T,T})^{-1}S_T']$ over S_T to verify the eligibility of this argument. This is straightforward: Forth order moments appear, but they are asymptotically negligible.

Thus, it remains to compute

$$E[\beta'\Psi(zI+\hat{\Psi}_{T,t,T})^{-1}\Psi\,\mathrm{tr}(\Sigma)\beta] \;\approx\; b_* P^{-1}\,\mathrm{tr}\,E[\Psi(zI+\hat{\Psi}_{T,t,T})^{-1}\Psi] \tag{270}$$

We have by the Woodbury matrix identity

$$(zI+\hat{\Psi}_T)^{-1} \;=\; (zI+\hat{\Psi}_{T,t})^{-1} - \frac{1}{NT}(zI+\hat{\Psi}_{T,t})^{-1}S_t'(I+\frac{1}{NT}S_t(zI+\hat{\Psi}_{T,t})^{-1}S_t')^{-1}S_t(zI+\hat{\Psi}_{T,t})^{-1} \tag{271}$$

and therefore

$$\begin{aligned}
\psi_{*,1} \;&=\; P^{-1}\,\mathrm{tr}(\Psi) \;=\; P^{-1}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_T)^{-1}(zI+\hat{\Psi}_T)\Psi] \\
&=\; c^{-1}z\xi(z) \;+\; P^{-1}\frac{1}{NT}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_T)^{-1}\sum_t S_t'S_t\Psi] \\
&=\; c^{-1}z\xi(z) \;+\; P^{-1}N^{-1}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_T)^{-1}S_t'S_t\Psi] \\
&=\; c^{-1}z\xi(z) \;+\; P^{-1}N^{-1}\,\mathrm{tr}\,E[\Bigg((zI+\hat{\Psi}_{T,t})^{-1} \\
&-\; \frac{1}{NT}(zI+\hat{\Psi}_{T,t})^{-1}S_t'(I+\frac{1}{NT}S_t(zI+\hat{\Psi}_{T,t})^{-1}S_t')^{-1}S_t(zI+\hat{\Psi}_{T,t})^{-1}\Bigg)S_t'S_t\Psi]
\end{aligned} \tag{272}$$

By Lemma 9,

$$C_T \;=\; \frac{1}{NT}S_t(zI+\hat{\Psi}_{T,t})^{-1}S_t' \;\to\; \xi_T(z)\,\Sigma\,.$$

in probability. Therefore, by a standard argument based on the Lebesgue dominated con-

vergence theorem implies that

$$
\begin{aligned}
&\psi_{*,1} \\
&\approx\ c^{-1}z\xi(z)\ +\ P^{-1}N^{-1}\operatorname{tr}E[\Bigg((zI+\hat{\Psi}_{T,t})^{-1} \\
&-\ \frac{1}{NT}(zI+\hat{\Psi}_{T,t})^{-1}S_t'(I+C_T)^{-1}S_t(zI+\hat{\Psi}_{T,t})^{-1}\Bigg)S_t'S_t\Psi] \\
&=\ c^{-1}z\xi(z)\ +\ P^{-1}N^{-1}\operatorname{tr}E[(zI+\hat{\Psi}_{T,t})^{-1}S_t'S_t\Psi] \\
&-P^{-1}N^{-1}\operatorname{tr}E[(zI+\hat{\Psi}_{T,t})^{-1}S_t'(I+C_T)^{-1}C_TS_t\Psi] \qquad (273)\\
&=\ c^{-1}z\xi(z)\ +\ P^{-1}N^{-1}\operatorname{tr}E[(zI+\hat{\Psi}_{T,t})^{-1}\Psi^2\operatorname{tr}(\Sigma)] \\
&-P^{-1}N^{-1}\operatorname{tr}E[S_t\Psi(zI+\hat{\Psi}_{T,t})^{-1}S_t'(I+C_T)^{-1}C_T] \\
&\to\ c^{-1}z\xi(z)\ +\ P^{-1}N^{-1}\operatorname{tr}E[(zI+\hat{\Psi}_{T,t})^{-1}\Psi^2\operatorname{tr}(\Sigma)] \\
&-P^{-1}N^{-1}\operatorname{tr}E[\Sigma\operatorname{tr}(\Psi^2(zI+\hat{\Psi}_{T,t})^{-1})(I+C_T)^{-1}C_T] \\
&\approx\ c^{-1}z\xi(z)\ +\ P^{-1}N^{-1}\operatorname{tr}E[(zI+\hat{\Psi}_{T,t})^{-1}\Psi^2\operatorname{tr}(\Sigma)] \\
&-N^{-1}\operatorname{tr}(\Sigma(I+\xi(z)\Sigma)^{-1}\xi(z)\Sigma)P^{-1}E[\operatorname{tr}(\Psi^2(zI+\hat{\Psi}_{T,t})^{-1})]\,,
\end{aligned}
$$

where we have used that

$$E[S_tXS_t']\ =\ \Sigma\operatorname{tr}(\Psi X)$$

for any matrix X independent of S_t, and the fact that $(I+C_T)^{-1}C_T$ is uniformly bounded.

Solving this gives us

$$\lim P^{-1}E[\operatorname{tr}(\Psi^2(zI+\hat{\Psi}_{T,t})^{-1})]\ =\ \frac{\psi_{*,1}-c^{-1}z\xi(z)}{N^{-1}(\operatorname{tr}(\Sigma)-\operatorname{tr}(\Sigma(I+\xi(z)\Sigma)^{-1}\xi(z)\Sigma))}\,, \qquad (274)$$

which is equivalent to

$$\lim P^{-1}E[\operatorname{tr}(\Psi^2(zI+\hat{\Psi}_{T,t})^{-1})]\ =\ \frac{\psi_{*,1}-c^{-1}z\xi(z)}{N^{-1}\operatorname{tr}(\Sigma(I+\xi(z)\Sigma)^{-1}))}\,. \qquad (275)$$

Summarizing, we get

$$
\begin{aligned}
&(NT)^{-1}E[\sum_{t=1}^{T}\hat{\beta}'S_t'R_{t+1}] \\
&=\ (N^2T)^{-1}\operatorname{tr}E[(S_T\beta\beta'S_T'+I)S_T(zI+\hat{\Psi})^{-1}S_T']\ +\ N^{-2}(1-T^{-1})E[\beta'S_T'S_T(zI+\hat{\Psi})^{-1}S_t'S_t\beta] \\
&\approx\xi(z)N^{-1}\operatorname{tr}(\Sigma(I+\xi(z)\Sigma)^{-1})\ +\ b_*\psi_{*,1}\xi(z)N^{-1}\operatorname{tr}(\Sigma^2(I+\xi(z)\Sigma)^{-1}) \\
&+\ \operatorname{tr}(\Sigma/N)\operatorname{tr}(\Sigma(I+\xi(z)\Sigma)^{-1}/N)b_*\frac{\psi_{*,1}-c^{-1}z\xi(z)}{N^{-1}\operatorname{tr}(\Sigma(I+\xi(z)\Sigma)^{-1}))}
\end{aligned} \qquad (276)
$$

□

L Dealing with High-Dimensional Features

Suppose we have a have a high dimensional feature matrix $S \in \mathbb{R}^{T\times P}$ where P is much larger than T. Our goal is to compute

$$\beta(z) \;=\; (S'S/T + zI)^{-1}S'y\,. \tag{277}$$

We start with the observation that

$$T^{-1}SS'v \;=\; \lambda v \tag{278}$$

implies

$$T^{-1}S'SS'v = \lambda S'v \tag{279}$$

and

$$\|S'v\|^2 \;=\; v'SS'v \;=\; T\lambda\,. \tag{280}$$

Thus, if

$$T^{-1}SS' = VDV' \tag{281}$$

we have

$$T^{-1}S'S \;=\; UDU' \tag{282}$$

where

$$U = S'VD^{-1/2}T^{-1/2} \tag{283}$$

and the quasi-inverse will be

$$(T^{-1}S'S + zI)^{-1} = U(D+zI)^{-1}U' = S'VD^{-1/2}T^{-1/2}(D+zI)^{-1}T^{-1/2}D^{-1/2}V'S \tag{284}$$

and we need

$$\begin{aligned}&(T^{-1}S'S + zI)^{-1}S'y = S'VD^{-1/2}T^{-1/2}(D+zI)^{-1}T^{-1/2}D^{-1/2}V'SS'y\\ &\;=\; S'Q\\ &Q \;=\; V(D+zI)^{-1}D^{-1}V'(SS'/T)y \;=\; V(D+zI)^{-1}D^{-1}V'VDV'y \;=\; V(D+zI)^{-1}V'y\end{aligned} \tag{285}$$

Now, we have our algorithm. Suppose we have that $S = (S_1, ..., S_K)$ comes in K blocks, each block being $T \times P_1$ where P_1 is small.

Required: shared object $\Psi \in \mathbb{R}^{T\times T}$.

For parallel loop through $k = 1, \cdots, K$:

Generate S_k. Compute $S_k S_k' \in \mathbb{R}^{T\times T}$. Add it to Ψ
Get the eigenvalue decomposition of the small matrix $\Psi = SS' = VDV'$.
Now we compute

$$Q = VD^{-1/2}(D+zI)^{-1})D^{-1/2}V'(SS')y \ \in \ \mathbb{R}^T$$

[Now comes the cost. We need to repeat this loop one more time.]
For parallel loop through $k = 1, \cdots, K$:
Generate S_k. compute $q_k = S_k' Q \in \mathbb{R}^{P_1}$. [and its number k!!! This is important we need to keep track of the identity].
Stack q_k in one vector. We are done.

M Inverting and Eigen-Decomposing the giant Matrix

So, we are buildling

$$\Psi_k \ = \ \sum_{\kappa=1}^{k} S_k S_k', \ \Psi = SS' = \sum_{\kappa=1}^{K} S_k S_k' \tag{286}$$

The key insight is the identity

$$\Psi_{k+1} \ = \ \Psi_k \ + \ S_k S_k' \tag{287}$$

now, let

$$\Psi_k \ = \ V_k D V_k' \tag{288}$$

where $V_k \in \mathbb{R}^{T\times\nu}$ where ν is the number of top eigenvalues we keep, and $D \in \mathbb{R}^{\nu\times\nu}$ is diagonal. Then, $V_k' V_k = I$, and

$$V_k'\Psi_{k+1}V_k \ = \ V_k'\Psi_k V_k \ + \ V_k' S_k S_k' V_k \ = \ D \ + \ V_k' S_k S_k' V_k \tag{289}$$

Now, we do the eigenvalue decomposition of $D + V_k' S_k S_k' V_k$ and this gives us $\lambda_1(k), \cdots, \lambda_\nu(k)$. Now, we do the orthogonal complement

$$(I-V_kV_k')\Psi_{k+1}(I-V_kV_k') \ = \ (I-V_kV_k')V_kDV_k'(I-V_kV_k') + (I-V_kV_k')S_kS_k'(I-V_kV_k') \ = \ (I-V_kV_k')S_kS_k'(I-V_k\tag{290}$$

Note however that $S_k \in \mathbb{R}^{T\times P_1}$ where P_1 is small and hence we can rather work with

$$\Gamma_k \ = \ S_k'(I-V_kV_k')(I-V_kV_k')S_k \ = \ S_k'(I-V_kV_k')S_k \tag{291}$$

It is a small $P_1 \times P_1$ matrix, whose eigenvalues $\mu_1, \cdots, \mu_{P_1}$ are trivial to compute. The corresponding eigenvector matrix is $\tilde{W} \in \mathbb{R}^{P_1\times P_1}$, and then defining $\tilde{S}_k = (I - V_k V_k')S_k$, we get defining $\tilde{D}$ =diag of eigenvalues of Γ_k, we get

$$W_k \ = \tag{292}$$

We will also use the Using the Woodbury matrix identity, we get

$$(\Psi_{k+1}+zI)^{-1} \;=\; (\Psi_k+zI)^{-1} \;-\; (\Psi_k+zI)^{-1}S_k(I+S_k'(\Psi_k+zI)^{-1}S_k)^{-1}S_k'(\Psi_k+zI)^{-1} \quad (293)$$

Thus, eigenvalues solve

$$x + S_k'(\Psi_k + zI)^{-1}S_k x \;=\; 0$$

for some x. Multiplying by S_k, we get

$$(I + S_k S_k'(\Psi_k + zI)^{-1})y = 0$$

Thus, the matrix

N z and $z*$

TODO: Put here the grid of all z and the way we approximate z* in practice.

O Additional Results

In Tables 4 to 3, we show the performance of the benchmark models (Linear, EW) using long-short deciles portfolio.

For both *Linear* and *Complex* models, we select the shrinkage parameters z ex-post by selecting the one producing the highest Sharpe ratio for the H-L portfolio.

For all models m we show the average returns, SD, Sharpe ratio, and α and T-values (T_α) of the following regression: $R_t^{(m)} = \alpha + \beta_1 R_t^{(m)} + \beta_2 * R_t^{(m)}$

P Features Definition

In Table 12 we list all the stock-level signals used in the empirical section.

Table 4: mega (please see definition at the beginning of the section. (part 1/2)

		Low(L)	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	High(H)	H-L
Linear	Avg	0.02	0.03	0.07	0.06	0.08	0.10	0.12	0.13	0.16	0.19	0.17
	SD	0.62	0.62	0.62	0.63	0.62	0.62	0.65	0.66	0.67	0.69	0.48
	SR	0.09	0.17	0.38	0.33	0.44	0.57	0.62	0.70	0.83	0.94	1.24
	α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	T_α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EW	Avg	-0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	SD	0.06	0.06	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.05
	SR	-0.02	0.22	0.27	0.38	0.44	0.39	0.53	0.58	0.70	0.71	0.65
	α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	T_α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P=100	Avg	0.05	0.05	0.07	0.08	0.09	0.10	0.11	0.12	0.12	0.15	0.10
	SD	0.64	0.63	0.63	0.63	0.60	0.64	0.63	0.66	0.66	0.70	0.53
	SR	0.25	0.30	0.36	0.46	0.53	0.53	0.63	0.65	0.65	0.74	0.67
	α	0.03	0.02	0.00	0.02	0.02	0.01	0.01	0.00	-0.02	-0.02	-0.04
	T_α	3.42	2.58	0.33	2.83	2.15	0.61	1.15	0.20	-2.13	-2.13	-2.22
P=1000	Avg	0.02	0.06	0.08	0.08	0.09	0.09	0.11	0.11	0.14	0.18	0.16
	SD	0.63	0.62	0.62	0.61	0.62	0.62	0.64	0.65	0.69	0.70	0.50
	SR	0.09	0.32	0.42	0.43	0.52	0.52	0.57	0.61	0.71	0.90	1.13
	α	0.00	0.03	0.01	0.02	0.02	0.00	-0.00	-0.01	-0.01	0.00	0.03
	T_α	0.29	3.16	1.72	2.21	2.02	0.43	-0.09	-0.74	-1.26	0.48	1.92
P=2000	Avg	0.02	0.06	0.06	0.07	0.09	0.09	0.12	0.13	0.14	0.18	0.17
	SD	0.63	0.62	0.62	0.60	0.62	0.62	0.64	0.67	0.68	0.70	0.52
	SR	0.10	0.33	0.32	0.39	0.52	0.51	0.62	0.66	0.72	0.90	1.10
	α	0.00	0.03	-0.00	0.01	0.02	0.00	0.01	0.00	-0.01	0.01	0.02
	T_α	0.41	3.33	-0.62	1.36	2.03	0.18	1.01	0.31	-1.05	0.57	1.61
P=5000	Avg	0.01	0.06	0.06	0.06	0.09	0.11	0.11	0.13	0.15	0.19	0.18
	SD	0.65	0.61	0.62	0.61	0.62	0.63	0.63	0.68	0.68	0.72	0.58
	SR	0.05	0.31	0.31	0.32	0.48	0.60	0.59	0.68	0.76	0.92	1.07
	α	-0.00	0.02	-0.01	-0.00	0.01	0.02	0.00	0.01	-0.00	0.01	0.02
	T_α	-0.48	2.90	-0.75	-0.06	1.08	2.08	0.20	0.72	-0.16	0.89	1.12

Table 5: mega (part 2/2)

		Low(L)	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	High(H)	H-L
P=10000	Avg	0.02	0.05	0.05	0.06	0.08	0.10	0.12	0.13	0.14	0.20	0.18
	SD	0.64	0.63	0.62	0.60	0.60	0.63	0.63	0.68	0.68	0.72	0.54
	SR	0.08	0.30	0.30	0.37	0.48	0.57	0.63	0.67	0.71	0.95	1.16
	α	0.00	0.02	-0.01	0.01	0.01	0.01	0.01	0.00	-0.01	0.01	0.03
	T_α	0.13	2.53	-0.94	0.88	1.07	1.38	1.17	0.45	-1.30	1.43	2.20
P=20000	Avg	0.02	0.05	0.06	0.07	0.07	0.09	0.11	0.14	0.15	0.19	0.17
	SD	0.64	0.62	0.63	0.60	0.61	0.61	0.63	0.67	0.68	0.72	0.54
	SR	0.10	0.26	0.33	0.43	0.42	0.53	0.61	0.70	0.74	0.93	1.11
	α	0.00	0.01	-0.00	0.02	-0.00	0.00	0.01	0.01	-0.01	0.01	0.03
	T_α	0.39	1.73	-0.29	2.19	-0.10	0.50	0.76	1.22	-0.67	1.05	1.74
P=50000	Avg	0.01	0.05	0.06	0.07	0.08	0.09	0.11	0.14	0.16	0.19	0.18
	SD	0.65	0.62	0.62	0.60	0.60	0.63	0.64	0.66	0.69	0.72	0.55
	SR	0.05	0.27	0.35	0.43	0.44	0.50	0.59	0.71	0.78	0.92	1.14
	α	-0.01	0.02	0.00	0.02	0.00	0.00	0.00	0.01	0.00	0.01	0.03
	T_α	-0.52	2.10	0.16	2.05	0.18	0.11	0.22	1.40	0.36	0.90	1.92
P=100000	Avg	0.01	0.05	0.06	0.07	0.08	0.09	0.11	0.13	0.15	0.19	0.18
	SD	0.65	0.62	0.63	0.59	0.61	0.62	0.64	0.67	0.69	0.72	0.55
	SR	0.06	0.26	0.36	0.40	0.46	0.53	0.57	0.69	0.78	0.94	1.16
	α	-0.00	0.01	0.00	0.01	0.01	0.01	-0.00	0.01	0.00	0.01	0.03
	T_α	-0.37	1.79	0.26	1.55	0.64	0.61	-0.13	0.94	0.25	1.28	2.11
P=250000	Avg	0.01	0.04	0.07	0.06	0.09	0.10	0.10	0.13	0.16	0.19	0.18
	SD	0.65	0.62	0.63	0.59	0.62	0.62	0.64	0.65	0.70	0.72	0.55
	SR	0.07	0.23	0.38	0.34	0.50	0.55	0.55	0.71	0.78	0.93	1.13
	α	-0.00	0.01	0.01	0.00	0.01	0.01	-0.00	0.01	0.00	0.01	0.03
	T_α	-0.26	1.18	0.88	0.20	1.49	0.93	-0.45	1.38	0.19	1.14	1.79

Table 6: large (part 1/2)

		Low(L)	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	High(H)	H-L
Linear	Avg	0.00	0.02	0.05	0.07	0.08	0.11	0.12	0.13	0.16	0.19	0.19
	SD	0.62	0.62	0.61	0.63	0.63	0.64	0.63	0.64	0.66	0.68	0.46
	SR	0.00	0.10	0.29	0.39	0.42	0.59	0.64	0.70	0.83	0.98	1.44
	α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	T_α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EW	Avg	-0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	SD	0.08	0.07	0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.05
	SR	-0.06	0.20	0.32	0.39	0.48	0.53	0.60	0.66	0.67	0.81	0.78
	α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	T_α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P=100	Avg	0.03	0.05	0.07	0.08	0.07	0.09	0.10	0.13	0.13	0.17	0.14
	SD	0.63	0.62	0.62	0.62	0.61	0.62	0.63	0.64	0.67	0.69	0.46
	SR	0.17	0.27	0.37	0.46	0.42	0.50	0.55	0.69	0.67	0.87	1.09
	α	0.03	0.03	0.01	0.02	-0.00	-0.01	-0.01	0.00	-0.02	-0.01	-0.01
	T_α	3.72	3.98	2.01	2.18	-0.17	-1.35	-1.53	0.59	-2.61	-1.24	-0.47
P=1000	Avg	0.01	0.04	0.05	0.07	0.09	0.09	0.11	0.12	0.15	0.19	0.17
	SD	0.63	0.61	0.61	0.61	0.61	0.62	0.63	0.65	0.66	0.70	0.43
	SR	0.06	0.22	0.26	0.42	0.49	0.52	0.60	0.65	0.81	0.92	1.40
	α	0.01	0.02	-0.01	0.01	0.01	-0.01	-0.00	-0.00	0.01	-0.00	0.03
	T_α	1.42	2.92	-0.76	1.13	1.83	-0.75	-0.05	-0.59	1.04	-0.15	2.90
P=2000	Avg	0.01	0.04	0.05	0.06	0.08	0.11	0.11	0.13	0.14	0.20	0.19
	SD	0.63	0.60	0.61	0.61	0.62	0.62	0.63	0.64	0.66	0.70	0.43
	SR	0.06	0.20	0.28	0.37	0.45	0.59	0.60	0.69	0.75	0.97	1.51
	α	0.01	0.02	-0.00	-0.00	0.00	0.01	-0.00	0.00	-0.00	0.01	0.05
	T_α	1.28	2.32	-0.41	-0.18	0.72	0.92	-0.17	0.54	-0.56	1.18	4.17
P=5000	Avg	0.01	0.04	0.06	0.07	0.09	0.10	0.10	0.12	0.15	0.19	0.18
	SD	0.62	0.61	0.60	0.61	0.61	0.62	0.63	0.64	0.66	0.70	0.39
	SR	0.07	0.20	0.32	0.42	0.51	0.54	0.54	0.67	0.78	0.93	1.55
	α	0.01	0.02	0.00	0.01	0.01	-0.00	-0.01	0.00	0.00	0.00	0.06
	T_α	1.58	2.38	0.68	1.04	2.10	-0.42	-1.68	0.16	0.36	0.26	5.16

Table 7: large (part 2/2)

		Low(L)	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	High(H)	H-L
P=10000	Avg	0.01	0.02	0.05	0.07	0.08	0.09	0.11	0.13	0.15	0.20	0.20
	SD	0.64	0.60	0.60	0.61	0.61	0.61	0.63	0.64	0.67	0.72	0.46
	SR	0.03	0.14	0.26	0.42	0.46	0.52	0.58	0.73	0.79	0.99	1.50
	α	0.01	0.00	-0.00	0.01	0.01	-0.01	-0.01	0.01	0.00	0.01	0.05
	T_α	0.71	0.70	-0.73	1.22	1.06	-0.80	-0.76	1.52	0.46	1.57	4.07
P=20000	Avg	0.00	0.03	0.05	0.07	0.08	0.09	0.11	0.14	0.15	0.20	0.20
	SD	0.64	0.61	0.59	0.62	0.61	0.62	0.63	0.64	0.67	0.71	0.45
	SR	0.00	0.16	0.29	0.38	0.47	0.52	0.58	0.74	0.79	0.99	1.55
	α	-0.00	0.01	0.00	0.00	0.01	-0.01	-0.01	0.01	0.00	0.01	0.06
	T_α	-0.04	1.23	0.07	0.05	1.15	-0.78	-0.73	1.83	0.54	1.60	4.64
P=50000	Avg	0.00	0.03	0.06	0.08	0.08	0.09	0.11	0.14	0.15	0.19	0.18
	SD	0.63	0.60	0.60	0.60	0.60	0.61	0.63	0.65	0.67	0.71	0.40
	SR	0.02	0.20	0.36	0.44	0.44	0.50	0.58	0.73	0.78	0.92	1.60
	α	0.00	0.02	0.01	0.01	0.00	-0.01	-0.01	0.01	0.00	0.00	0.07
	T_α	0.39	2.26	1.67	1.67	0.37	-1.47	-0.77	1.49	0.53	0.03	5.89
P=100000	Avg	0.01	0.04	0.06	0.07	0.08	0.09	0.11	0.12	0.14	0.18	0.17
	SD	0.63	0.60	0.60	0.60	0.61	0.62	0.63	0.65	0.67	0.70	0.37
	SR	0.06	0.26	0.34	0.43	0.48	0.49	0.61	0.67	0.75	0.91	1.62
	α	0.01	0.03	0.01	0.01	0.01	-0.01	-0.00	0.00	-0.00	0.00	0.08
	T_α	1.25	3.51	1.25	1.34	1.38	-1.58	-0.08	0.08	-0.27	0.01	6.42
P=250000	Avg	0.00	0.04	0.06	0.07	0.08	0.09	0.11	0.14	0.15	0.19	0.19
	SD	0.63	0.60	0.60	0.61	0.61	0.61	0.63	0.65	0.67	0.71	0.40
	SR	0.00	0.22	0.36	0.38	0.43	0.50	0.62	0.73	0.78	0.94	1.65
	α	-0.00	0.02	0.01	0.00	0.00	-0.01	0.00	0.01	0.00	0.00	0.07
	T_α	-0.04	2.62	1.75	0.06	0.20	-1.43	0.15	1.50	0.47	0.52	6.21

Table 8: small (part 1/2)

		Low(L)	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	High(H)	H-L
Linear	Avg	-0.03	0.01	0.03	0.05	0.07	0.08	0.12	0.14	0.15	0.20	0.23
	SD	0.54	0.56	0.56	0.57	0.58	0.60	0.61	0.62	0.63	0.68	0.40
	SR	-0.17	0.08	0.17	0.31	0.39	0.47	0.65	0.76	0.81	1.03	1.98
	α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	T_α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EW	Avg	-0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02
	SD	0.09	0.08	0.07	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.06
	SR	-0.18	0.07	0.21	0.35	0.47	0.62	0.67	0.69	0.79	0.87	0.92
	α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	T_α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P=100	Avg	0.02	0.04	0.05	0.04	0.07	0.08	0.09	0.11	0.14	0.18	0.16
	SD	0.57	0.56	0.57	0.57	0.58	0.59	0.60	0.61	0.64	0.68	0.37
	SR	0.11	0.26	0.28	0.26	0.40	0.48	0.50	0.62	0.77	0.90	1.48
	α	0.04	0.03	0.02	-0.01	0.00	-0.00	-0.02	-0.02	-0.00	-0.01	0.01
	T_α	6.37	4.84	2.99	-1.16	0.18	-0.14	-4.15	-2.83	-0.42	-1.55	0.62
P=1000	Avg	0.01	0.03	0.03	0.05	0.06	0.08	0.09	0.11	0.15	0.20	0.19
	SD	0.57	0.56	0.56	0.57	0.57	0.58	0.59	0.62	0.64	0.70	0.38
	SR	0.03	0.20	0.20	0.34	0.39	0.45	0.55	0.63	0.78	0.99	1.78
	α	0.03	0.02	0.00	0.00	-0.00	-0.01	-0.01	-0.02	-0.00	0.00	0.04
	T_α	4.50	3.22	0.66	0.79	-0.24	-1.23	-2.26	-2.74	-0.05	0.48	3.32
P=2000	Avg	0.01	0.03	0.04	0.06	0.06	0.08	0.09	0.11	0.14	0.20	0.19
	SD	0.57	0.56	0.57	0.57	0.57	0.58	0.59	0.61	0.64	0.69	0.35
	SR	0.03	0.17	0.23	0.35	0.39	0.46	0.54	0.65	0.76	0.99	1.91
	α	0.03	0.02	0.01	0.01	-0.00	-0.01	-0.02	-0.01	-0.00	0.00	0.05
	T_α	4.57	2.58	1.50	1.24	-0.14	-0.87	-2.75	-2.28	-0.67	0.54	4.71
P=5000	Avg	0.01	0.03	0.05	0.06	0.07	0.07	0.09	0.11	0.13	0.19	0.18
	SD	0.57	0.56	0.56	0.58	0.57	0.58	0.58	0.61	0.63	0.68	0.33
	SR	0.07	0.16	0.29	0.37	0.40	0.44	0.55	0.63	0.72	0.97	1.91
	α	0.04	0.01	0.02	0.01	0.00	-0.01	-0.02	-0.02	-0.01	0.00	0.06
	T_α	5.23	2.32	3.05	1.71	0.23	-1.30	-2.70	-2.46	-1.79	0.19	5.40

Table 9: small (part 2/2)

		Low(L)	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	High(H)	H-L
P=10000	Avg	-0.00	0.02	0.04	0.05	0.05	0.07	0.09	0.12	0.16	0.21	0.21
	SD	0.58	0.56	0.56	0.56	0.57	0.58	0.59	0.62	0.64	0.71	0.40
	SR	-0.01	0.15	0.23	0.32	0.32	0.45	0.54	0.65	0.85	1.02	1.83
	α	0.03	0.01	0.01	0.00	-0.01	-0.01	-0.02	-0.01	0.01	0.01	0.04
	T_α	3.51	1.98	1.45	0.39	-1.88	-1.11	-2.53	-2.16	1.82	1.26	3.68
P=20000	Avg	0.01	0.03	0.04	0.05	0.07	0.07	0.09	0.11	0.14	0.19	0.19
	SD	0.58	0.56	0.56	0.58	0.56	0.57	0.59	0.62	0.64	0.69	0.34
	SR	0.04	0.21	0.25	0.31	0.42	0.43	0.56	0.64	0.77	0.98	1.89
	α	0.03	0.02	0.01	0.00	0.00	-0.01	-0.01	-0.01	-0.00	0.00	0.06
	T_α	4.52	3.31	1.93	0.04	0.73	-1.60	-2.29	-2.21	-0.59	0.25	5.26
P=50000	Avg	0.00	0.02	0.04	0.04	0.06	0.08	0.10	0.11	0.15	0.21	0.20
	SD	0.58	0.56	0.55	0.57	0.57	0.57	0.59	0.62	0.65	0.70	0.38
	SR	0.01	0.15	0.27	0.27	0.34	0.48	0.56	0.63	0.83	1.01	1.88
	α	0.03	0.01	0.01	-0.01	-0.01	-0.00	-0.01	-0.02	0.01	0.01	0.05
	T_α	3.87	2.00	2.48	-1.07	-1.62	-0.14	-2.23	-2.75	1.22	1.10	4.31
P=100000	Avg	0.00	0.02	0.05	0.05	0.06	0.07	0.10	0.11	0.14	0.20	0.19
	SD	0.58	0.56	0.55	0.57	0.57	0.57	0.59	0.61	0.64	0.70	0.35
	SR	0.03	0.14	0.30	0.32	0.36	0.44	0.58	0.64	0.77	0.99	1.91
	α	0.03	0.01	0.02	0.00	-0.01	-0.01	-0.01	-0.01	-0.00	0.00	0.06
	T_α	4.26	1.72	3.41	0.35	-1.04	-1.47	-1.66	-2.34	-0.35	0.63	5.26
P=250000	Avg	0.00	0.03	0.04	0.05	0.07	0.07	0.09	0.11	0.14	0.19	0.19
	SD	0.58	0.56	0.56	0.56	0.57	0.58	0.59	0.61	0.64	0.69	0.34
	SR	0.03	0.21	0.26	0.33	0.41	0.44	0.54	0.64	0.75	0.98	1.95
	α	0.03	0.02	0.01	0.00	0.00	-0.01	-0.02	-0.01	-0.01	0.00	0.07
	T_α	4.26	3.38	2.17	0.63	0.54	-1.44	-2.79	-2.10	-1.05	0.31	5.79

Table 10: micro (part 1/2)

		Low(L)	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	High(H)	H-L
Linear	Avg	-0.07	-0.03	-0.00	0.02	0.05	0.06	0.09	0.11	0.15	0.21	0.28
	SD	0.43	0.44	0.45	0.47	0.48	0.47	0.49	0.51	0.53	0.59	0.36
	SR	-0.56	-0.24	-0.00	0.18	0.34	0.45	0.62	0.77	0.97	1.25	2.71
	α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	T_α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EW	Avg	-0.01	-0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.02	0.03
	SD	0.10	0.08	0.08	0.07	0.06	0.06	0.05	0.05	0.05	0.05	0.07
	SR	-0.37	-0.08	0.03	0.22	0.37	0.57	0.66	0.78	0.90	1.13	1.32
	α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	T_α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P=100	Avg	-0.00	0.02	0.02	0.04	0.04	0.06	0.07	0.09	0.10	0.15	0.15
	SD	0.48	0.49	0.48	0.47	0.48	0.47	0.47	0.48	0.48	0.52	0.27
	SR	-0.00	0.14	0.18	0.31	0.31	0.42	0.52	0.63	0.75	0.97	1.87
	α	0.07	0.05	0.02	0.02	-0.00	-0.01	-0.01	-0.02	-0.03	-0.03	-0.00
	T_α	11.75	8.14	5.06	3.44	-0.48	-1.19	-2.64	-4.09	-5.25	-5.83	-0.30
P=1000	Avg	-0.02	0.01	0.03	0.03	0.05	0.06	0.07	0.10	0.11	0.17	0.19
	SD	0.48	0.47	0.47	0.46	0.46	0.47	0.47	0.48	0.50	0.54	0.27
	SR	-0.15	0.04	0.19	0.23	0.35	0.44	0.51	0.69	0.79	1.07	2.41
	α	0.05	0.03	0.03	0.01	0.00	-0.00	-0.01	-0.01	-0.02	-0.02	0.03
	T_α	9.02	7.15	4.64	1.32	0.65	-0.54	-3.09	-1.82	-3.74	-3.85	2.76
P=2000	Avg	-0.01	0.00	0.02	0.04	0.05	0.06	0.08	0.09	0.11	0.16	0.17
	SD	0.47	0.46	0.47	0.46	0.47	0.47	0.48	0.48	0.50	0.54	0.25
	SR	-0.11	0.01	0.18	0.28	0.38	0.46	0.55	0.65	0.75	1.03	2.43
	α	0.06	0.03	0.02	0.01	0.01	0.00	-0.01	-0.01	-0.03	-0.03	0.03
	T_α	9.65	5.99	5.16	2.53	1.32	0.10	-1.68	-2.87	-5.49	-4.91	3.74
P=5000	Avg	-0.02	0.00	0.02	0.04	0.04	0.06	0.07	0.10	0.11	0.16	0.18
	SD	0.48	0.46	0.47	0.46	0.46	0.47	0.47	0.48	0.50	0.54	0.25
	SR	-0.14	0.03	0.16	0.27	0.34	0.47	0.54	0.70	0.77	1.04	2.58
	α	0.05	0.03	0.02	0.01	0.00	0.00	-0.01	-0.01	-0.02	-0.03	0.05
	T_α	8.61	6.64	4.32	2.45	0.19	0.46	-1.71	-1.40	-4.52	-4.69	5.28

Table 11: micro (part 2/2)

		Low(L)	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	High(H)	H-L
P=10000	Avg	-0.03	-0.01	-0.00	0.03	0.05	0.06	0.07	0.10	0.13	0.20	0.23
	SD	0.48	0.47	0.46	0.46	0.47	0.47	0.47	0.48	0.51	0.56	0.31
	SR	-0.23	-0.05	-0.00	0.19	0.36	0.43	0.55	0.71	0.91	1.22	2.59
	α	0.04	0.02	-0.00	0.00	0.00	-0.00	-0.01	-0.01	-0.00	0.00	0.04
	T_α	6.85	4.64	-0.06	0.34	0.76	-0.95	-1.70	-1.18	-0.49	0.13	3.90
P=20000	Avg	-0.03	-0.01	0.01	0.03	0.05	0.05	0.07	0.10	0.13	0.19	0.22
	SD	0.48	0.47	0.46	0.46	0.47	0.47	0.47	0.48	0.50	0.57	0.29
	SR	-0.22	-0.06	0.08	0.22	0.37	0.41	0.55	0.69	0.90	1.17	2.65
	α	0.04	0.02	0.01	0.01	0.01	-0.01	-0.01	-0.01	-0.00	-0.01	0.05
	T_α	7.07	4.20	2.25	1.08	0.99	-1.24	-1.73	-1.78	-0.81	-1.30	4.99
P=50000	Avg	-0.04	-0.01	-0.00	0.02	0.04	0.05	0.08	0.10	0.13	0.21	0.24
	SD	0.48	0.47	0.46	0.46	0.46	0.47	0.47	0.49	0.50	0.57	0.31
	SR	-0.26	-0.05	-0.01	0.17	0.30	0.40	0.56	0.73	0.92	1.27	2.71
	α	0.04	0.02	-0.00	-0.00	-0.00	-0.01	-0.01	-0.00	-0.00	0.01	0.05
	T_α	6.31	4.81	-0.25	-0.27	-0.61	-1.38	-1.29	-0.46	-0.07	1.58	5.17
P=100000	Avg	-0.04	-0.01	0.00	0.02	0.04	0.06	0.07	0.10	0.13	0.21	0.24
	SD	0.49	0.47	0.46	0.46	0.46	0.46	0.47	0.49	0.50	0.57	0.30
	SR	-0.25	-0.07	0.03	0.18	0.29	0.43	0.54	0.72	0.93	1.25	2.75
	α	0.04	0.02	0.00	0.00	-0.00	-0.00	-0.01	-0.00	0.00	0.01	0.06
	T_α	6.39	4.21	0.83	0.17	-0.85	-0.73	-2.02	-0.65	0.16	1.03	5.63
P=250000	Avg	-0.03	-0.00	0.01	0.02	0.04	0.06	0.07	0.10	0.13	0.20	0.23
	SD	0.49	0.47	0.46	0.46	0.46	0.46	0.47	0.49	0.51	0.57	0.28
	SR	-0.22	-0.04	0.06	0.18	0.33	0.47	0.53	0.68	0.90	1.20	2.79
	α	0.04	0.02	0.01	0.00	0.00	0.00	-0.01	-0.01	-0.00	-0.00	0.07
	T_α	6.75	5.01	1.64	0.14	0.15	0.28	-2.32	-1.94	-0.84	-0.30	6.53

Table 12: Features Definition

Signal	Description	Signal	Description
cowc_gr1a	current operating working capital growth	oaccrals_at	operating accruals
oaccruals_ni	operating accruals to net income	taccruals_at	total accruals to total asset
taccruals_ni	total accruals to net income	capex_abn	abnormal corporate investment
debt_gr3	3 year book debt growth	fnl_gr1a	change in financial liabilities
ncol_gr1a	change in noncurrent operating liabilties	nfna_gr1a	change in net financial assets
noa_at	net operating assets to total asset	aliq_at	liquid asset to total asset
at_gr1	asset growth	be_gr1a	change in book equity
capx_gr1	1 year CAPEX growth	capx_gr2	2 year CAPEX growth
capx_gr3	3 year CAPEX growth	coa_gr1a	change in current operating asset
col_gr1a	change in current operating liabilities	emp_gr1	employment growth
inv_gr1a	inventory growth	lnoa_gr1a	Change in long-term net operating asset
mispricing_mgmt	mispricing factor: management	ncoa_gr1a	change in noncurrent operating asset
noa_gr1a	change in net operating asset	ppeinv_gr1a	change in PPE and inventory
ret_60_12	long-term reversal	sale_gr1	1 year sales growth
sale_gr3	3 year sales growth	sale_gr1q	1 quarter sales growth
seas_2_5_na	years 2-5 lagged returns, nonannual	age	firm age
aliq_mat	liquid asset to market value	at_be	book leverage
bidaskhl_21d	21 day high-low bid-ask spread	cash_at	cash-to-asset
netdebt_me	net debt-to-price	ni_vol	earnings volatility
rd_sale	R&D to sales	rd5_at	R&D to total asset
tangibility	asset tangibility	z_score	Altman Z-score
beta_60m	market beta from 60m rolling regression	beta_simson_21d	realized monthly beta
betabab_1260d	realized 5 year beta	betadown_252d	Downside beta
chcsho_12m	12 month net stock issuance	earnings_variability	earnings variability
eqnetis_at	net equity issuance to total asset	fcf_me	free cash flow to price
ivol_capm_21d	CAPM idiosyncratic volatility (21 days)	ivol_capm_252d	CAPM idiosyncratic volatility (252 days)
ivol_ff3_21d	FF3F idiosyncratic volatility (21 days)	ivol_hxz4_21d	q-factor model idiosyncratic volatility (21 days)
netis_at	net total issuance to asset	ocfq_saleq_std	cash flow volatility
rmax1_21d	max daily return in 1 month	rmax5_21d	max 5 day return in 1 month
rvol_21d	return volatility in 1 month	turnover_126d	share turnover
zero_trades_126d	number of zero trades (6 months)	zero_trades_21d	number of zero trades (1 months)
zero_trades_252d	number of zero trades (12 months)	prc_highprc_252d	high-low in last year
resff3_12_1	residual momentum t-12 to t-1	resff3_6_1	residual momentum t-6 to t-1
ret_12_1	price mometum t-12 to t-1	ret_3_1	price mometum t-3 to t-1
ret_6_1	price mometum t-6 to t-1	ret_9_1	price mometum t-9 to t-1
seas_1_1_na	1 year lagged return, nonannual	dsale_dinv	sales growth minus inventory growth
dsale_drec	sales growth minus receivable growth	dsale_dsga	sales growth minus SG&A growth
niq_at_chg1	quarterly return to asset growth	niq_be_chg1	quarterly return to equity growth
niq_su	standardized earnings surprise	ocf_at_chg1	operating cash flow to asset growth
qmj_safety	quality minus junk: safety	ret_12_7	price momentum t-12 to t-7
sale_emp_gr1	labor force efficiency change	saleq_su	standardized revenue surprise
seas_1_1an	year 1-lagged return, annual	sti_gr1a	change in short-term investment
dolvol_var_126d	dollar volume variation	ebit_bev	EBIT to enterprise value
ebit_sale	EBIT to sales	f_score	Pitroski F-score
intrinsic_value	intrinsic value to market	ni_be	return to equity
niq_be	quarterly return to equity	o_score	Ohlson O-score
ofc_at	operating cash flow to asset	ope_be	operating profits to book equity
ope_bel1	operating profits to lagged book equity	turnover_var_126d	turnover variation
at_turnover	capital turnover	cop_at	cash operating profit to asset
cop_atl1	cash operating profit to lagged asset	dgp_dsale	gross margin change minus sales change
gp_at	gross profit to asset	gp_atl1	gross profit to lagged asset
mispricing_perf	mispricing factor: performance	ni_inc8q	number of consecutive earnings increase
niq_at	quarterly return to asset	op_at	operating profit to asset
op_atl1	operating profit to lagged asset	opex_at	operating leverage
qmj	quality minus junk	qmj_growth	quality minus junk: growth
qmj_prof	quality minus junk: profitability	sales_bev	asset turnover
tax_gr1a	tax expense surprise	corr_1260d	market correlation
coskew_21d	coskewness	dbnetis_at	net debt issuance
kz_index	Kaplan-Zingales index	lti_gr1a	change in long-term investment
ni_ar1	earnings persistence	pi_nix	taxable income-to-book
seas_11_15an	years 11-15 lagged returns, annual	seas_11_15na	years 11-15 lagged returns, nonannual
seas_16_20an	years 16-20 lagged returns, annual	seas_16_20na	years 16-20 lagged returns, nonannual
seas_6_10an	years 6-10 lagged returns, annual	seas_6_10na	years 6-10 lagged returns, nonannual
seas_2_5_an	years 2-5 lagged returns, annual	ami_126d	Amihud liquidity measure
dolvol_126d	dollar trading volume	market_equity	market equity
prc	price	rd_me	R&D to market
iskew_capm_21d	CAPM idiosyncratic skewness	iskew_ff3_21d	FF3F idiosyncratic skewness
iskew_hxz4_21d	q-factor model idiosyncratic skewness	ret_1_0	short-term reversal
rmax5_rvol_21d	scaled max 5 day return	rskew_21d	total skewness
at_me	asset to market	be_me	book to market
bev_mev	book to market enterprise value	debt_me	debt to market
div12m_me	dividend yield	ebitda_mev	EBITDA to market enterprise value
eq_dur	equity duration	eqnpo_12m	equity net payout
eqnpo_me	net payout yield	eqpo_me	payout yield
ni_me	earnings to price	ocf_me	operaiting cash flow to price
sale_me	sales to price		