Statistical Methods for Asset Pricing and Portfolio Management



Yuantao Shi
The Queen's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
May 2026

Table of Contents

List of Figures **3**

1 Acknowledgment **1**

2 Introduction **1**

3 Virtue of Complexity with Noisy Features **4**
3.1 Introduction . . . 4
3.2 Model . . . 6
3.2.1 Asset dynamics . . . 6
3.2.2 Portfolio: timing strategy and performance evaluation . . . 10
3.2.2.1 Timing strategy . . . 10
3.2.2.2 Out-of-sample R-squared . . . 12
3.3 Ridge regression . . . 12
3.4 Theoretical properties of ridge regression . . . 14
3.4.1 Random matrix theory . . . 14
3.4.2 Asymptotic results . . . 16
3.4.3 Independence among features . . . 18
3.4.4 A note on the convexity of performance measures . . . 20
3.5 Empirical analysis . . . 21
3.5.1 Data . . . 21
3.5.2 Empirical design of noise incorporation . . . 21
3.5.3 Out-of-sample performance . . . 24
3.6 Conclusion . . . 27

4 The Limited Virtue of Complexity in a Noisy World **29**
4.1 Introduction . . . 29
4.1.1 Model misspecification . . . 36
4.1.2 Growing noise along the feature-discovery process . . . 40
4.2 Empirical analysis . . . 43
4.2.1 Data . . . 43
4.2.2 Autoencoder as a noise-reduction tool . . . 43
4.2.3 Out-of-sample performance of market timing strategy . . . 46
4.3 Conclusion . . . 47

5 The Mechanical Virtue of Random Fourier Features and Model Complexity **49**
5.1 Introduction . . . 49

5.2 Model . . . 52
5.2.1 Empirical properties of RFF with $T = 1$ . . . 52
5.2.2 A closer look at the theoretical mechanism of the RFF method . . . 54
5.3 Numerical simulation . . . 58
5.3.1 Scenario 1: pure momentum . . . 59
5.3.2 Scenario 2: pure reversal . . . 59
5.3.3 KMZ dataset with sign-flipped CRSP returns . . . 60
5.3.4 Number of observations $T > 1$ . . . 60
5.4 Conclusion and future work . . . 62

Bibliography **64**

Appendix A Virtue of Complexity with Noisy Features **71**
A.1 Proofs . . . 71
A.1.1 Proof of Proposition 1. . . . 71
A.1.2 Proof of Theorem 1. . . . 77
A.1.3 Proof of Proposition 2. . . . 108

Appendix B The Limited Virtue of Complexity in a Noisy World **116**
B.1 Proofs . . . 116
B.1.1 Proof of Corollary 2. . . . 116
B.1.2 Proof of Corollary 3. . . . 118
B.1.3 Proof of Proposition 3. . . . 121
B.2 Motivating examples . . . 122
B.3 Autoencoder details . . . 123

Appendix C The Mechanical Virtue of Random Fourier Features and Model Complexity **126**
C.1 Proofs . . . 126
C.1.1 Proof of Proposition 4 . . . 126
C.1.2 Proof of Proposition 5 . . . 129
C.1.3 Proof of Proposition 6 . . . 130
C.1.4 Proof of Proposition 7 . . . 131

List of Figures

3.1 Out-of-sample market timing performance for $\log_{10} z = 0$. Training window is $T = 12$ months and RFF count P ranges from 2 to 4,000 with $\gamma = 2$, $\log_{10} z = 0$, and noise variance ranging from 0 to 0.25. 25
3.2 Out-of-sample market timing performance for $\log_{10} z = 3$. . . . 25
3.3 Out-of-sample market timing performance for $\log_{10} z = 4$. . . . 26
3.4 Best out-of-sample market timing performance for adding noise to the data. 26
3.5 Best out-of-sample market timing performance for adding noise to the feature. 27

4.1 Prediction of stock returns using simulated features that are observed without noise. The x-axis shows the number of features, the y-axis is the out-of-sample R-squared. The results are averaged over 50 repetitions. 31
4.2 Stock return prediction with a neural network with data from NYSE, AMEX, and Nasdaq from 1990 to 2024. The x-axis shows model complexity (i.e., number of neurons), and the y-axis depicts out-of-sample R-squared. The results are averaged over 20 repetitions. 32
4.3 Complexity and model performance with simulated features observed with noise. In the left-hand panel, the noise level in each feature remains fixed as model complexity increases, whereas in the right-hand panel, the noise level for each feature increases linearly as model complexity increases, and the "noise scale" in the right-hand panel is the magnitude of this linear growth. The results are averaged over 50 repetitions. 33
4.4 Out-of-sample R-squared and Sharpe ratio for $f(x) = 10\,x^{\alpha}$ under different values of α and ridge parameter z. 44
4.5 Out-of-sample R^2 (left panel) and Sharpe ratio (right panel) of the market timing strategy using de-noised features from the autoencoder. Results are averaged across all stocks and five repetitions. Shaded bands represent ± 1 standard error across the five repetitions. 46

5.1 Performance metrics as functions of number of features P, with number of observations $T = 1$, and ridge penalty $\lambda = 1$. . 54

5.2 Performance metrics for scenario 1 as functions of number of features P, with number of observation $T = 1$, and ridge penalty $\lambda = 1$. 59
5.3 Performance metrics for scenario 2 as functions of number of features P, with number of observation $T = 1$, and ridge penalty $\lambda = 1$. 60
5.4 Performance of the KMZ experiment with CRSP returns sign-flipped at odd indices, as functions of number of features P, with number of observation $T = 1$, and ridge penalty $\lambda = 1$. . 61
5.5 Performance metrics for scenario 1 as functions of number of features P, with number of observation $T = 3, 6, 12, 36$, and ridge penalty $\lambda = 1$. 61
5.6 Performance metrics for scenario 2 as functions of number of features P, with number of observation $T = 3, 6, 12, 36$, and ridge penalty $\lambda = 1$. 61
5.7 Performance of the KMZ experiment with CRSP returns sign-flipped at odd indices, as functions of number of features P, with number of observation $T = 3, 6, 12, 36$, and ridge penalty $\lambda = 1$. 62

1 | Acknowledgment

I would like to express my deep gratitude to my supervisor, Álvaro Cartea, for his support and encouragement throughout the past 32 months. His guidance, patience, and unwavering belief in my work have been instrumental during my DPhil journey. I am especially grateful for the freedom he gave me to explore my ideas, and for his genuine enthusiasm and support for the topics I chose to pursue.

I am also deeply grateful to my research collaborator, Qi Jin. His clarity of thought, precision in writing, and strong intuition for empirical work have greatly shaped our research. I am especially thankful for his support in refining ideas and improving our work, and for his friendship, which has made this journey all the more enjoyable.

I am also especially thankful to Matthieu R. Meunier, Sergio Calvo Ordonez, and Francesco Piatti, whose rigor, energy, and openness made our collaborations both productive and enjoyable. It is particularly meaningful to see our joint work contribute to their future, and I look forward to seeing their research continue to grow. I would also like to thank the examiners of my viva, Leandro Sanchez-Betancourt and Ruixun Zhang, for their profound expertise, careful reading, and insightful comments on my thesis.

I owe my deepest gratitude to my parents for their unconditional love and unwavering support throughout my years studying abroad. I am also grateful to my family for their constant encouragement and understanding, which have been a steady source of strength.

I am grateful to the Oxford-Man Institute of Quantitative Finance (OMI), Man Group, and The Queen's College for funding my studies and for their administrative support. I would also like to extend my gratitude to all my colleagues at the OMI for inspiring discussions and encouragement. I am especially thankful to Ruotong Cao, Lei Fang, Zichuan Guo, Huidong Liang, Sergio Calvo Ordonez, Jonathan Plenk, Danni Shi, Shifan Yu, and Jiexiu Zhu. Their friendship and support made my time here far more meaningful.

I would like to thank Sirui Huang, Chongxuan Zhang, Guoyun Zhang, Jiangshan Zhang, and many other friends from my high school and undergraduate

years. It means a lot to me to have shared so much time together in Shanghai, my hometown.

Finally, I would like to acknowledge the artists and creators whose work has been a constant source of inspiration and comfort. The games of Hideo Kojima; the music of Eminem, Nujabes, Low Roar, Shiro Sagisu, Tatsuro Yamashita and Kanno Yoko; the films and anime of Anno Hideaki and Satoshi Kon; and the manga of Tite Kubo, Fujimoto Tatsuki and Gege Akutami have accompanied me through many moments of this journey. Their work has shaped the way I see the world and supported me in ways that are difficult to fully express.

Abstract

Recent advances in machine learning have led to the widespread use of highly complex models in asset return prediction. While a growing literature documents improved out-of-sample performance as model complexity increases, the mechanisms underlying these gains remain an open question. This thesis studies the limits of the virtue of complexity in financial prediction through three related projects. The first develops a high-dimensional model with noisy predictive features, derives asymptotic expressions for prediction and portfolio performance, and shows that under optimal regularization performance is decreasing and convex in the level of noise. The second studies a partial-observation setting and characterizes conditions under which increasing complexity eventually reduces predictive accuracy and portfolio performance, giving rise to an optimal level of complexity and a limited virtue of complexity. The third derives analytical results for random Fourier feature models showing that rising performance with complexity can arise mechanically, even in environments where meaningful learning is implausible. Together, the results show how feature quality, partial observability, and the mechanics of high-dimensional estimators shape the relationship between model complexity and performance.

2 | Introduction

Understanding and predicting asset returns is a central goal in financial economics and quantitative asset management. In recent years, this task has increasingly relied on high-dimensional models and machine-learning methods. These approaches use large sets of predictors and flexible nonlinear transformations to extract predictive signals from financial data [65, 31]. A growing body of work argues that, when combined with appropriate regularization, increasing model complexity can improve out-of-sample predictive performance. This idea is commonly referred to as the "virtue of complexity" [47, 36, 7, 6].

At the same time, the interpretation of predictive success in financial settings remains challenging. Financial predictors are often noisy, persistent, and only weakly informative about future returns. Moreover, performance is typically evaluated not only through statistical measures such as out-of-sample R^2, but also through economic outcomes such as portfolio returns and Sharpe ratios. These features distinguish return prediction from many standard machine-learning problems and motivate a growing debate about whether observed gains from complex models reflect learning or arise from other mechanisms [56, 48].

This thesis studies the limits of the virtue of complexity in return prediction. It brings together three related projects that examine different channels through which model complexity affects predictive and economic performance.

The first part of the thesis focuses on the role of data quality and measurement error. In practice, predictive features are rarely observed without noise, and improving feature quality requires costly investments in data collection, infrastructure, and human resources. When features are measured with error, additional features may contribute more noise than signal. Chapter 3 develops a theoretical framework in which an investor faces noisy features, and shows that, under appropriate regularization, portfolio performance is decreasing and convex in the level of noise, highlighting the value of high-quality features.

The second part of the thesis studies a partial-observation setting in which investors observe only a subset of the true features, and these features may

themselves be noisy. In this environment, increasing complexity introduces a trade-off between incorporating additional predictive information and amplifying measurement error. Chapter 4 shows that this trade-off leads to a *limited virtue of complexity*, where performance may deteriorate beyond an optimal level of complexity.

The third part of the thesis studies a complementary mechanism. It examines environments in which increases in model complexity appear to improve performance even when learning is implausible. Using random Fourier feature representations of predictive regressions, we show that monotonic relationships between model size and out-of-sample performance can arise mechanically from the structure of the estimator itself. These patterns persist even in extreme small-sample settings where the model cannot extract information from the data. Related concerns about mechanical or non-informational sources of predictability have also been raised in recent critiques of high-dimensional return prediction [56, 13].

More specifically, the main mathematical contributions of the thesis are as follows. In Chapter 3, we derive high-dimensional asymptotic expressions for the expected return, leverage, out-of-sample R-squared and Sharpe ratio of a timing strategy with ridge regression when predictive features are observed with measurement error. For the case of independent features with homogeneous noise, we obtain explicit expressions for these quantities, characterize the optimal ridge parameter, and show that the optimally regularized out-of-sample R-squared and Sharpe ratio are decreasing and convex in the noise level. In Chapter 4, we extend this framework to partial observation and derive asymptotic expressions for portfolio performance as a function of the fraction of observed features. Under a specification in which average measurement error increases with model complexity, we characterize the optimal ridge parameter and derive conditions under which there exists an interior optimal level of model complexity. We also derive the corresponding asymptotic characterization for a feature-discovery setting in which later-discovered features are increasingly noisy. In Chapter 5, we derive finite-sample expressions for the expected out-of-sample R-squared and expected timing return in the single-observation setting, together with an asymptotic characterization of the Sharpe ratio as the number of random Fourier features grows. We further show that reversing the signs of the relevant return products reverses the timing return and Sharpe ratio for every number of features, providing a mathematical characterization of the mechanical nature of the observed complexity-performance relationship.

Taken together, the three projects provide a unified perspective on when complex prediction models succeed in financial applications and when they do not. Rather than treating the virtue of complexity as a general principle, this thesis shows that performance gains may reflect either economic information or mechanical properties of high-dimensional estimators. Distinguishing between these forces is crucial for interpreting empirical evidence based on large and flexible forecasting models.

The rest of this thesis is organized as follows: Chapter 3 studies noisy features; Chapter 4 develops the limited virtue of complexity; Chapter 5 analyzes the mechanical virtue of complexity. The appendix contains proofs and details of the empirical settings.

3 | Virtue of Complexity with Noisy Features

3.1 Introduction

[47] (KMZ) explore the relationship between return forecasts and portfolios formed from heavily parameterized models. Their work establishes a theoretical basis for machine learning models in portfolio construction, and they show that portfolio performance strictly increases in model complexity under appropriate shrinkage of ridge regression. This result is known as the virtue of complexity. We extend the work of KMZ by considering a setting closer to the real world where the feature matrix is observed with noise.

To provide a gentle introduction to the main ideas and notation, we first present the model informally and defer the precise dimensions of the variables and the formal probabilistic assumptions to Section 3.2.

As in KMZ, consider a true predictive model of the form

$$R_{t+1} = f(G_t) + \epsilon_t \,, \tag{3.1}$$

where R_t is the return of a single asset, G_t is a fixed class of predictive signals, and f is a continuous function. Next, use the universal approximation theorem (see [40]) to approximate f as a wide neural network with P neurons

$$f(G_t) = \sum_{i=1}^{P} F_{i,t}\, \beta_i \,, \tag{3.2}$$

where $F_{i,t} = g(w_i^\intercal\, G_t)$ is the output of the ith neuron, characterized by a fixed non-linear activation function g and learnable weight vector w_i. The term $w_i^\intercal\, G_t$ is the inner product between the weight vector w_i and the predictive signals G_t, where the weights w_i determine the influence of each signal on the neuron's activation.

Therefore, one can write the true data generating process of returns R_t as a

classic regression model

$$R_{t+1} = \sum_{i=1}^{P} F_{i,t}\, \beta_i + \epsilon_t \,, \tag{3.3}$$

where $F_{i,t}$ are features, β_i are unobserved parameters, and ϵ_t is innovation.

In KMZ, the authors assume that the investor uses noiseless observations of $F_{i,t}$ to estimate the parameter β_i. However, in practice, quantitative investors are often constrained by their domain expertise, computational resources, and the quality of data, all of which may introduce significant noise to their observation of the feature F_t. In particular, when training neural networks to approximate the true data generation process $f(G_t)$, they will very likely not obtain the best approximation function.

In this chapter, the investors observe $\tilde{F}_t$, which is a noisy version of the true feature F_t. The infrastructure and resources available to the investor determine how close $\tilde{F}_t$ is to F_t. Investors must decide how much to invest in better infrastructure and resources to reduce the noise in the features they observe. This investment could include upgrading computational resources, hiring additional research personnel, or acquiring more sophisticated and alternative data sources. At the core of this decision is the trade-off between the costs associated with these resources and the benefits of obtaining observations $\tilde{F}_t$ that are closer, in some sense, to the true features F_t. For example, a shorter Euclidean distance between $\tilde{F}_t$ and F_t may lead to more accurate estimations of asset returns and, consequently, better portfolio performance.

Throughout this chapter, we assume that the investor observes noisy measurements corresponding to the full set of true features. This allows us to isolate the effect of feature noise on prediction and portfolio performance. In Chapter 4, we relax this full-observation assumption and consider a partial-observation setting in which the investor observes only a subset of the true features, which may themselves be measured with noise.

Our central research question explores this trade-off: **Should investors invest in improving their infrastructures and resources to reduce the noise in the features they observe, thereby enhancing portfolio performance?** We investigate the relationship between the performance of the optimal portfolio under optimal choice of parameters and the level of noise in the observed features. This perspective introduces another facet of

the "virtue of complexity", where "complexity" refers to model sophistication and to the allocation of more computational resources, additional human capital, and higher-quality and alternative data sources.

In practice, sophisticated investors often face significantly higher marginal costs when attempting to enhance their infrastructure compared to the marginal costs faced by less sophisticated investors. This difference arises because advanced strategies and complex systems require greater investments in data acquisition, computational resources, and specialized talent. As a result, even small improvements require substantial resources and these are increasing as the level of sophistication is already higher [2, 12]. However, we observe that the most successful quantitative fund managers are often more incentivized to spend substantial resources on purchasing data from vendors, offering competitive compensations to attract talent, and investing in computing power [19, 10, 12]. These real-world practices exemplify the *noisy virtue of complexity* that we formalize in our analysis.

3.2 Model

3.2.1 Asset dynamics

We consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which all random variables introduced below are defined. For each P, let X_t and Y_t be sequences of $\mathbb{R}^P$-valued random vectors, let ϵ_t be a sequence of real-valued random variables, and let $\beta \in \mathbb{R}^P$ be a random vector of predictive coefficients. The sequences X_t, Y_t, and ϵ_t are each independent and identically distributed over time and are mutually independent. Moreover, β is independent of these three sequences. The distributional and moment conditions imposed on these random variables are stated in Assumption 1 below.

Consider a financial market where the data generating process (DGP) of the **excess** returns of a single asset is given by the predictive model

$$R_{t+1} = F_t^\intercal \, \beta + \epsilon_{t+1} \, . \tag{3.4}$$

Here, $F_t \in \mathbb{R}^{P \times 1}$ is the true feature vector, $\beta \in \mathbb{R}^{P \times 1}$ is a vector of predictive coefficients, P is the number of features, $\intercal$ is the transpose operator, and $\epsilon_{t+1} \in \mathbb{R}$ is the innovation in returns.

The investors use historical data to estimate the predictive coefficients β and

to construct optimal portfolios. Unlike the literature, which assumes that the investor observes the true feature matrix F_t, here, the investors observe a noisy version $\tilde{F}_t$ of the true feature matrix due to imperfections in the investor's infrastructure. Therefore, the investors use the model

$$R_{t+1} = \tilde{F}_t^\intercal \beta + \tilde{\epsilon}_{t+1}\,, \tag{3.5}$$

where $\tilde{\epsilon}_{t+1} = \epsilon_{t+1} + F_t^\intercal \beta - \tilde{F}_t^\intercal \beta$. Let $T \in \mathbb{Z}$ be the length of the look-back window. Under (3.5) the investors use $\left\{R_{\tau+1}, \tilde{F}_\tau\right\}_{\tau=t-T}^{t-1}$, instead of $\{R_{\tau+1}, F_\tau\}_{\tau=t-T}^{t-1}$, to estimate the excess returns R_{t+1}.

Below, we introduce a series of technical assumptions that ensure the analytical tractability of the model we introduced above.

Assumption 1.

(a) The error term ϵ_{t+1} in (3.4) is independent and identically distributed (i.i.d.), and the moments $\mathbb{E}[\epsilon_{t+1}] = \mathbb{E}[\epsilon_{t+1}^3] = 0$, $\mathbb{E}[\epsilon_{t+1}^2] = 1$, and $\mathbb{E}[\epsilon_{t+1}^4] < \infty$.

(b) There exists a random vector $X_t \in \mathbb{R}^P$ and a symmetric P-dimensional positive semidefinite matrix Ψ such that the true feature matrix F_t satisfies

$$F_t = \Psi^{1/2} X_t\,,$$

where X_t is i.i.d. for all t. Furthermore, the ith and jth components of X_t, denoted by $X_{i,t}$ and $X_{j,t}$, are independent, with $\mathbb{E}[X_{i,t}] = \mathbb{E}\left[X_{i,t}^3\right] = 0$, $\mathbb{E}\left[X_{i,t}^2\right] = 1$, and the fourth moments $\mathbb{E}\left[X_{i,t}^4\right]$ are uniformly bounded, i.e., there exists a constant $K_X < \infty$ such that

$$\sup_{P \geq 1} \sup_{1 \leq i \leq P} \sup_{t \in \mathbb{Z}} \mathbb{E}\left[|X_{i,t}|^4\right] \leq K_X.$$

(c) There exists a random vector $Y_t \in \mathbb{R}^P$ and a symmetric P-dimensional positive semidefinite matrix Φ such that the noisy feature matrix $\tilde{F}_t$ satisfies

$$\tilde{F}_t = \Psi^{1/2} X_t + \Phi^{1/2} Y_t\,,$$

where Y_t is i.i.d. for all t and independent of X_t. Furthermore, $Y_{i,t}$ and $Y_{j,t}$ are independent, with $\mathbb{E}[Y_{i,t}] = \mathbb{E}\left[Y_{i,t}^3\right] = 0$, $\mathbb{E}\left[Y_{i,t}^2\right] = 1$, and the fourth moments $\mathbb{E}\left[Y_{i,t}^4\right]$ are uniformly bounded, i.e., there exists a constant $K_Y < \infty$

such that

$$\sup_{P\geq 1}\sup_{1\leq i\leq P}\sup_{t\in\mathbb{Z}}\mathbb{E}\left[|Y_{i,t}|^4\right]\leq K_Y.$$

(d) The true predictive coefficients $\beta=(\beta_i)_{i=1}^P\in\mathbb{R}^P$ *are arranged in a vector whose coordinates* β_i *are i.i.d. and independent of* X_t*,* Y_t*, and* ϵ_t*. Additionally,* $\mathbb{E}[\beta]=\mathbf{0}\in\mathbb{R}^P$*, and* $\mathbb{E}\left[\beta\,\beta^\intercal\right]=P^{-1}\,b_{*,P}\,I_P$ *for some constant* $b_{*,P}=\mathbb{E}\left[||\beta||^2\right]$*, where* $b_{*,P}\xrightarrow{\mathbb{P}}b_*$ *for some* $b_*>0$*, and* I_P *is the* P*-dimensional identity matrix. Furthermore,* $\mathbb{E}\left[\beta_i^4\right]\leq K\,P^{-2}$ *for some* $K>0$*.*

Assumptions 1(a), (b), and (d) are standard in the literature; see, for example, [36], [64], and [22]. Assumptions 1(a) and (b) rule out more general time-series dependence in the return innovations and features. While such dependence may be present in financial data, these assumptions simplify the analysis and allow us to focus on the effect of high dimensionality and noisy features. Assumption 1(b) still allows for dependence across the true features through the covariance matrix Ψ. Assumption 1(d) assumes that the predictive coefficients β have i.i.d. coordinates and a spherical covariance structure. This rules out systematic differences in predictability across individual features, but avoids imposing a particular structure on which features are more important and allows us to study the effect of increasing the number of features in a symmetric setting.

Assumption 1(c) imposes a simple additive noise structure and assumes that the noise is independent of the true features. This rules out more general forms of measurement error in which the noise may depend on the underlying feature values. However, the covariance matrix Φ still allows different features to have different levels of noise and permits the noise across features to be correlated. The specification therefore provides sufficient flexibility for studying the effect of feature noise while keeping the analysis tractable.

Since X_t and Y_t are independent, the covariance matrix of the observed features is

$$\mathbb{E}\left[\tilde{F}_t\tilde{F}_t^\intercal\right]=\Psi+\Phi,$$

while their cross-covariance with the true features is

$$\mathbb{E}\left[\tilde{F}_t F_t^\intercal\right]=\Psi.$$

For convenience, we denote the covariance matrix of the observed features

by

$$\tilde{\Psi} := \Psi + \Phi.$$

Thus, $\tilde{\Psi}$ incorporates both the variation in the true features and the additional variation induced by feature noise. The independence between X_t and Y_t is not essential to the main mechanism of the model, but is imposed to simplify the analysis and obtain clearer expressions.

We next impose an assumption on the asymptotic spectral behavior of the observed-feature covariance matrix $\tilde{\Psi}$.

Assumption 2. *Let $\lambda_k(\tilde{\Psi})$, for $k = 1, \ldots, P$, denote the eigenvalues of an arbitrary P-dimensional matrix $\tilde{\Psi}$. The spectral distribution*

$$\boldsymbol{F}^{\tilde{\Psi}}(x) = P^{-1} \sum_{k=1}^{P} \mathbf{1}_{\lambda_k(\tilde{\Psi}) \leq x}$$

of the eigenvalues of $\tilde{\Psi}$ converges to a nonrandom probability distribution $\tilde{H}$ supported on $[0, +\infty)$ as $P \to \infty$. Furthermore, $\tilde{\Psi}$ is uniformly bounded as $P \to \infty$, in the sense of the Spectral norm. We use

$$\tilde{\psi}_{*,k} = \lim_{P \to \infty} P^{-1} \operatorname{tr}\left(\tilde{\Psi}^k\right) k \geq 1$$

to denote the kth asymptotic moments of the eigenvalues of $\tilde{\Psi}$. Similarly, $\psi_{*,k}$ denotes asymptotic moments of the eigenvalues of Ψ.

With this assumption, one can estimate $\tilde{\Psi} = \Psi + \Phi$, which is essential in deriving the limiting behavior of the Sharpe ratio and R-squared of portfolios in high-dimensional settings where $P, T \to \infty$ and $P/T \to c > 0$.

Assumptions 1 and 2 lead us to the following lemma [47].

Lemma 1. *As the number of features becomes arbitrarily large, i.e., $P \to \infty$, we have*

$$\beta^\intercal A_P \beta - P^{-1} b_* \operatorname{tr}(A_P) \xrightarrow{\mathbb{P}} 0$$

for any bounded sequence of matrices A_P. In particular, $\beta^\intercal \Psi \beta \xrightarrow{\mathbb{P}} b_* \psi_{*,1}$.

In other words, Lemma 1 shows that the variance of the asset return converges to the product of the first moment of its covariance matrix scaled by

a constant.

3.2.2 Portfolio: timing strategy and performance evaluation

The investors use the noisy features $\tilde{F}_t$ in (3.5) to estimate the predictive coefficients β and to predict stock returns. Here, we study the Sharpe ratio of a timing investment strategy and the out-of-sample R-squared of predictions on the future returns of the asset, and the objective of the investor is to maximize the Sharpe ratio of their portfolio.

Recall that we model the predictive coefficients β as a random vector drawn from a distribution. In our analysis, we consider a single realization of β, so β is fixed throughout the period under study. This approach reflects the practical assumption that the predictive coefficients remain constant within the horizon of our analysis, see, e.g., [15], [57].

From this point onward, all expectations are conditional on the random variable β, unless explicitly stated otherwise. In particular, under Assumption 1, the future return R_{t+1} is independent of the current return R_t conditional on β.

3.2.2.1 Timing strategy

At time t, the investors' *timing strategy* uses predictive signals to choose the portfolio weight π_t assigned to the risky asset. Timing strategies involve adjusting the portfolio's exposure to the risky asset at each time point t to exploit predictable variations in returns, see, e.g., [55] and [39].

The portfolio return from time t to $t+1$ is given by

$$R^{\pi}_{t+1} = \pi_t \, R_{t+1} \,,$$

where π_t denotes the portfolio weight and recall that in (3.4) we work with excess returns.

The goal is to maximize the unconditional Sharpe ratio

$$\mathrm{SR} = \mathbb{E}[R^{\pi}_{t+1}] \Big/ \sqrt{\mathbb{E}[(R^{\pi}_{t+1})^2]} \,. \tag{3.6}$$

In (3.6), we use the second moment $\mathbb{E}[(R^{\pi}_{t+1})^2]$ instead of the variance $\mathrm{Var}[R^{\pi}_{t+1}]$. We make this choice following the arguments in KMZ that using the second moment of returns is consistent with the broader framework of mean-variance portfolio theory, where they argue that using the second moment does not affect the optimal portfolio weight π_t.

[34] show that under the true DGP in (3.4), the optimal weight π_t that maximizes the unconditional Sharpe ratio (3.6) is

$$\pi^{\mathbf{u}}_t = F_t^{\intercal}\,\beta/(1+(F_t^{\intercal}\,\beta)^2)\,.$$

With Assumptions 1 and 2, KMZ show that $(F_t^{\intercal}\,\beta)^2 \xrightarrow{\mathbb{P}} b_*\,\psi_{*,1}$ when the number of features $P \to \infty$, where b_* is the limiting Euclidean norm of the predictive coefficients β, and $\psi_{*,1}$ is the asymptotic first moment of the eigenvalues of Ψ, as defined in Assumption 1(d) and Assumption 2. Asymptotically, this convergence implies that the unconditional optimal weight $\pi^{\mathbf{u}}_t$ is the same as the conditional optimal weight

$$\pi^{\mathbf{c}}_t = F_t^{\intercal}\,\beta\,,$$

which maximizes the conditional Sharpe ratio

$$\mathbb{E}_t[R^{\pi}_{t+1}]\Big/\sqrt{\mathrm{Var}_t[(R^{\pi}_{t+1})^2]}\,,$$

where $\mathbb{E}_t$ and Var_t are the conditional expectation and variance of R^{π}_{t+1} with respect to the information set $\sigma(\beta, F_\tau, R_\tau, \tau = t, ..., t-T)$. Therefore, we focus on the conditional optimal weight $\pi^{\mathbf{c}}_t$.

In practice, an investor cannot compute the optimal weight $\pi^{\mathbf{c}}_t$ because she only observes the noisy model (3.5) instead of the true model (3.4), and hence she does not observe the true predictive coefficients β. Therefore, the investor uses

$$\hat{\pi}_t = \tilde{F}_t^{\intercal}\,\hat{\beta} \tag{3.7}$$

to estimate the conditional optimal portfolio weight, where $\hat{\beta}$ is adapted to the history $\sigma(\tilde{F}_\tau, R_{\tau+1}, \tau = t-1, ..., t-T)$. We refer to the portfolio with weight $\hat{\pi}_t$ as the "feasible" strategy.

3.2.2.2 Out-of-sample R-squared

The feasible strategy portfolio weight $\hat{\pi}_t = \tilde{F}_t^\intercal \hat{\beta}$ is an estimator of the expected return $\mathbb{E}[R_t] = F_t^\intercal \beta$ of the asset in the true model (3.4). Therefore, it is essential to study the predictive accuracy of this estimator on the future returns of the asset. Consider the mean squared error (MSE)

$$\text{MSE}(\hat{\beta}) = \mathbb{E}\left[\left(R_{t+1} - \tilde{F}_t^\intercal \hat{\beta}\right)^2 \mid \hat{\beta}\right] = \mathbb{E}\left[R_{t+1}^2\right] - 2\,\mathbb{E}\left[\hat{\pi}_t\, R_{t+1} \mid \hat{\beta}\right] + \mathbb{E}\left[\hat{\pi}_t^2 \mid \hat{\beta}\right], \tag{3.8}$$

where the expectation in (3.8) is also conditional on β; therefore, the first term in the decomposition in (3.8) is independent of $\hat{\beta}$ because R_{t+1} is independent of $\sigma(F_\tau, R_{\tau+1}, \tau = t-1, ..., t-T)$ given β, hence R_{t+1} is also independent of $\hat{\beta}$ given β. The remaining two terms on the right-hand side of (3.8) are the conditional expected return and the conditional squared value of the feasible timing strategy. Therefore, we refer to the last two terms in equation (3.8) as "Timing Expected Return" and "Timing Leverage", respectively. In particular, a timing strategy with larger expected return has lower MSE in predicting future returns, and a higher expected leverage corresponds to a higher MSE in predictions. Here, the term "Timing Leverage" is appropriate because it captures the effect of dynamically adjusting the position in an asset based on timing signals. By varying leverage, the strategy aims to exploit perceived short-term opportunities. The out-of-sample R-squared of the prediction in predictive coefficients is given by

$$\text{R}^2(\hat{\beta}) = 1 - \frac{\text{MSE}(\hat{\beta})}{\mathbb{E}\left[R_{t+1}^2\right]}. \tag{3.9}$$

3.3 Ridge regression

One input of the feasible timing strategy is the estimate of the predictive coefficients β. However, one cannot use the traditional least squares method to estimate β because it fails in high-dimensional settings, i.e., when $P/T \to c > 1$ as $P, T \to \infty$.

Instead, an investor can consider a Bayesian approach. Suppose her prior belief is that the distribution of the predictive coefficients β is $N\left(0, (z\,T)^{-1}\, I_P\right)$, where z is the ridge parameter, and the distribution of the error term ϵ_{t+1} in (3.4) is $N(0,1)$. Then, if she were to observe the true DGP in (3.4), the

posterior mean of her estimation of the predictive coefficients would be

$$\mathbb{E}[\beta(z)] = \hat{\beta}(z) = \left(z\, I_P + T^{-1} \sum_{\tau} F_\tau\, F_\tau^\intercal \right)^{-1} T^{-1} \sum_{\tau} F_\tau\, R_{\tau+1}\,, \qquad (3.10)$$

which coincides with the estimator of a ridge-regularized least square estimator, see [52].

However, the investor observes the noisy model in (3.5), so her estimate of the predictive coefficients is[1]

$$\hat{\beta}(z) = \left(z\, I_P + T^{-1} \sum_{\tau} \tilde{F}_\tau\, \tilde{F}_\tau^\intercal \right)^{-1} T^{-1} \sum_{\tau} \tilde{F}_\tau\, R_{\tau+1}\,. \qquad (3.11)$$

Alternatively, if one has other priors on the predictive coefficients β, one could derive other estimators for β. For example, if the prior is that β follows a Laplace distribution, one can use LASSO regression to estimate β. Previous studies suggest that assuming that the predictive coefficients β are sparse can result in unreliable estimations of the predictive coefficients, see, e.g., [28]. LASSO's tendency to force the value of coefficients to be zero can distort the estimation when the true model does not meet sparsity assumptions. Therefore, in this chapter, we use ridge regression instead of LASSO regression.

The ridge parameter $z_* = c/b_*$ maximizes the out-of-sample R-squared and the Sharpe ratio of the timing strategy $\hat{\pi}_t$ in (3.7) when the investor observes the true DGP in (3.4), see KMZ. From a Bayesian perspective, this optimal choice of ridge parameter is equivalent to taking the prior variance $(z\,T)^{-1}$ equal to the true variance b_*/P. However, the investors only see the noisy DGP in (3.5), so the optimal ridge parameter z_* takes a more complicated form.

[1] In [48], the authors state that our model is a special case of their Theorems 2 and 3. Their claim is incorrect because one cannot separate the noise term in the ridge regression. In particular, in our setting, the investor cannot observe $\Phi^{1/2} Y_t$ and therefore $\Phi^{1/2} Y_t$ must appear in equation (3.11).

3.4 Theoretical properties of ridge regression

In this section, we use the ridge estimator in (3.11) to study both the asymptotic behavior of the unconditional Sharpe ratio in (3.6) and the out-of-sample R-squared in (3.9) of the feasible timing strategy. Recall that the feasible timing strategy is a function of a noisy version of the DGP in (3.5).

3.4.1 Random matrix theory

To understand the limiting behavior of the ridge estimator in (3.11) as $P, T \to \infty$ and $P/T \to c > 1$, we study the limiting behavior of the sample covariance matrix $\hat{\tilde{\Psi}}_T = T^{-1} \sum_\tau \tilde{F}_\tau \tilde{F}_\tau^\intercal$. Unfortunately, as with most random covariance matrices, it is very difficult to characterize the limiting distribution of the matrix. Luckily, we need to consider only the limiting distribution of the eigenvalues of $\hat{\tilde{\Psi}}_T$ to derive all necessary results in this section.

For any symmetric matrix $\tilde{\Psi}$, a useful property is that

$$\begin{aligned} P^{-1} \operatorname{tr}\left((\tilde{\Psi} - u\, I)^{-1}\right) &= P^{-1} \sum_{i=1}^{P} \left(\lambda_i(\tilde{\Psi}) - u\right)^{-1} \\ &= \int_{-\infty}^{\infty} \frac{1}{x-u}\, d\boldsymbol{F}^{\tilde{\Psi}}(x)\,, \quad u < 0\,, \end{aligned} \tag{3.12}$$

where $\lambda_i(\tilde{\Psi})$ are the eigenvalues of $\tilde{\Psi}$, and $\operatorname{tr}(\,\cdot\,)$ is the trace operator. The right-hand side of (3.12) is the Stieltjes transform of the distribution function $\boldsymbol{F}^{\tilde{\Psi}}$ of the eigenvalues of $\tilde{\Psi}$. By Assumption 2, this distribution converges to a non-random distribution $\tilde{H}$ as $P \to \infty$. Therefore, we have

$$m_{\tilde{\Psi}}(u) = \int_{-\infty}^{\infty} \frac{1}{x-u}\, d\tilde{H}(x) = \lim_{P\to\infty} P^{-1} \operatorname{tr}\left((\tilde{\Psi} - u\, I_P)^{-1}\right)\,.$$

The function $m_{\tilde{\Psi}}(u)$ is the limiting Stieltjes transform of the eigenvalue distribution of $\tilde{\Psi}$. In practice, the matrix $\tilde{\Psi}$ is not observed, so one uses the sample covariance matrix $\hat{\tilde{\Psi}}_T$ to compute the empirical representation of $m_{\tilde{\Psi}}(u)$, which is given by

$$m(u; c) = \lim_{P,T\to\infty, P/T\to c} P^{-1} \operatorname{tr}\left(\left(\hat{\tilde{\Psi}}_T - u\, I_P\right)^{-1}\right)\,.$$

The following lemma from [5] and [60] shows the relationship between $m(u;c)$ and $m_{\tilde{\Psi}}(u)$.

Lemma 2. *Let Assumptions 1 and 2 hold, and let $c > 0$ and $u < 0$. Then, as $P, T \to \infty$ and $P/T \to c$, the distribution of the eigenvalues of $\hat{\tilde{\Psi}}_T$ converges to a distribution whose Stieltjes transform $m(u;c)$ is the unique positive solution to the equation*

$$m(u;c) = \frac{1}{1 - c - c\,u\,m(u;c)}\, m_{\tilde{\Psi}}\left(\frac{u}{1 - c - c\,u\,m(u;c)}\right). \tag{3.13}$$

Lemma 2 states that the empirical Stieltjes transform $m(u;c)$ derived from the sample covariance matrix is a non-random function which depends on the true Stieltjes transform distribution $m_{\tilde{\Psi}}(u)$. With this convergence result, one can characterize the trace of the matrix $T^{-1}\left(\hat{\tilde{\Psi}}_T + z\,I_P\right)^{-1}\tilde{\Psi}$.

Proposition 1. *Let Assumptions 1 and 2 hold. Then, as $P, T \to \infty, P/T \to c$, for any $z > 0$, we have*

$$T^{-1}\,\mathrm{tr}\left(\left(z\,I_P + \hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\right) \xrightarrow{\mathbb{P}} \xi(z;c)\,, \tag{3.14}$$

where

$$\xi(z;c) = \frac{1 - z\,m(-z;c)}{c^{-1} - 1 + z\,m(-z;c)}\,. \tag{3.15}$$

Proposition 1 shows that the dependence of the trace of $T^{-1}\left(\hat{\tilde{\Psi}}_T + z\,I_P\right)^{-1}\tilde{\Psi}$ on the Stieltjes transform $m(-z;c)$ is non-linear. If the true DGP in (3.4) is observed and $\xi(z;c)$ is known, one can derive asymptotic results for the R-squared and Sharpe ratio of the timing strategy, see, for example, KMZ.

However, in our setting, where one observes a noisy version of the model in (3.5), the expected return $\mathbb{E}\left[\hat{\pi}_t\,R_{t+1} \mid \hat{\beta}\right]$ and the expected leverage term $\mathbb{E}\left[\hat{\pi}_t^2 \mid \hat{\beta}\right]$ of the timing strategy are different from those in KMZ. In particular, $\mathbb{E}\left[\hat{\pi}_t^2 \mid \hat{\beta}\right]$ is equal to $\hat{\beta}^\intercal\,\tilde{\Psi}\,\hat{\beta}$ instead of $\hat{\beta}^\intercal\,\Psi\,\hat{\beta}$ as in KMZ, which makes the limit in (3.14) insufficient for our results. Below, we introduce two limits

that capture the interaction between the covariance structure of the observed features and their covariance with the true features.

Assumption 3. *The following limits exist and are finite:*

$$\xi_{\Psi}(z;c) = \lim_{P,T\to\infty,P/T\to c} T^{-1}\ \mathrm{tr}\left(\mathbb{E}\left[\left(z\,I_P + \hat{\tilde{\Psi}}_T\right)^{-1}\Psi\,\Psi^{\intercal}\right]\right) \geq 0\,, \tag{3.16}$$

$$\eta_{\Psi}(z;c) = \lim_{P,T\to\infty,P/T\to c} T^{-1}\ \mathrm{tr}\left(\mathbb{E}\left[\left(z\,I_P + \hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\left(z\,I_P + \hat{\tilde{\Psi}}_T\right)^{-1}\Psi\,\Psi^{\intercal}\right]\right) \geq 0\,. \tag{3.17}$$

The quantities $\xi_{\Psi}(z;c)$ and $\eta_{\Psi}(z;c)$ are analogous to the auxiliary limits introduced by KMZ (i.e, equation 17 in KMZ) in their analysis. Unlike $\xi(z;c)$, which depends only on the covariance structure of the observed features, these quantities additionally capture the relationship between the observed and true features through $\Psi = \mathbb{E}[\widetilde{F}_t F_t^{\intercal}]$. In general, explicit expressions for these limits are difficult to obtain because they depend on the joint covariance structure represented by Ψ and $\widetilde{\Psi}$. However, recent analysis provides conditions under which limits of this type exist for fairly general covariance structures, see, e.g, Theorem 6 and Lemma 2 in [48]. Assumption 3 is therefore mainly a technical regularity condition rather than a strong restriction on the model.

3.4.2 Asymptotic results

With the notations and assumptions above, we prove the following limiting results for the out-of-sample R-squared and Sharpe ratio of the timing strategy.

Theorem 1. *Let Assumptions 1, 2, and 3 hold. Then, as $P, T \to \infty, P/T \to c$, the expected return of the timing strategy is given by*

$$\mathcal{E}(z;c) = \lim_{P,T\to\infty,P/T\to c} \mathbb{E}\left[\hat{\pi}_t(z)\,R_{t+1} \mid \hat{\beta}\right] = \frac{b_*\,c^{-1}\,\xi_{\Psi}(z;c)}{1+\xi(z;c)}\,, \tag{3.18}$$

and the expected leverage of the timing strategy converges to

$$\begin{aligned}\mathcal{L}(z;c) &= \lim_{P,T\to\infty,P/T\to c} \mathbb{E}\left[\hat{\pi}_t(z)^2 \mid \hat{\beta}\right] \\ &= (1+b_*\,\psi_{*,1})\,(\xi(z;c)+z\,\partial\xi(z;c)/\partial z)+\frac{b_*\,c^{-1}\,\eta_\Psi(z;c)}{(1+\xi(z;c))^2} \\ &\quad -\frac{2\,b_*\,c^{-1}\;(\xi(z;c)+z\,\partial\xi(z;c)/\partial z)\;\xi_\Psi(z;c)}{1+\xi(z;c)}\,, \end{aligned} \tag{3.19}$$

and the out-of-sample R-squared of the timing strategy converges to

$$R^2(z;c)=\frac{2\,\mathcal{E}(z;c)-\mathcal{L}(z;c)}{1+b_*\,\psi_{*,1}}\,. \tag{3.20}$$

The conditional second moment of the feasible timing strategy return converges to

$$\mathcal{V}(z;c)=\lim_{P,T\to\infty,P/T\to c}\mathbb{E}\left[(\hat{\pi}_t(z)\,R_{t+1})^2\mid\hat{\beta}\right]=2\,\mathcal{E}(z;c)^2+(1+b_*\,\psi_{*,1})\,\mathcal{L}(z;c)\,,$$

and the Sharpe ratio of the feasible timing strategy is

$$SR(z;c)=\frac{\mathcal{E}(z;c)}{\sqrt{2\,\mathcal{E}(z;c)^2+(1+b_*\,\psi_{*,1})\;\mathcal{L}(z;c)}}\,. \tag{3.21}$$

In all results, convergence is in probability.

The theorem above shows the convergence and explicit forms of the limiting behaviors of the timing strategy mean return, timing strategy expected leverage, and the out-of-sample R-squared and Sharpe ratio of the timing strategy. These results are not easy to interpret. In particular, (3.20) and (3.21) involve complex functions, so it is challenging to interpret the impact of parameters (e.g., the ridge parameter z) on the Sharpe ratio and R-squared of the timing strategy.

Below we discuss a special case when $\Psi = I_P$ and $\Phi = a\,I_P$, where $a \geq 0$ represents the noise level in the features. In this case, one can obtain explicit expressions for $\xi_\Psi(z;c)$ and $\eta_\Psi(z;c)$.

3.4.3 Independence among features

Consider $\Psi = I_P$ and $\Phi = a\, I_P$, that is, the true features are independent from each other, and the order of magnitude of the measurement error for each feature is the same and the noise is independent from each other. This special case is of empirical significance. In practice, quantitative portfolio managers often ensure that any new feature they add to their existing pool of features is independent from any of the previously functioning ones [35]. Here, the investor invests in infrastructures to reduce the noise level a, and conditional on the noise level a the investor chooses the ridge parameter z to maximize the Sharpe ratio of the timing strategy. In this case, the limiting Stieltjes transform of $\tilde{\Psi}$ is $m_{\tilde{\Psi}}(u;a) = \frac{1}{1+a-u}$ and the limiting Stieltjes transform of $\hat{\tilde{\Psi}}_T$ is

$$m(u;c;a) = \frac{(1+a)\,(1-c) - u - \sqrt{(u-(1+c)\,(1+a))^2 - 4\,c\,(1+a)^2}}{2\,(1+a)\,c\,u}\,,$$

see [53]. Additionally, the limit of the trace of $T^{-1}\left(\hat{\tilde{\Psi}}_T + z\, I_P\right)^{-1}\tilde{\Psi}$ as the sample size T becomes arbitrarily large, is given by

$$\xi(z;c;a) = \frac{1 - z\, m(-z;c;a)}{c^{-1} - 1 + z\, m(-z;c;a)} = \lim_{P,T\to\infty, P/T\to c} T^{-1}\ \mathrm{tr}\left((z\, I_P + \hat{\tilde{\Psi}}_T)^{-1}(1+a)\right)\,.$$

Similarly, one can write ξ_Ψ and η_Ψ as

$$\xi_\Psi(z;c;a) = \lim_{P,T\to\infty, P/T\to c} T^{-1}\,\mathrm{tr}\left(\mathbb{E}\left[\left(z\, I_P + \hat{\tilde{\Psi}}_T\right)^{-1}\right]\right) = \xi(z;c;a)/(1+a)\,,$$

$$\begin{aligned}\eta_\Psi(z;c;a) &= \lim_{P,T\to\infty, P/T\to c} T^{-1}\ \mathrm{tr}\left(\mathbb{E}\left[\left(z\, I_P + \hat{\tilde{\Psi}}_T\right)^{-1}\left(z\, I_P + \hat{\tilde{\Psi}}_T\right)^{-1}(1+a)\right]\right)\\ &= -\frac{\partial \xi(z;\, c;\, a)}{\partial z}\,.\end{aligned}$$

Substituting the expressions into (3.18) and (3.19), we have the following corollary.

Corollary 1. *Let the assumptions in Theorem 1 hold, and assume that* $\Psi = I_P$ *and* $\Phi = a\, I_P$*. Then, the timing strategy's expected return is*

$$\mathcal{E}(z;c;a) = \frac{b_*\, c^{-1}\, \xi(z;c;a)}{(1+\xi(z;c;a))\,(1+a)}\,,$$

and its expected leverage is

$$\mathcal{L}(z;c;a) = (1+b_*)\ (\xi(z;\ c;\ a) + z\,\partial\xi(z;\ c;\ a)/\partial z) - \frac{b_*\,c^{-1}\,\partial\xi(z;\ c;\ a)/\partial z}{(1+\xi(z;\ c;\ a))^2} - \frac{2b_*\,c^{-1}\ (\xi(z;\ c;\ a) + z\,\partial\xi(z;\ c;\ a)/\partial z)\ \xi(z;\ c;\ a)}{(1+\xi(z;\ c;\ a))\,(1+a)}\,.$$

Next, with

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

the R-squared and Sharpe ratio of the timing strategy are

$$R^2(z;c;a) = \frac{2\,\mathcal{E}(z;c;a) - \mathcal{L}(z;c;a)}{1+b_*} \tag{3.22}$$

and

$$SR(z;c;a) = \frac{\mathcal{E}(z;c;a)}{\sqrt{2\,\mathcal{E}(z;c;a)^2 + (1+b_*)\ \mathcal{L}(z;c;a)}}\,. \tag{3.23}$$

Now, with (3.22) and (3.23), we derive further results regarding the optimal choice for the ridge parameter z and the properties of the R-squared and Sharpe ratio as a function of the noise level a.

Proposition 2. *Let the assumptions in Corollary 1 hold. Then the following statements hold:*

(i) The expected return $\mathcal{E}(z;c;a)$ of the timing strategy is greater than zero for any ridge parameter $z \geq 0$, noise level $a \geq 0$, and $c \geq 0$. Furthermore, $\mathcal{E}(z;c;a)$ is monotonically decreasing and convex in the noise level a.

(ii) The out-of-sample R-squared $R^2(z;c;a)$ and Sharpe ratio $SR(z;c;a)$ of the timing strategy are both monotonically increasing in z for $z < z_* = c\,(1+a)\,(1+a+a\,b_*)/b_*$ and monotonically decreasing in z for $z > z_*$. Furthermore, with the optimal choice z_*, both $R^2(z_*;c;a)$ and $SR(z_*;c;a)$ are monotonically decreasing and convex in the noise level a.

(iii) In the ridgeless limit $z \to 0$, the expected return of the timing strategy is

$$\mathcal{E}(0;c;a) = \begin{cases} \frac{b_*}{(1+a)\,c} & \text{if} \quad c > 1\,, \\ \frac{b_*}{(1+a)} & \text{if} \quad c < 1\,, \end{cases}$$

and the expected leverage of the timing strategy is

$$\mathcal{L}(0;c;a) = \begin{cases} \frac{-b_*+(1+a)\,(1+b_*)\,c}{(1+a)\,(c-1)\,c} & \textit{if} \quad c>1\,, \\ -\frac{b_*+(1+a+(a-1)\,b_*)\,c}{(1+a)\,(c-1)} & \textit{if} \quad c<1\,. \end{cases}$$

In particular, when $a = 0$, the optimal ridge parameter z_* is c/b_*, in which case our results are identical to those in Proposition 3 of KMZ.

3.4.4 A note on the convexity of performance measures

Proposition 2 shows that when an investor observes a noisy version of the features and uses ridge regression to estimate the predictive coefficients, one can obtain positive expected returns and positive Sharpe ratio. In particular, the Sharpe ratio of the timing strategy increases monotonically as the ridge parameter is adjusted from zero to z_* and decreases as the value of the ridge parameter increases beyond z_*, where z_* is the optimal level of shrinkage.

When investors know the optimal value of the ridge parameter z_*, they can achieve the maximum Sharpe ratio of the timing strategy. Recall that the noise level a is determined by the quality of infrastructure, which is determined by the efforts of the investor into hiring staff, acquiring information, and building better models. Therefore, better infrastructure is associated with lower levels of a and less sophisticated infrastructure corresponds to higher levels of a.

For the optimal level of shrinkage z_*, Proposition 2 states that the expected return, out-of-sample R-squared, and Sharpe ratio of the timing strategy are all monotonically decreasing and convex in the noise level a. Monotonicity in a shows that the performance of every investor's portfolio increases as the investor invests more in infrastructure. However, the investor must consider if the increase in portfolio performance justifies the additional infrastructure costs.

The convexity result in Proposition 2 addresses this core research question. When the infrastructure provided by the investor is good (i.e., level of noise a is low), she benefits more from a marginal improvement in her infrastructure; on the other hand, if the infrastructure is poor (i.e., level of noise a is high) the benefits from a marginal improvement in the infrastructure are small.

The third result in Proposition 2 discusses the ridgeless limit of the mean return and the expected leverage of the timing strategy. When the dimension of the true DGP is low (i.e., $c < 1$), the mean return of the timing strategy is constant in c and monotonically decreases as the value of the noise level a increases. In high-dimensional settings (i.e., $c > 1$), we apply no shrinkage to the estimated coefficients in the ridgeless case. In the absence of shrinkage, the expected return of the timing strategy decreases monotonically in c, so the return of the timing strategy decreases as the dimension of the true DGP increases. A lower expected return does not necessarily correspond to a lower Sharpe ratio in the ridgeless setting. In the high-dimensional ridgeless setting, the Sharpe ratio first increases and then decreases as the complexity c of the true model increases.

3.5 Empirical analysis

In this section, we investigate our theoretical results via empirical analysis.

3.5.1 Data

We use the same dataset and follow the same data preprocessing approach as that in KMZ. Our forecast target is the monthly excess return of the CRSP value-weighted index. Our predictors are 15 variables from [63], available monthly from 1926 to 2023.

Returns are adjusted based on their rolling 12-month standard deviation. Like KMZ, we calculate the standard deviation from the uncentered second moment. We use an expanding window over the historical data of the predictors. To ensure stable predictor standardization, we use at least 36 months of data. Therefore, our analysis employs data from 1930 onwards.

3.5.2 Empirical design of noise incorporation

Empirically, to analyze the noisy DGP (3.5) one requires two steps. First, we examine how empirical results change from low-dimensional cases where $P \ll T$ to high-dimensional cases where $P \gg T$. Second, in our empirical analysis, as in KMZ, we use Random Fourier Features (RFF) ([58], [59]) to add noise to the features. Let G_t represent our 15×1 vector of predictors. The RFF method transforms G_t into a new set of features $F_{i,t}$ for $i = 1, \ldots, P/2$

as follows:

$$F_{i,t} = [\sin(\gamma\,\omega_i^\intercal\,G_t)\,,\,\cos(\gamma\,\omega_i^\intercal\,G_t)]^\intercal\,, \quad \omega_i \sim \text{ i.i.d. } N(0, I_{15})\,.$$

In this transformation, ω_i is a 15×1 vector of random weights drawn independently from a multivariate standard normal distribution, for each i the transformation creates a random linear combination of the original predictors G_t. The scaling parameter γ controls the frequency of the sine and cosine functions, so it affects the smoothness of the resulting features. The index i ranges from 1 to $P/2$, the total number of features equals P because each i yields two features (the sine and cosine components).

This transformation maps the original predictors G_t into a higher-dimensional, nonlinear feature space. The benefit of RFF is that, for a given input data set G_t, one can vary the value of P to generate arbitrarily small and large sets of features to explore low-dimensional and high-dimensional settings in our analysis.

To add noise to the features, we use two approaches. The first approach adds noise directly to the observed predictor, G_t. In this case, the noisy observation is

$$\tilde{G}_t = G_t + a\,Y_t, \quad Y_t \sim N(0, I_{15})\,, \tag{3.24}$$

so the corresponding noisy features are

$$\tilde{F}_{i,t} = \left[\sin\left(\gamma\,\omega_i^\intercal\,\tilde{G}_t\right),\,\cos\left(\gamma\,\omega_i^\intercal\,\tilde{G}_t\right)\right]^\intercal, \quad \omega_i \sim \text{ i.i.d. } N(0, I_{15})\,. \tag{3.25}$$

Here, when $a > 0$, the noise arises from limited access to high-quality data by investors. [2]

The second approach adds noise after generating the features. Here, we first use the noiseless G_t to create the RFFs, $F_{i,t}$. Then, we add noise to the

[2]For instance, in intra-day trading, more sophisticated investors may invest significantly into acquiring expensive, clean tick-level data, whereas less sophisticated ones only have access to either less granular snapshot data or tick-level data that are less accurate and less processed.

generated features, so

$$\tilde{F}_{i,t} = [\sin(\gamma\,\omega_i^\intercal\,G_t) + a\,\mu_{i,t},\ \cos(\gamma\,\omega_i^\intercal\,G_t) + a\,\eta_{i,t}]^\intercal\,,$$
$$(\omega_i^\intercal, \mu_{i,t}, \eta_{i,t})^\intercal \sim \text{ i.i.d. } N(0, I_{17})\,. \tag{3.26}$$

As above, a represents the noise level. This approach models a scenario where noisy observations of the feature stem from the skills and infrastructure of the investors. Investors with better technical skills (e.g., machine learning expertise and financial knowledge) and superior infrastructure (e.g., greater computing resources and efficient coding practices) experience less noise in their features, even when using the same data.[3]

We use one-year rolling training windows ($T = 12$ months) and a large number of RFFs, with P as high as 4000. We mention two advantages from using short training periods, e.g., $T = 12$ months. One, the complexity of the model is high with fewer RFFs, thus reducing computational effort, and two, it demonstrates that the advantages of model complexity can be obtained with small sample sizes of observations.

We generate a series of out-of-sample forecasts and trading strategies for different levels of model complexity, where the number of RFFs is $P \in \{2, 100, 500, 1000, 2000, 4000\}$. We also explore a wide range of ridge shrinkage parameters, with $\log_{10} z$ ranging from -3 to 5, placing greater emphasis on larger values. Our theoretical findings suggest that the optimal level of shrinkage should be high in settings where the value of $c = P/T$ is large. For the noise level, we vary the value of a from 0 to 0.5 when adding noise directly to the data and from 0 to $\sqrt{5}$ when adding noise to the generated features. We repeat the experiment 100 times and average the results across all repetitions. The steps for each iteration of the experiment are:

(i) **Noise generation**: In the first approach, use (3.24) to add noise directly to the observation G_t and generate $\tilde{G}_t$, then use (3.25) to create 4000 RFFs. In the second approach, use (3.26) to generate the 4000 RFFs where the noise is added after one generates the features. The bandwidth parameter is $\gamma = 2$, as in KMZ.

(ii) **Model setup**: In both approaches, we use a selected number of features, with $P \in \{2, 100, 500, 1000, 2000, 4000\}$, and a ridge shrinkage param-

[3]Our approaches to introducing noise align with the approach discussed by [51], where they suggest that the noise may already exist in the data, or it may be added by the user.

eter $\log_{10} z \in \{-3, \ldots, 5\}$. The predictors $\tilde{F}_t$ for the regression in (3.5) are constructed by selecting P features from the set of RFFs generated in step (i).

(iii) **Out-of-sample predictions**: We use the estimations from step (ii) to perform recursive out-of-sample predictions and to implement a market timing strategy. For each $t \in \{12\,\text{months}, \ldots, 1127\,\text{months}\}$, we estimate (3.5) with the training data

$$\left\{R_\tau, \tilde{F}_{\tau-1}, \tau = t, \cdots, t-T+1\right\}.$$

We compute the out-of-sample return forecast $\hat{\beta}^\intercal \tilde{F}_t$, and the timing strategy return $\hat{\beta}^\intercal \tilde{F}_t R_{t+1}$ from the estimated coefficients. Prior to estimation, we standardize both the training and out-of-sample RFFs by their volatility.

(iv) **Performance evaluation**: Finally, we calculate the out-of-sample R^2, along with the average return, volatility, and Sharpe ratio of the timing strategy from the sequence of out-of-sample predictions and strategy returns generated in step (iii).

3.5.3 Out-of-sample performance

First, we analyze the performance of the approach where noise is added directly to the data. We investigate how noise levels impact two key metrics: out-of-sample R-squared and Sharpe ratio. To do this, we first examine the results under fixed ridge shrinkage values. Specifically, we look at shrinkage levels of $\log_{10} z = 0$, $\log_{10} z = 3$, and $\log_{10} z = 4$, see Figures 3.1, 3.2, and 3.3.

The average return and Sharpe ratio of the strategy exhibit nearly convex trends in Figure 3.1, 3.2, and 3.3, which is consistent with our theoretical predictions. The R-squared metric does not follow the same convex pattern. With small shrinkage ($\log_{10} z = 0$), the R-squared values remain relatively flat as the variance of noise increases, implying that the model's capacity to capture the true underlying signal is compromised when regularization is insufficient. However, as shrinkage increases to $\log_{10} z = 3$ and $\log_{10} z = 4$, more convex patterns emerge in the R-squared curves. This indicates that regularization helps larger models perform better by improving generalization when dealing with noise.

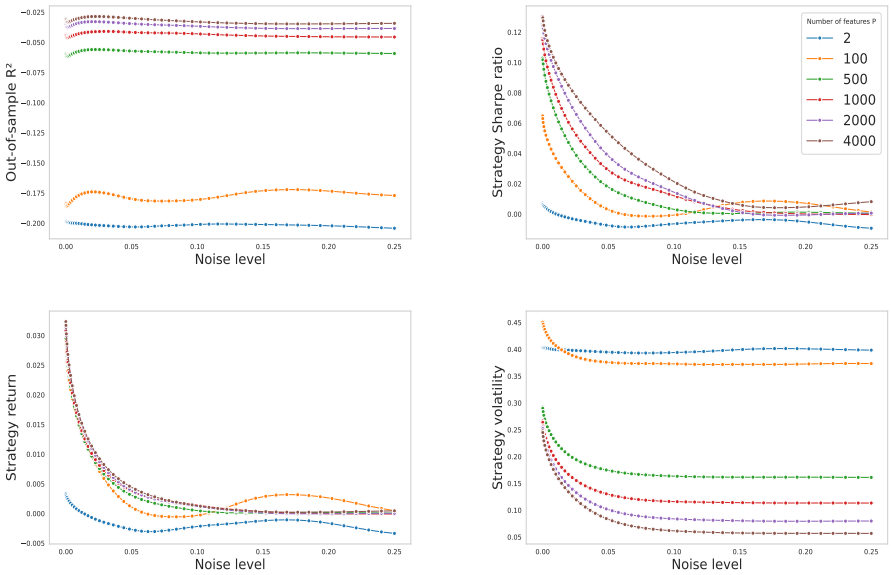


Figure 3.1: Out-of-sample market timing performance for $\log_{10} z = 0$. Training window is $T = 12$ months and RFF count P ranges from 2 to 4,000 with $\gamma = 2$, $\log_{10} z = 0$, and noise variance ranging from 0 to 0.25.

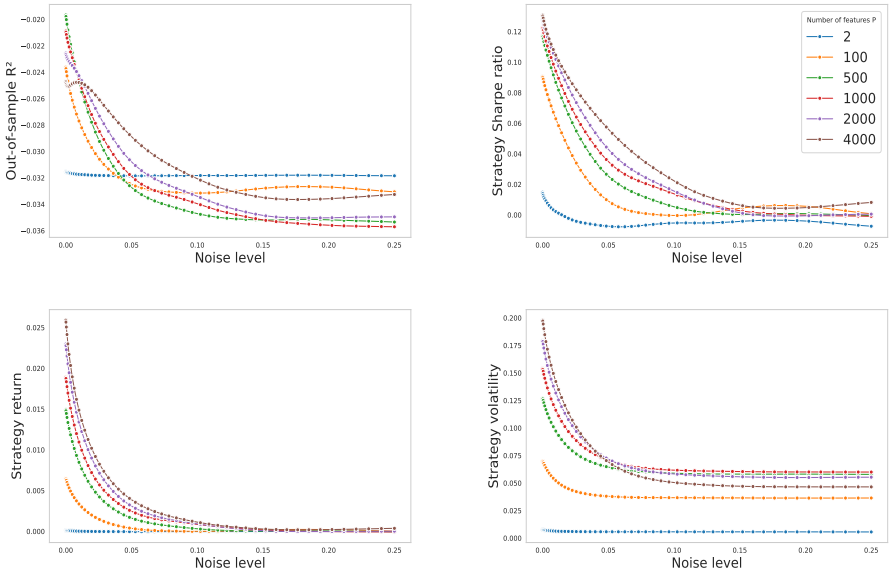


Figure 3.2: Out-of-sample market timing performance for $\log_{10} z = 3$.

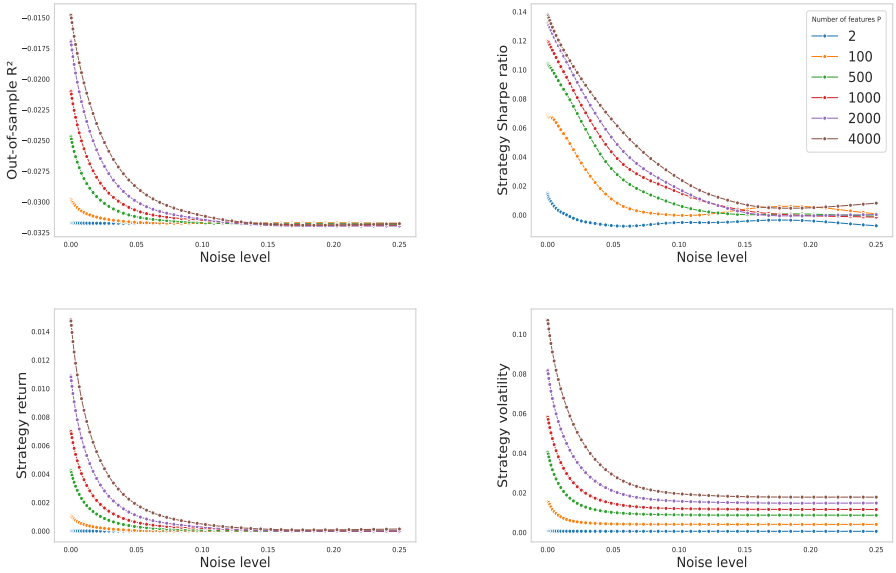


Figure 3.3: Out-of-sample market timing performance for $\log_{10} z = 4$.

To further understand the role of shrinkage, we plot the maximum R-squared and Sharpe ratio values for various shrinkage levels. In particular, for each value of P, we cross-validate over the shrinkage level $\log_{10} z$ and take the maximum average value of R-squared and Sharpe ratio. Figure 3.4 shows the optimal performance achievable with appropriate regularization. We observe a more convex behavior for various values of P. Compared to Figure 3.1, 3.2, and 3.3, for all P, the maximum R-squared and maximum Sharpe-ratio are monotonically decreasing and convex in the noise level.

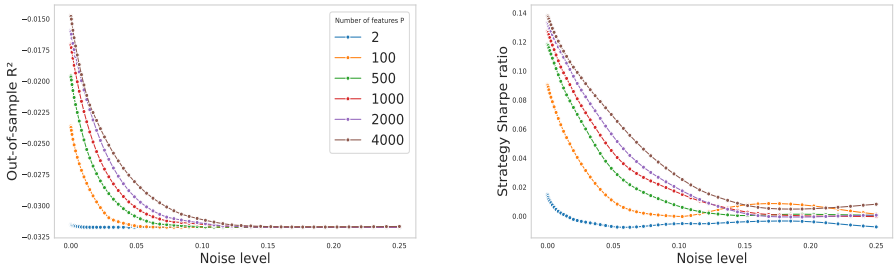


Figure 3.4: Best out-of-sample market timing performance for adding noise to the data.

Next, we focus on the approach where noise is added into the generated features. When we keep the value of the ridge shrinkage constant, the results

closely resemble those from the first approach where noise is added directly to the data. Therefore, to streamline the analysis, we present the results for the best average out-of-sample performance for various shrinkage levels.

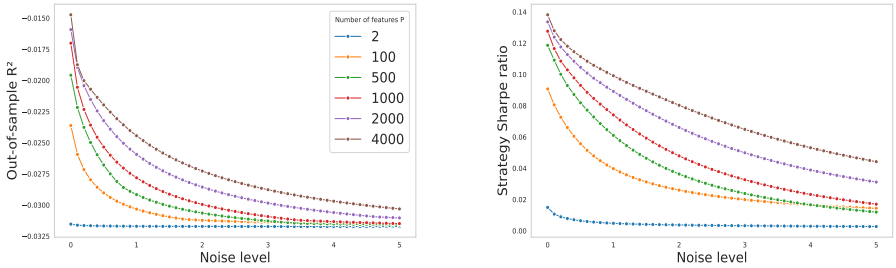


Figure 3.5: Best out-of-sample market timing performance for adding noise to the feature.

Figure 3.5 illustrates the best average out-of-sample performance when noise is directly added to the features, showing clearer convex trends in the performance metrics compared to the first approach where noise is added to the data. One possible explanation for this improvement is that, in the second approach, the simulated model is more closely aligned with the theoretical structure described in (3.5). By adding noise directly to the "true" features, we simulate a more realistic scenario where noise affects the underlying signal processing rather than the raw data itself.

Our empirical analysis provides strong support for the theoretical predictions. By varying noise levels and model complexity, we observe a clear convex relationship between performance metrics, such as Sharpe ratio and R-squared, and the level of noise, as predicted by our theory. Portfolios constructed with higher noise consistently underperform those based on sharper features, and this effect is observed both when noise is added at the data level or at the feature level. The adverse impact of noise is more noticeable in simpler models, but performance improves when noise is reduced, particularly with appropriate regularization.

3.6 Conclusion

In this chapter, we study a noisy version of the virtue of complexity. Unlike KMZ, where investors observe the true feature matrix, we assume that

investors observe noisy features. This setting captures the practical role of infrastructure, data quality, computational resources, and research skill in getting the features used for portfolio construction.

We show that, under appropriate shrinkage, the out-of-sample R-squared and Sharpe ratio of the feasible timing strategy are decreasing and convex in the noise level. This result implies that reducing noise improves portfolio performance, but the marginal benefit is larger when the investor already has relatively high-quality infrastructure. Our empirical analysis supports this prediction.

The analysis in this chapter provides a baseline setting in which investors observe the full set of predictive features, but these features are measured with noise. In the next chapter, we relax this full-observation assumption and consider a partial-observation setting in which investors observe only a subset of the true features, which may themselves be noisy. This extension allows us to study the trade-off that arises as model complexity increases: observing more features provides additional predictive information, but may also introduce more measurement error. We show that, in this setting, the monotonicity result in KMZ may break down, leading to a *limited virtue of complexity*.

4 | The Limited Virtue of Complexity in a Noisy World

4.1 Introduction

Advances in large-scale data acquisition, algorithmic trading, and machine learning techniques are reshaping modern financial markets. Today's asset pricing and portfolio allocation problems involve an ever-growing array of potential features, ranging from traditional macroeconomic variables to text-based sentiment indicators, high-dimensional firm-level characteristics, and alternative datasets. However, should investors include as many feature-based predictors as possible in their return prediction and investment models? Or is there a threshold beyond which adding more predictive features adversely affects model performance?

The traditional econometric view, rooted in Occam's razor and articulated by Box's dictum that "all models are wrong, but some are useful," advocates for parsimonious models that prioritize interpretability and guard against overfitting [9]. Yet, recent advances in statistical learning theory challenge this view. The double descent phenomenon shows that model performance may improve even in highly overparameterized models, provided the models are appropriately regularized [7, 6, 36, 46]. These developments give rise to two seemingly opposing views, one of which favors simplicity, while the other embraces complexity.

This chapter reconciles these two schools of thought. We develop a theoretical framework in which an investor predicts excess returns with a large number of features that are observed with noise. Noise arises from issues such as inconsistent data collection, processing errors, or insufficient computing power, all of which are common challenges in real-world settings. Our theoretical framework proves a limited virtue of complexity in financial forecasting where the performance of portfolios depends on the noise level in features, and we show that more complex models do not necessarily lead to better performance when features are not perfectly observed.

This chapter builds directly on the noisy-feature framework developed in Chapter 3. There, we assume that the investor observes the full set of predictive features with measurement error. Here, we relax this full-observation

assumption and consider a partial-observation setting in which the investor observes only a subset of the true features, which may themselves be noisy. This extension allows us to study the trade-off between the additional predictive information gained from observing more features and the measurement error associated with increasing model complexity.

We provide three examples to motivate our framework. The first example replicates the results of [47], henceforth KMZ, and the other two examples focus on the relationship between model complexity and predictive performance in practical settings when features may be observed with noise.

First, to illustrate how a model performs in an ideal complex environment, we start with an example following the same approach as that in KMZ.[1] Specifically, we use market data to simulate features and assume that the investor perfectly observes as many of these features as desired and uses them in a ridge regression to predict stock returns. In this ideal setup, Figure 4.1 shows that, consistent with the findings of KMZ, increasing model complexity, i.e., including more features, improves the out-of-sample R-squared of the predictions.[2]

However, investors do not have off-the-shelf access to all (often latent) market features, nor are there guarantees that relevant features can be perfectly observed or that features can be computed or learned without noise. In practice, numerous studies suggest that neural networks are well suited for learning complex, data-driven patterns that map real-world inputs to stock returns [see, e.g., 31, 29, 44, 20].[3] Hence, instead of simulating market features, one can train a neural network on market data to predict returns, which is a more realistic approach to learn and exploit the diverse and interdependent features in financial markets.[4]

Therefore, our second example proceeds as in [31], where we use market data from NYSE, AMEX, and Nasdaq between 1990 and 2024 to fit a neural

[1] For training details, see Appendix B.2.

[2] [56] argues that the timing strategy based on simulated features in KMZ mainly reflects a volatility-timed momentum strategy rather than genuine return predictability. [18] show that the empirical "virtue of complexity" documented in KMZ can arise mechanically from the structure of random Fourier features, rather than from learning. Here, our analysis focuses on the trade-off between model complexity and noise in features.

[3] Additional examples include [38], [61], [32], [33], [50], [3], [11], [4], [16], [26], and [46].

[4] In our empirical analysis in Section 4.2, we take a different route: instead of simulating features, we denoise observed features with an autoencoder before forecasting.

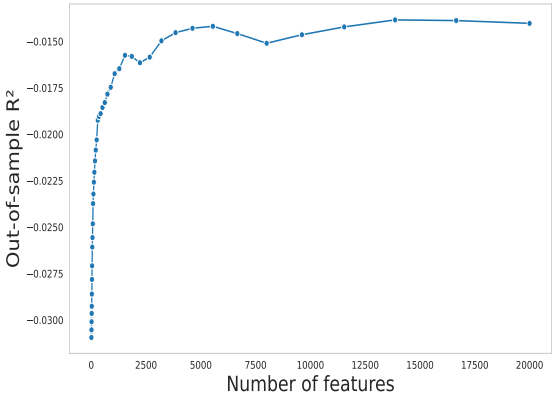


Figure 4.1: Prediction of stock returns using simulated features that are observed without noise. The x-axis shows the number of features, the y-axis is the out-of-sample R-squared. The results are averaged over 50 repetitions.

network to predict monthly stock returns.[5] Our findings show that the out-of-sample performance of the model does not monotonically increase as the model becomes more complex, see Figure 4.2. Indeed, as model complexity increases, the predictive performance of the model first increases when complexity is low and then decreases when complexity is high. These observations are not in line with the results of KMZ in Figure 4.1,[6] where out-of-sample model performance should increase as the model becomes more complex.[7]

When learning features, practical constraints can explain why the performance of a model does not improve when the complexity of the model increases. For example, due to limited resources such as the amount and quality of data, limited access to computational power, and the skill of investment research staff (collectively referred to as "infrastructure"), increasing the complexity of the feature space often also increases the average measurement error in each feature. In particular, in many models, when additional parameters or layers are introduced, the entire feature structure must be re-estimated, which inadvertently magnifies errors in finite-sample data. In fact, [37] and

[5]For training details, see Appendix B.2.

[6]In much of the literature predicting stock returns with neural networks, final performance metrics are commonly averaged over three to five runs; see, e.g., [31] and [44].

[7]Training neural networks to predict stock returns is analogous to fitting a high-dimensional linear model to the data because the last layer of a neural network is a linear layer, see KMZ. As stated in KMZ, one would expect the out-of-sample performance in predicting stock returns to improve as the neural network's complexity increases.

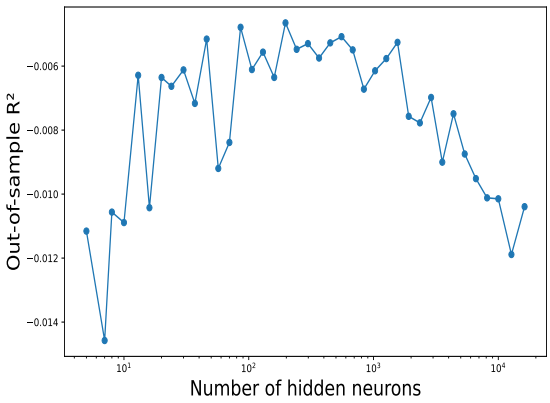


Figure 4.2: Stock return prediction with a neural network with data from NYSE, AMEX, and Nasdaq from 1990 to 2024. The x-axis shows model complexity (i.e., number of neurons), and the y-axis depicts out-of-sample R-squared. The results are averaged over 20 repetitions.

[43] show that increasing the complexity of the model in a finite sample setting tends to increase the variance of the parameter estimates, effectively amplifying any noise already present. Therefore, we conjecture that as model complexity increases, including more features allows the model to fit more complex patterns, but in the meantime, each feature becomes noisier, on average, due to re-estimation on the same data sample.[8] As the average level of noise in each feature increases with model complexity, the harm to model performance from increased noise may ultimately outweigh the gains from increased model complexity.

Our third motivating example in Figure 4.3 uses the same data and feature-generating method as that used in Figure 4.1, but now the simulated features are observed with noise. Figure 4.3 shows how the noise in the artificial features affects the link between complexity and the performance of the model. Both panels of the figure show the out-of-sample R-squared of the model as a function of model complexity, where the difference between the two panels is the noise in the features as model complexity increases.[9] In the left-hand panel, the noise level of each feature does not change as model complexity

[8]In practice, the increase in noise need not be uniform across features. When features are discovered sequentially rather than learned jointly, later-discovered features may be harder to estimate accurately from the same finite-sample data, resulting in higher measurement error. We explore this form of measurement-error growth in Section 4.1.2.

[9]For more details of adding the noise and training process, see Appendix B.2.

increases, and we see that the relationship between model performance and model complexity is 'approximately' monotonic. Nonetheless, it is clear that as the level of noise increases, all else being equal, the performance of the model decreases, that is, the performance function shifts down as the fixed noise level of the features increases.

However, in the right-hand panel, the noise in each feature increases linearly as the number of features increases, which provides a more realistic view of how expanding the feature space harms the average quality of each feature [37, 43]. Here, model performance initially improves with model complexity, but ultimately declines as the average noise level of each feature outweighs the benefit from model complexity. In line with the findings depicted in Figure 4.2, the results in the right-hand panel of Figure 4.3 show that as the quality of features continues to decrease, model complexity does not always increase the out-of-sample predictive performance of the model, so the potential gains from adding more features are limited, regardless of how features are learned.

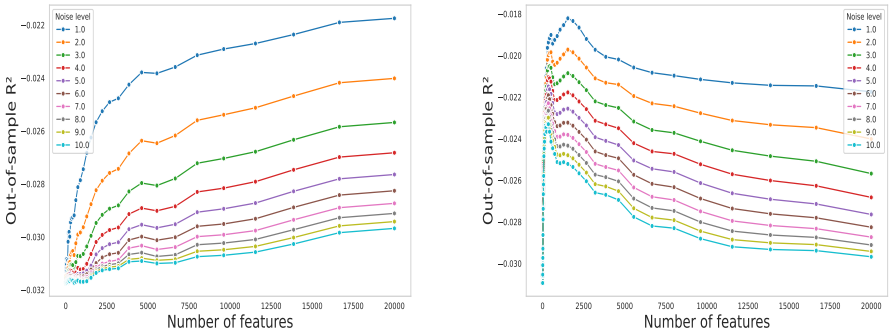


Figure 4.3: Complexity and model performance with simulated features observed with noise. In the left-hand panel, the noise level in each feature remains fixed as model complexity increases, whereas in the right-hand panel, the noise level for each feature increases linearly as model complexity increases, and the "noise scale" in the right-hand panel is the magnitude of this linear growth. The results are averaged over 50 repetitions.

This tradeoff between the benefits of model complexity and the challenges posed by noisy features raises a critical question: *How can a large set of predictive features be used when the "true features" are imperfectly observed due to infrastructure limitations?* To answer this question, this chapter proposes a theoretical framework to show the effect of noisy observations on the performance of complex prediction models. Specifically, we study how the

measurement noise in features affects the predictive power of feature-based return models, and we document what we refer to as the "limited virtue of complexity." Although increasing model dimensionality under appropriate shrinkage can improve return predictability—as proposed by the literature that favors complex models—we show that if there is noise in the observed features, as often is the case in real-world environments, increasing model complexity can exacerbate the detrimental impact of noise on the performance of return prediction and the Sharpe ratio of the portfolio.

We study this partial-observation setting in two ways. First, we assume that the average measurement error per feature increases with model size, that is, the more features the investor tries to learn, the noisier each becomes on average. This allows us to characterize analytically how the trade-off between additional predictive information and increasing measurement error determines the optimal level of complexity. Then, we study an alternative and more economically meaningful case in which later-discovered features are noisier than earlier-discovered features. In both cases, we show that an increase in model complexity can eventually become detrimental. In other words, even if additional features are available, an investor may not always wish to include them when measurement error degrades feature quality. This result complements the "virtue of complexity" with a pragmatic view of an optimal level of complexity beyond which prediction and portfolio performance deteriorate.

Our analysis does not contradict the theory developed in KMZ. Their results establish a virtue of complexity under the assumption that features are observed without error. Instead, our contribution studies how this relationship is altered when features are observed with measurement error, which is common in real-world data environments. In this sense, our results complement KMZ by showing that monotonic improvements in performance with model complexity need not hold once measurement error in features is taken into account.

Empirically, we verify our theoretical findings on U.S. equities by learning the features with an autoencoder, motivated by [65], who show that out-of-sample reconstruction error increases with feature dimensionality. Our empirical results show that out-of-sample R-squared and portfolio Sharpe ratio increase when model complexity initially increases, but these two metrics of performance eventually decrease as more features are added. This pattern is aligned with our theoretical results: when features are learned with error,

increasing complexity yields diminishing, and eventually negative, marginal gains.

A growing strand of empirical asset pricing research explores how high-dimensional methods and machine learning can uncover patterns in large, complex data sets. Similar to this chapter, much of this work builds on random matrix theory and high-dimensional asymptotics to demonstrate that suitably-tuned regularization—often in the form of ridge or related shrinkage approaches—can counter the risk of overfitting models when the number of features is of the same order as that of the sample size. Specifically, [22], [64], and [36] detail when, under certain assumptions, the out-of-sample performance of overparameterized or "ridgeless" models is good. Similarly, [32] and [44] show that advanced learning architectures can significantly improve out-of-sample forecasts. The contribution of this chapter to this strand of the literature is to incorporate *measurement noise* into the features and explore its interaction with model complexity. While the literature above typically treats predictors as noise-free, our work shows that noisy features can substantially affect the benefits of high-dimensional predictive models. Our empirical design also differs from traditional end-to-end prediction (e.g., [32, 29, 44, 20]): we separate representation from prediction by first denoising the observed feature set with an autoencoder and then estimating ridge forecasts on the reconstructed features. This separation clarifies how measurement error interacts with model complexity in practice.

Recent innovations in high-dimensional econometrics and empirical asset pricing have driven a proliferation of candidate features and anomalies [35, 25, 54]. Many of the proposed features rely on data sources subject to reporting lags or other data imperfections, which deteriorates the empirical performance of statistical methods across various frequencies [1, 63, 42, 19, 41]. Recently, researchers are examining how feature measurement errors interact with large-scale forecasting approaches. Here, we incorporate such errors directly into the features and provide a systematic way to assess the interaction of data imperfections and the performance of high-dimensional models.

Our study also connects to broader questions about partial observability and misspecified feature structures in finance. Classic works, including [57] and [15], show that excluding relevant variables can harm predictability and portfolio performance, while more recent studies (e.g., [45]) show that one may overlook some features due to practical limitations on data processing or modeling capacity. In a high-dimensional environment, noise in features

further complicates this landscape: even if there are additional features, the cumulative effect of noise may undermine the incremental gains from including them in the model. Consequently, our framework offers a more comprehensive and pragmatic perspective on feature proliferation, highlighting that pursuing additional features may be counterproductive when the search for them also raises the noise in each feature.

We also relate to a growing debate on the limits of model complexity in return prediction. [56] argues that the performance of high-dimensional models in KMZ may not reflect predictability, while [13] shows that averaging forecasts across random draws weakens the Sharpe-complexity relation. From a theoretical perspective, [23] highlights fundamental limits to learning in small samples, and [24] shows that complex models may instead rely on projection learning rather than recovering the true return function. We contribute to this debate by demonstrating how feature noise influences the effectiveness of model complexity in financial prediction.

4.1.1 Model misspecification

We now extend the noisy-feature model from the previous chapter to a partial-observation setting. The data-generating process, timing strategy, ridge estimator, and regularity assumptions are the same as in Chapter 3 unless stated otherwise. The key difference is that the investor no longer observes the full feature vector. Instead, she observes only a subset of the true features, and these observed features may be measured with noise. While Chapter 3 isolates the effect of measurement error under full observation, the partial-observation setting considered here allows us to study how feature noise interacts with model complexity as the investor expands the set of observed features.

This setting is motivated by the model misspecification problem studied in KMZ, where the investor cannot observe the full set of features but the observed features are measured without error. Suppose that the true DGP is given by (3.4) with P features. The investor observes only P_1 of these features. Denote the observed true features by $F_{t,(1)} = (F_{i,t})_{i=1}^{P_1}$, and the remaining unobserved features by $F_{t,(2)} = (F_{i,t})_{i=P_1+1}^{P}$. The corresponding covariance structure of $F_{t,(1)}$ and $F_{t,(2)}$ is

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

The investor uses the misspecified ridge estimator

$$\hat{\beta}_1 = \left(z\, I_{P_1} + \hat{\Psi}_{1,1} \right)^{-1} \frac{1}{T} \sum_t F_{t,(1)}\, R_{t+1}\,,$$

where

$$\hat{\Psi}_{1,1} = \frac{1}{T} \sum_t F_{t,(1)}\, F_{t,(1)}^{\intercal}\,,$$

to estimate the true predictive coefficients β, so the timing strategy is given by

$$\hat{\pi}_{1,t} = F_{t,(1)}^{\intercal}\, \hat{\beta}_1\,. \tag{4.1}$$

KMZ prove that, under certain regularity conditions, the out-of-sample R-squared and Sharpe ratio of (4.1), both with the optimal shrinkage parameter z_*, are monotonically increasing and concave in the ratio $q = P_1/P$. KMZ refers to this result as the *Virtue of Complexity*; see Theorem 1 in KMZ.

We investigate if this monotonic and concave relationship holds when features are observed with noise. Specifically, suppose the investor observes both an incomplete and noisy set of features, denoted as $\tilde{F}_{t,(1)} = (\tilde{F}_{i,t})_{i=1}^{P_1} = F_{t,(1)} + \Phi_{1,1}^{1/2}\, Y_{t,(1)}$ and employs the estimator

$$\hat{\tilde{\beta}}_1 = \left(z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1} \right)^{-1} \frac{1}{T} \sum_t \tilde{F}_{t,(1)}\, R_{t+1}\,,$$

where

$$\hat{\tilde{\Psi}}_{1,1} = \frac{1}{T} \sum_t \tilde{F}_{t,(1)}\ \tilde{F}_{t,(1)}^{\intercal}\,,$$

so the corresponding timing strategy is

$$\hat{\tilde{\pi}}_{1,t} = \tilde{F}_{t,(1)}^{\intercal}\, \hat{\tilde{\beta}}_1\,. \tag{4.2}$$

We analyze the properties of the out-of-sample R-squared and Sharpe ratio of (4.2) evaluated at the optimal shrinkage parameter $z_{*,1}$. Intuitively, the monotonicity result established in KMZ may not hold when there is measurement error in the features. As more features are observed, the accumulation of noise could potentially outweigh the benefits of incorporating additional features, thereby adversely affecting the Sharpe ratio.

To investigate this, we assume that Ψ and $\Phi_{1,1}$ are diagonal matrices, and the covariance matrices of $\tilde{F}_{t,(1)}$ is $\tilde{\Psi}_{1,1} = \Psi_{1,1} + \Phi_{1,1}$ because the investor observes a partial model. However, instead of assuming $\Phi_{1,1} = a\, I_{P_1}$, we assume $\Phi_{1,1} = a\, q\, I_{P_1}$, where $q = P_1/P$ is the ratio of observed features. This is an empirically motivated assumption where the investor repeatedly discovers new features using the same set of available information, but as the number of features increases, the average quality of these features deteriorates because of the limited size of data.[10] In this case, we have the following corollary:

Corollary 2. *Let the assumptions in Theorem 1 hold, and assume that* $\Psi = I_P$ *and* $\Phi_{1,1} = a\, q\, I_{P_1}$*, where* $q = P_1/P$ *and* $P_1 < P$*. Then, as* $P, T, P_1 \to \infty$ *and* $P/T \to c, P_1/P = q$*, we have*

$$T^{-1}\ \mathrm{tr}\left((z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1})^{-1}\, \tilde{\Psi}_{1,1}\right) \xrightarrow{\mathbb{P}} \xi(z; c\, q; a\, q)\,.$$

The timing strategy's expected return is

$$\mathcal{E}(z; c; a; q) = \frac{b_*\, c^{-1}\, \xi(z; c\, q; a\, q)}{(1 + \xi(z; c\, q; a\, q))\,(1 + a\, q)}\,,$$

and its expected leverage is

$$\begin{aligned}\mathcal{L}(z; c; a; q) = (1 + b_*)\ &(\xi(z;\ c\, q;\ a\, q) + z\, \partial\xi(z;\ c\, q;\ a\, q)/\partial z) \\ &- \frac{b_*\, c^{-1}\, \partial\xi(z;\ c\, q;\ a\, q)/\partial z}{(1 + \xi(z;\ c\, q;\ a\, q))^2} \\ &- \frac{2 b_*\, c^{-1}\ (\xi(z;\ c\, q;\ a\, q) + z\, \partial\xi(z;\ c\, q;\ a\, q)/\partial z)\ \xi(z;\ c\, q;\ a\, q)}{(1 + \xi(z;\ c\, q;\ a\, q))\,(1 + a\, q)}\,.\end{aligned}$$

The R-squared and Sharpe ratio of the timing strategy are

$$R^2(z; c; a; q) = \frac{2\, \mathcal{E}(z; c; a; q) - \mathcal{L}(z; c; a; q)}{1 + b_*}\,, \tag{4.3}$$

and

$$SR(z; c; a; q) = \frac{\mathcal{E}(z; c; a; q)}{\sqrt{2\, \mathcal{E}(z; c; a; q)^2 + (1 + b_*)\ \mathcal{L}(z; c; a; q)}}\,. \tag{4.4}$$

[10] An example of this assumption would be using auto-encoder as a de-noising method. [65] proved that the out-of-sample error bound of the de-noised features will increase with the number of features fed into the auto-encoder, see Theorem 2 of [65] for more details. For tractability, in our analysis, we assume this relationship is linear.

Next, we derive results regarding the optimal choice of the ridge parameter z and the properties of the R-squared and Sharpe ratio of the timing strategy as a function of $q = P_1/P$.

Proposition 3. *Let the assumptions in Corollary 2 hold. Then the following statements hold:*

(i) Both the out-of-sample R-squared $R^2(z; c; a; q)$ and Sharpe ratio $SR(z; c; a; q)$ of the timing strategy are monotonically increasing in z for $z < z_{*,1} = c\,(a\,q + 1)\,(a\,b_*\,q + a\,q - b_*\,q + b_* + 1)/b_*$ and monotonically decreasing in z for $z > z_{*,1}$.

(ii) With the optimal choice $z_{*,1}$, for both out-of-sample R-squared $R^2(z; c; a; q)$ and Sharpe ratio $SR(z; c; a; q)$, we have,

- *If $a > 1$ and $c > \frac{a\,b_*}{(a-1)\,(a^2\,(b_*+1)+a+b_*)}$, there exists a $q_* \in (0, 1)$ such that both $R^2(z; c; a; q)$ and $SR(z; c; a; q)$ are monotonically increasing in q for $q < q_*$ and monotonically decreasing in q for $q > q_*$.*

- *Otherwise, $R^2(z; c; a; q)$ and $SR(z; c; a; q)$ are monotonically increasing in q.*

The first result in Proposition 3 shows that when noise scales linearly with the complexity of our partial information model, there exists an optimal shrinkage that maximizes the Sharpe ratio of our timing strategy.

The second result in Proposition 3, on the other hand, shows that the results in KMZ do not necessarily hold when noise scales linearly with the complexity of the partial model. When the value of the scaling coefficient in noise a is small, investors are incentivized to discover more features and ultimately approach a full-information model in which all relevant features are discovered, despite the increasing level of noise. However, when the value of a is large and when the value of c, the true complexity of the world, is also large, there exists an optimal level of complexity q_* in a partial information model where investors no longer seek to find more features beyond this level of complexity.[11]

[11] One can prove that there is also an optimal level of complexity when noise grows in proportion to the logarithm of the model's complexity (i.e., $a \log(1 + q)$). Thus, there are other ways in which noise increases as complexity increases.

An intuitive explanation of the second result in Proposition 3 is that when the world is complex (i.e., when c is large), the explanatory power of each individual feature is very weak. Due to limited access to data, the increasing level of noise associated with discovering more features may deteriorate the average quality of features to a point that it adversely affects the Sharpe ratio of the timing strategy.

4.1.2 Growing noise along the feature-discovery process

In the previous section, we assumed that $\Phi_{1,1} = a\, q\, I_{P_1}$, so the level of measurement error scales linearly with the observed-feature ratio $q = P_1/P$. While analytically convenient, this assumption is restrictive: it forces all observed features to share the same noise variance and ties noise mechanically to q.

In this section, we consider an alternative and more economically natural specification in which noise increases along the feature-discovery sequence. Specifically, we assume

$$\Phi_{1,1} = \mathrm{diag}\big(f(1/P), f(2/P), \ldots, f(P_1/P)\big)\,,$$

where $f : [0,1] \to \mathbb{R}_+$ is continuous and increasing. This formulation reflects a natural feature-discovery process in which earlier features are easier to identify and therefore measured with greater precision, while later features are increasingly difficult to extract and consequently subject to higher measurement error. In this case, we have the following corollary:

Corollary 3. *Let the assumptions in Theorem 1 hold, and assume that* $\Psi = I_P$ *and* $\Phi_{1,1} = \mathrm{diag}(f(1/P), f(2/P), \cdots, f(P_1/P))$*, where* $P_1 < P$ *and* $f : [0,1] \to \mathbb{R}_+$ *is the noise function and is continuous and increasing on the interval* $[0,1]$*. Then, as* $P, T, P_1 \to \infty$ *and* $P/T \to c, P_1/P = q$*, we have*

$$T^{-1}\ \mathrm{tr}\left((z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1})^{-1}\, \tilde{\Psi}_{1,1}\right) \xrightarrow{\mathbb{P}} \xi_1(z; c; q)\,,$$

where

$$\xi_1(z; c; q) = \frac{1 - z\, m_1(-z; c; q)}{(c\, q)^{-1} - 1 + z\, m_1(-z; c; q)}\,,$$

and $m_1(u;c;q)$ *solves*[12]

$$m_1(u;c;q) = \frac{1}{1-c\,q-c\,q\,u\,m_1(u;c;q)}\, m_{\tilde{\Psi}_{1,1}}\left(\frac{u}{1-c\,q-c\,q\,u\,m_1(u;c;q)}\right), \tag{4.5}$$

where

$$m_{\tilde{\Psi}_{1,1}}(u) = \lim_{P_1\to\infty} P_1^{-1}\ \mathrm{tr}\left((\tilde{\Psi}_{1,1}-u\,I_{P_1})^{-1}\right) = \frac{1}{q}\int_0^q \frac{1}{1+f(x)-u}dx \tag{4.6}$$

is the limiting Stieltjes transform of the eigenvalue distribution of $\tilde{\Psi}_{1,1}$.

The timing strategy's expected return is

$$\mathcal{E}(z;c;q) = \frac{b_*\,q\,m_1(-z;c;q)}{1+\xi_1(z;c;q)},$$

and its expected leverage is

$$\mathcal{L}(z;c;q) = (1+b_*)\,(\xi_1(z;c;q)+z\,\partial\xi_1(z;c;q)/\partial z) - \frac{b_*\,c^{-1}\,\partial\xi_1(z;c;q)/\partial z}{(1+\xi_1(z;c;q))^2} - \frac{2b_*\,q\,(\xi_1(z;c;q)+z\,\partial\xi_1(z;c;q)/\partial z)\,m_1(-z;c;q)}{(1+\xi_1(z;c;q))}.$$

The R-squared and Sharpe ratio of the timing strategy are

$$R^2(z;c;q) = \frac{2\,\mathcal{E}(z;c;q)-\mathcal{L}(z;c;q)}{1+b_*}, \tag{4.7}$$

and

$$SR(z;c;q) = \frac{\mathcal{E}(z;c;q)}{\sqrt{2\,\mathcal{E}(z;c;q)^2+(1+b_*)\,\mathcal{L}(z;c;q)}}. \tag{4.8}$$

Relative to Corollary 2, the expressions for the out-of-sample R-squared and Sharpe ratio are substantially more complex. In particular, the expression for $m_{\tilde{\Psi}_{1,1}}(u)$ in equation (4.6) will not be, in general, in closed form, in which cases the expression for m_1 in equation (4.5) will not be in closed form either and must be solved numerically. As a result, there is not always an explicit analytical characterization for the optimal ridge parameter z. Nevertheless, the model can still be analyzed numerically. In particular, we can study

[12]Existence and uniqueness of such m_1 follow from standard random matrix theory results, see, e.g, [60] and [5].

the behavior of the out-of-sample R-squared (4.7) and Sharpe ratio (4.8) numerically and assess how the "virtue of complexity" depends on the rate at which feature quality deteriorates as model complexity increases.

In Figure 4.4, we consider the parametric family

$$f(x) = 10\,x^{\alpha},$$

with $\alpha \in \{0, 0.5, 1, 10\}$, and plot the out-of-sample R-squared and Sharpe ratio for different values of the ridge shrinkage parameter z. The parameter α controls how quickly noise increases as more features are discovered, from flat noise growth ($\alpha = 0$) to rapidly increasing noise in later-discovered features ($\alpha = 10$).

Figure 4.4 reveals a clear and economically intuitive pattern. When noise grows slowly with feature complexity ($\alpha = 0$ or $\alpha = 0.5$), the Sharpe ratio is monotonically increasing in q, with a concave shape that closely resembles the "virtue of complexity" result in KMZ. In this case, adding features is beneficial. By contrast, once noise increases sufficiently rapidly with model complexity ($\alpha = 1$ and $\alpha = 10$), an inverted U-shaped Sharpe ratio emerges. Performance initially improves as additional features are added, but beyond a given level of complexity it begins to decline, reflecting the fact that the marginal features are increasingly dominated by measurement error.

The behavior of the out-of-sample R-squared aligns closely with these results predicted by our theory when noise in features starts to increase for later-discovered features. When noise grows slowly, the R-squared does not show an interior optimum for q and is largely determined by the ridge parameter z, emphasizing the role of regularization in high-dimensional settings. When noise grows more rapidly, the R-squared exhibits a clear interior maximum. This is consistent with the pattern observed for the Sharpe ratio where performance initially improves with complexity but eventually declines once noise in marginal features dominates the benefits of additional features.

Overall, Figure 4.4 shows that the monotonic and concave "virtue of complexity" result in KMZ does not generally hold when the quality of marginal features deteriorates sufficiently quickly as model complexity increases. When noise grows slowly, adding more features is beneficial. When noise grows rapidly, however, investors face an economically meaningful trade-off: increasing model complexity improves performance only up to a certain level,

after which measurement error in marginal features outweighs the benefits of adding new features.

4.2 Empirical analysis

In this section, we investigate our theoretical results empirically. Inspired by [65], we first apply an autoencoder to reduce noise in the features, then use ridge regression to forecast returns and construct the timing strategy. We use out-of-sample R^2 and the Sharpe ratio of the timing strategy to evaluate the model.

4.2.1 Data

Following [31], we collect monthly stock returns from CRSP for NYSE, AMEX, and Nasdaq firms from March 1957 to December 2024. We use the Treasury bill rate as a proxy for the risk-free rate to compute individual excess returns. We use 94 firm characteristics from [30] and 8 macroeconomic predictors from [63]. As in [31], every month we rank each stock characteristic cross-sectionally and map the ranks to the $[-1, 1]$ interval, as in [49] and [27]. We define the stock-level feature vector for firm i at time t as

$$F_{i,t} = x_t \otimes w_{i,t}\,, \tag{4.9}$$

where $\otimes$ denotes the Kronecker product, $w_{i,t}$ is a 94×1 vector of firm characteristics, and x_t is a 9×1 vector of macro predictors (including a constant). Thus, the dimension of $F_{i,t}$ is $P = 94 \times 9 = 846$. This method of feature construction is the same as that in [31], except that we exclude industry dummies because they do not require denoising.

To ensure enough training data, we keep only stocks with at least 30 years of data. We randomly select 100 stocks from the eligible universe for our analysis.

4.2.2 Autoencoder as a noise-reduction tool

To evaluate the impact of model complexity on forecasting performance when features are measured with error, we use autoencoders to de-noise the features. This approach reflects the model misspecification case introduced in Section 4.1.1. Our approach is motivated by the practical observation that investors often face a high-dimensional feature space constructed from firm

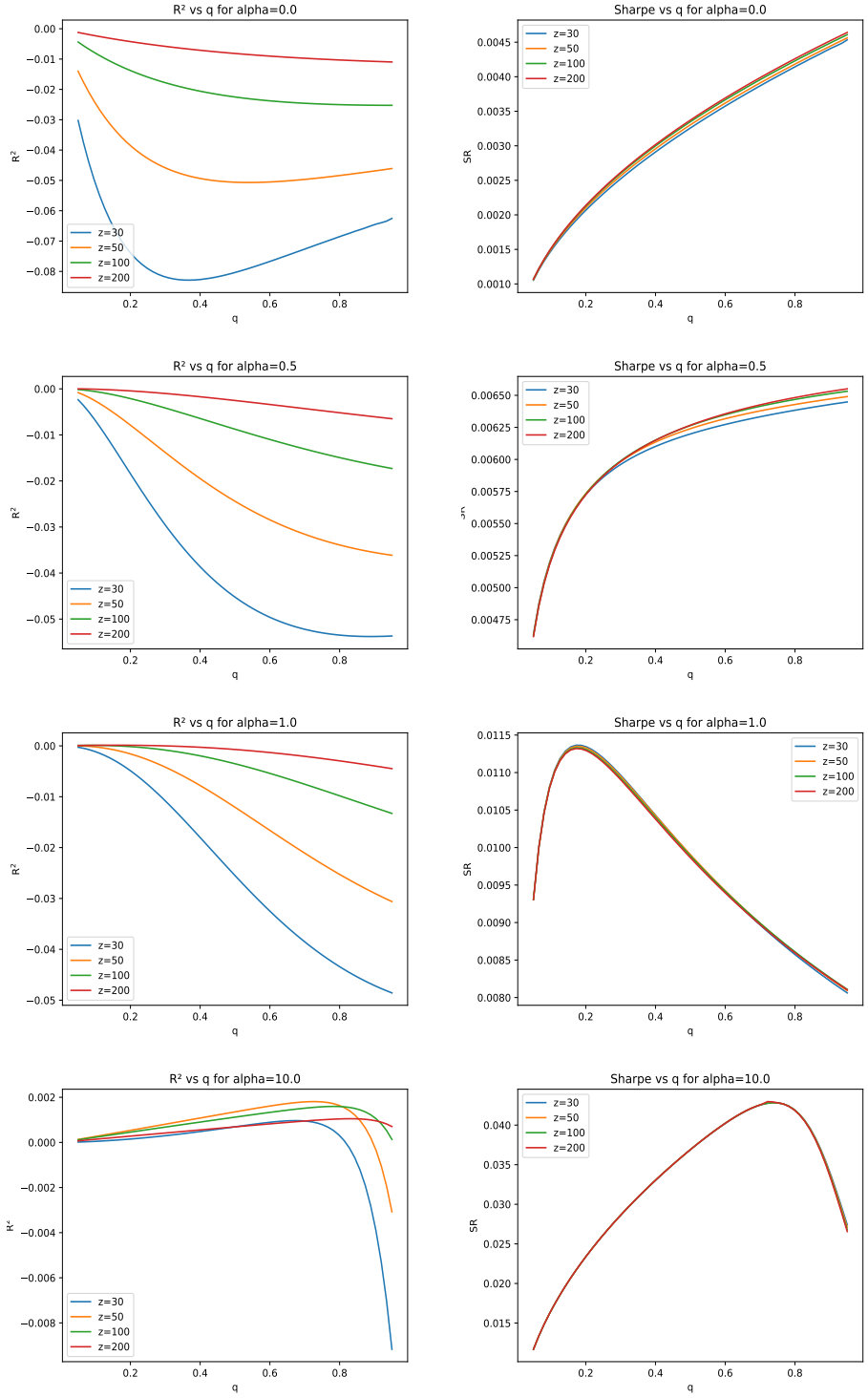


Figure 4.4: Out-of-sample R-squared and Sharpe ratio for $f(x) = 10\,x^{\alpha}$ under different values of α and ridge parameter z.

characteristics, macroeconomic variables, or other signals, and they must decide how many features to incorporate into their models. While including more features may help capture relevant return predictors, it also introduces more noise, particularly when the estimation is performed using flexible, data-driven methods such as deep learning. This trade-off becomes especially important in financial contexts, where high dimensionality often coincides with limited data availability.

Specifically, we assume that for each stock i, the excess returns follow a linear regression model

$$R_{t+1,i} = \beta_i^\intercal \, F_{i,t}^* + \varepsilon_{t+1,i} \,,$$

where $F_{i,t}^* \in \mathbb{R}^P$ represents the true, unobserved signals and $\varepsilon_{t+1,i}$ is the innovation term in the returns. In our empirical analysis, we treat the feature vector $F_{i,t}$ constructed from 94 firm characteristics and 8 macroeconomic variables (as defined in (4.9)) as the observable counterpart to $F_{i,t}^*$. That is, $F_{i,t} \in \mathbb{R}^P$ with $P = 846$ is our observed noisy feature vector and serves as a noisy proxy for the true feature vector.

The available features may be noisy, so investors may wish to estimate a lower noise representation of $F_{i,t}$ before forecasting the returns. To achieve this, we use a deep autoencoder to recover useful patterns from noisy high-dimensional data and generate a de-noised version of the features.

Specifically, let $F_{i,t}^{(P_1)} \in \mathbb{R}^{P_1}$ denote a fixed subset of P_1 features from the full vector $F_{i,t}$. We treat P_1 as a proxy for model complexity and vary $P_1 \in \{100, 200, \cdots, 846\}$, where increasing the value of P_1 expands the feature space but also makes the de-noising task more difficult due to the higher dimensionality of the feature space.

Theorem 2 of [65] provides critical motivation for this empirical design. The authors show that while autoencoders can achieve low in-sample reconstruction error when trained on noisy features, their out-of-sample reconstruction error increases with the number of input features. In particular, if P_1 denotes the input dimension and T the training sample size, the out-of-sample error bound is of order P_1/T. This result implies that although richer feature sets contain more predictive information, the learning and de-noising of features becomes increasingly difficult as complexity increases, especially when the training sample size, T, is fixed. In our setting, this suggests a fundamental trade-off: adding features may increase the amount of useful in-

formation captured by the model, but it also introduces reconstruction error, potentially degrading prediction performance. Thus, to analyze this trade-off, we adopt a two-step approach: first de-noise the observed features with an autoencoder, then estimate a ridge regression using the de-noised features and predict the returns and construct the timing-strategy. For the detailed training process and implementation, see Appendix B.3.

4.2.3 Out-of-sample performance of market timing strategy

Figure 4.5 reports the average out-of-sample R^2 and Sharpe ratio of the market timing strategy that uses the features de-noised by the autoencoder. Results are averaged over all stocks and five repetitions for feature dimensions P_1, where $P_1 \in \{100, 200, \cdots, 846\}$ is the number of features, which serves as a proxy for model complexity.

As Figure 4.5 shows, the out-of-sample R^2 and Sharpe ratio both increase as the feature dimension increases from low to moderate, but eventually decline once the model becomes more complex. Performance improves at first as additional features help capture predictive structure, yet the benefit is offset at higher dimensions, where the reconstruction error dominates and leads to weaker forecasts. This pattern illustrates the limited virtue of complexity: incorporating more features helps to improve out-of-sample performance, but the benefit diminishes as the autoencoder faces increasing difficulty in accurately de-noising high-dimensional features. In fact, beyond a certain point, adding more features reduces out-of-sample prediction performance and portfolio Sharpe ratio.

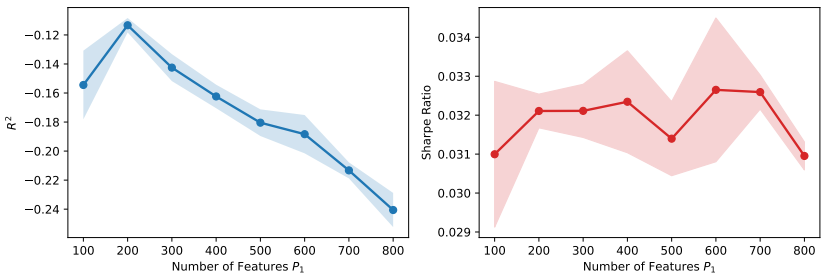


Figure 4.5: Out-of-sample R^2 (left panel) and Sharpe ratio (right panel) of the market timing strategy using de-noised features from the autoencoder. Results are averaged across all stocks and five repetitions. Shaded bands represent ± 1 standard error across the five repetitions.

4.3 Conclusion

We study the role of model complexity in portfolio optimization when predictive features are imperfectly observed. Across Chapters 3 and 4, we develop a framework that highlights how the benefits of complexity depend critically on the quality of the features.

In Chapter 3, we consider a setting in which investors observe the full set of features with measurement error. We show that, under appropriate regularization, portfolio performance can be improved despite noise. However, the gains from reducing noise are *convex*: investors with already high-quality infrastructure benefit more from marginal improvements, whereas those facing substantial noise experience diminishing returns. This result provides a refined view of the "virtue of complexity," emphasizing the importance of feature quality.

In Chapter 4, we extend this framework to a partial-observation setting in which investors observe only a subset of the true features, and these features may themselves be noisy. In this environment, increasing model complexity introduces a trade-off between incorporating additional predictive information and amplifying measurement error. We show that this trade-off gives rise to a *limited virtue of complexity*: performance initially improves with complexity but may eventually deteriorate when noise dominates. As a result, there exists an optimal level of model complexity that depends on both data quality and available resources.

Taken together, our findings reconcile the classical preference for parsimonious models with recent evidence supporting highly complex models. While complexity can enhance predictive performance, its benefits are not universal and depend on the interaction between model dimensionality and noise in the features. These results provide a more pragmatic perspective for investors, suggesting that the optimal level of complexity is shaped not only by the richness of available data but also by the ability to measure and process the data accurately.

Chapters 3 and 4 study how imperfect feature quality can limit the benefits of model complexity. The next chapter turns to a complementary question: whether improvements in empirical performance as model complexity increases necessarily reflect better learning of economic information. In particular, we examine whether the empirical virtue of complexity can arise

mechanically from the construction of high-dimensional features, even when these features do not necessarily contain additional economic information. This provides a complementary perspective to the noisy-feature framework developed in the previous two chapters by showing that observed gains from complexity may reflect not only the quality of the underlying features, but also the mechanics of the estimator itself.

5 | The Mechanical Virtue of Random Fourier Features and Model Complexity

5.1 Introduction

Chapters 3 and 4 study how imperfect feature quality can limit the benefits of model complexity. In this chapter, we consider a complementary question: whether an observed improvement in model performance as complexity increases necessarily reflects better learning of economic information. Specifically, we examine whether the empirical "virtue of complexity" can arise mechanically from the construction of high-dimensional random features and the resulting estimator, even in settings where meaningful learning is limited.

The recent debate on the "virtue of complexity" starts from a striking empirical claim: in forecasting stock returns, highly parameterized prediction models with more parameters than observations often outperform traditional low-dimensional specifications, see [47], which we refer to as KMZ throughout. In their experiment, KMZ generate random Fourier features (RFFs) from the [63] predictors, which they use to estimate a ridge regression to forecast CRSP market returns, and then use a market-timing investment strategy to assess the performance of an overparameterized model. They show that thousands of such nonlinear features deliver higher out-of-sample R^2 of the return forecast, higher mean timing returns, and higher Sharpe ratios than those of much simpler models, even when the training window contains only twelve monthly observations. KMZ interpret these findings as evidence that complex, overparameterized models can learn economically meaningful structure.

[56] challenges this interpretation. Nagel uses a kernel-ridge representation of the KMZ-RFF model to show that in the ridgeless, high-complexity limit, the market-timing strategy can be written as a particular form of volatility-timed momentum. When predictors are persistent and the training window is short, the kernel places most weight on recent observations of the same return series, so the timing strategy effectively extrapolates its own past with positive weights. Nagel uses this observation, along with simulations, to argue that the small-T results in KMZ do not demonstrate learning about expected returns.

Several other papers contribute to this debate. [13] argues that the upward-sloping Sharpe-complexity relation in KMZ disappears when forecasts are averaged across random draws. As studied in Chapters 3 and 4 and in [17], measurement error in features can also limit or overturn the virtue of complexity. Finally, [23] combines a kernel-level analysis of standardization with information-theoretic lower bounds, placing KMZ's training window in a signal-limited regime.

[48] respond that, while these critiques raise useful conceptual questions, they do not overturn the core KMZ message that large models can outperform small ones. Their response stresses the distinction between nominal and effective complexity, the role of implicit regularization, and the idea of limits to learning when P/T is large, where P is the number of features and T is the number of observations. They present additional simulations in which virtue-of-complexity patterns persist under various data constructions.

This chapter contributes to the debate by examining how KMZ use RFFs as an example of their theory; specifically, we follow their RFF-based example design in its most extreme form. While KMZ estimate each regression on a rolling window of $T = 12$ observations, we consider a rolling window with $T = 1$ observation, so each regression is fitted to one data point. In our setting, the model cannot extract any dynamic information from the time series, and the only randomness stems from the RFFs. This environment provides a clean test of whether the empirical "virtue of complexity" reflects learning or simply a mechanical property of the estimator when the number of RFFs increases.

We show that the main empirical patterns emphasized by KMZ, including rising out-of-sample R^2 of the return forecast, increasing returns, and increasing Sharpe ratios as the number of features P grows, persist even when the rolling window contains only one data point, i.e., when $T = 1$. To interpret this fact, we analyze the estimator in a fixed-data, random-feature framework and derive expressions for the expected performance measures. These results reveal that the monotonic relationships between model size and forecast performance arise mechanically from the structure of the random features rather than from learning.

To better understand this result, we run a simple sign-flip experiment in which we reverse the sign of some returns. This keeps the predictors the same but changes the relation between past and future returns. We find

that this change reverses the direction of the relationship between model complexity and performance. In other words, if performance increases with complexity in the original data, it will decrease with complexity after the sign flip. This shows that the pattern emphasized by KMZ can be reversed by a small change in the data, which suggests that it does not come from learning.

We also study the case where the training window contains more than one observation ($T > 1$). In this setting, the patterns become more complex and can vary with the choice of T. However, the main insight from the sign-flip experiment remains: changing the sign of returns still reverses the relationship between model complexity and performance. This suggests that the mechanism we highlight is not specific to the extreme case $T = 1$.

Our findings do not contradict the theory developed in KMZ. Their theoretical result is correct under their stated assumptions, and our analysis does not claim that overparameterized models cannot learn in richer environments. However, using RFFs as an illustrative example to showcase the benefits of overparameterized models may not provide conclusive evidence that such models outperform simpler alternatives when applied to market data. Rather, our contribution is diagnostic in nature. We show that when macroeconomic predictors and returns are held fixed, increases in performance measures (such as out-of-sample R^2, average strategy returns, and Sharpe ratios) as the number of features grows can arise mechanically from the structure of the RFF model. As a result, monotonic relationships between model complexity and performance observed in this setting are not, by themselves, conclusive evidence of learning in financial data.

The remainder of the chapter is as follows. Section 5.2 introduces the empirical results when $T = 1$ and provides the theoretical characterization of the performance metrics of the RFF regression, including closed-form expressions for the expected out-of-sample R^2 of the return forecast and for the expected return of the market-timing strategy, as the number of features increases. Section 5.3 provides simulation evidence and a modified version of the KMZ experiment to highlight the mechanical nature of the observed patterns between model performance and model complexity, and also considers the case where $T > 1$. Section 5.4 concludes the chapter and proposes future work.

5.2 Model

5.2.1 Empirical properties of RFF with $T = 1$

In their work, KMZ generate RFFs from the [63] predictors, estimate a ridge regression to forecast CRSP market returns and design a market-timing investment strategy, where they use out-of-sample R^2, average return and Sharpe ratio of the market-timing strategy to assess model performance. The central empirical finding in KMZ is that out-of-sample performance improves monotonically with model complexity, where model complexity is measured by the number of RFFs relative to the number of observations. Their experiment uses rolling windows of $T = 12$ months, and here we examine the extreme case in which the training window contains only one observation, i.e., $T = 1$. In this environment, the model cannot extract dynamic information from the data, and one cannot use the within-window standardization step in KMZ because each feature has only one observation in the training set.[1] Therefore, the deliberately extreme case $T = 1$ provides a transparent setting in which to assess whether the empirical "virtue of complexity" reflects learning or instead arises mechanically.

Let G_t denote the 15 predictor variables in [63] which are used in KMZ, and let R_t denote the CRSP market return used in KMZ. In their work, they draw $\omega_i \sim N(0, I_{15})$ and generate feature vector

$$F_t = (F_{1,t}^\intercal, \cdots, F_{P,t}^\intercal)^\intercal , \tag{5.1}$$

where each feature $F_{i,t}$ is generated with the nonlinear transformation

$$F_{i,t} = \begin{bmatrix} \sin(\gamma\, \omega_i^\intercal\, G_t) \\ \cos(\gamma\, \omega_i^\intercal\, G_t) \end{bmatrix} , \qquad i = 1, \ldots, P/2 , \tag{5.2}$$

so that the total number of features equals P. Here, I_{15} is the 15×15 identity matrix and γ is the bandwidth parameter.

To study how model performance varies with model complexity when $T = 1$, we follow the experiment design in KMZ and adapt it to our setting, where we vary P from 10 to 20000[2]. Therefore, we use $\{R_t, F_{t-1}\}$ to run the ridge

[1] As shown in equation (5.1), the dimension of each feature vector is $P \times 1$. KMZ's standardization step relies on computing the standard deviation of each feature $F_{i,t}$ within the training window, which is not defined with only one observation.

[2] The key differences relative to KMZ are that estimation is based on one observation

Algorithm 1: RFF-Based Forecasting with a Single-Observation Training Window

1. **RFF Library Generation.** Use (5.2) to construct a library of 20000 RFFs.
2. **Feature Selection.** For a given number P of features, use the *first* P *entries* of the RFF library, i.e., the first $P/2$ sine-cosine pairs to form the predictor vector F_t.
3. **Estimation with** $T = 1$**.** At each month t, use the single observation $\{R_t, F_{t-1}\}$ and ridge penalty $\lambda = 1$ to estimate the ridge regression and obtain the estimate $\hat{\beta}_t$ with (5.3). Next, compute the out-of-sample return forecast $\hat{R}_{t+1} = \hat{\beta}_t^\intercal F_t$ and the return $\hat{R}_{t+1} R_{t+1}$ of the timing strategy.
4. **Performance Evaluation.** Use the resulting sequence of forecasts and strategy returns to compute the out-of-sample R^2, average return, volatility, and Sharpe ratio of the timing strategy.
5. **Monte Carlo Averaging.** Repeat steps 1–4 for 500 independent draws of the RFF library and average the results for each value of P.

regression and obtain the estimate

$$\hat{\beta}_t = (F_{t-1} F_{t-1}^\intercal + \lambda I)^{-1} F_{t-1} R_t \,, \tag{5.3}$$

with which we compute the out-of-sample return forecast $\hat{R}_{t+1} = \hat{\beta}_t^\intercal F_t$ and construct the return $\hat{R}_{t+1} R_{t+1}$ of the timing strategy. Algorithm 1 summarizes the procedure to generate return forecasts and strategy returns.[3]

Figure 5.1 reports the results of this experiment. Strikingly, even with only one observation in each training window, the performance improves monotonically with the number of features: the out-of-sample R^2 of the return forecast, the mean return of the timing strategy, and the Sharpe ratio all increase with model complexity. This pattern is in accordance with the main empirical result of KMZ despite the fact that $T = 1$ leaves little scope for learning from the data.

The emergence of the same pattern between model performance and number of features in such a simplified environment suggests that the source of the

when constructing the ridge regression (5.3), and that we do not have the within-window standardization step in KMZ. See Algorithm 1 for details.

[3]Algorithm 1 uses the ridge parameter 1 for clearer illustration.

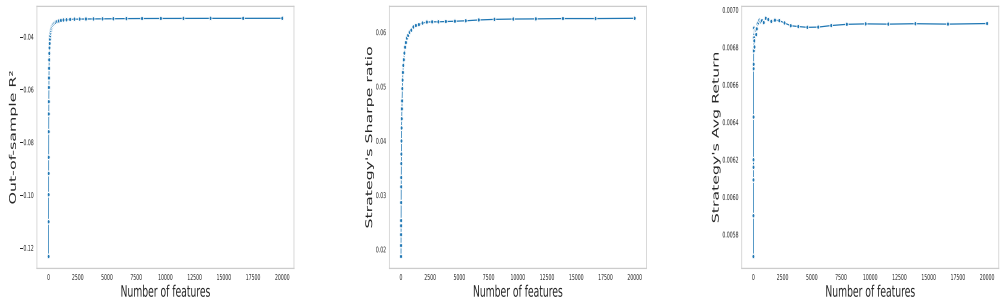


Figure 5.1: Performance metrics as functions of number of features P, with number of observations $T = 1$, and ridge penalty $\lambda = 1$.

"virtue of complexity" cannot be predictive structure uncovered by nonlinear estimation. Instead, the pattern must arise from a mechanical property of the high-dimensional estimator itself. The next subsection develops this point formally and shows precisely how performance metrics of the estimator (5.3) simplify when $T = 1$.

5.2.2 A closer look at the theoretical mechanism of the RFF method

A key observation is that the empirical exercise in the previous subsection repeatedly generates RFFs while holding the return data R_t and macroeconomic predictors G_t fixed. Thus, across repetitions, the only source of randomness is the set of RFF weights $\{\omega_j\}$. Next, we formalize this setting to explain why increasing the number of features produces these results.

Consider N fixed training-testing pairs

$$\{(x_{i,\text{train}},\, y_{i,\text{train}}),\ (x_{i,\text{test}},\, y_{i,\text{test}})\}_{i=1}^{N},$$

where $x_{i,\text{train}}, x_{i,\text{test}} \in \mathbb{R}^d$ and $y_{i,\text{train}}, y_{i,\text{test}} \in \mathbb{R}$. These pairs correspond to the predictor and return observations used in each window of the empirical experiment. Importantly, these data points are treated as constants, matching the empirical setting where the macroeconomic predictors and CRSP returns do not vary across repetitions of the experiment. The only randomness comes from the construction of RFFs.

For a given number of features P and bandwidth parameter γ, we generate

random weights $\omega_1, \ldots, \omega_{P/2} \sim \mathcal{N}(0, I_d)$, and define the RFF vectors

$$z_{i,\text{train}} = \begin{pmatrix} \cos(\gamma\, \omega_1^\intercal\, x_{i,\text{train}}) \\ \sin(\gamma\, \omega_1^\intercal\, x_{i,\text{train}}) \\ \vdots \\ \cos(\gamma\, \omega_{P/2}^\intercal\, x_{i,\text{train}}) \\ \sin(\gamma\, \omega_{P/2}^\intercal\, x_{i,\text{train}}) \end{pmatrix}, \qquad z_{i,\text{test}} = \begin{pmatrix} \cos(\gamma\, \omega_1^\intercal\, x_{i,\text{test}}) \\ \sin(\gamma\, \omega_1^\intercal\, x_{i,\text{test}}) \\ \vdots \\ \cos(\gamma\, \omega_{P/2}^\intercal\, x_{i,\text{test}}) \\ \sin(\gamma\, \omega_{P/2}^\intercal\, x_{i,\text{test}}) \end{pmatrix}.$$

The weights $\omega_1, \ldots, \omega_{P/2}$ are shared across all i within a given repetition, see steps 1 and 2 in Algorithm 1. For each pair i, use the single observation $(z_{i,\text{train}}, y_{i,\text{train}})$ to fit a ridge regression and get

$$\hat{\beta}_i = (z_{i,\text{train}}\, z_{i,\text{train}}^\intercal + \lambda\, I)^{-1}\, z_{i,\text{train}}\, y_{i,\text{train}}\,,$$

where $\hat{\beta}_i$ corresponds to $\hat{\beta}_t$ in (5.3). The corresponding prediction of the ridge regression is given by

$$\hat{y}_i = \hat{\beta}_i^\intercal\, z_{i,\text{test}}\,.$$

The performance measures are defined as follows. The prediction error is $(\hat{y}_i - y_{i,\text{test}})^2$, the out-of-sample R^2 is

$$R^2 = 1 - \frac{\sum_{i=1}^N (\hat{y}_i - y_{i,\text{test}})^2}{\sum_{i=1}^N (y_{i,\text{test}} - \overline{y}_{\text{test}})^2}\,, \quad \text{where} \quad \overline{y}_{\text{test}} = \frac{1}{N} \sum_{i=1}^N y_{i,\text{test}}\,.$$

The return contribution from test pair i is

$$Ret_i = \hat{y}_i\, y_{i,\text{test}}\,.$$

The average return is

$$Ret = \frac{1}{N} \sum_{i=1}^N Ret_i\,,$$

and the Sharpe ratio is

$$SR = \frac{Ret}{\sqrt{\frac{1}{N} \sum_{i=1}^N (Ret_i - Ret)^2}}\,.$$

The expectations are taken over the RFF randomness because the data are fixed. In this setting, the expected prediction error can be computed exactly, leading to the following characterization of the expected R^2.

Proposition 4. *Fix* $\{(x_{i,train}, y_{i,train}), (x_{i,test}, y_{i,test})\}_{i=1}^{N}$ *with* $x_{i,train}, x_{i,test} \in \mathbb{R}^d$ *and* $y_{i,train}$, $y_{i,test} \in \mathbb{R}$, *and consider the ridge regression with the RFFs described above. The expected* R^2, *taken over the randomness of the RFFs, is given by*

$$\mathbb{E}[R^2] = 1 - \frac{1}{\sum_{i=1}^{N}(y_{i,test} - \overline{y}_{test})^2} \sum_{i=1}^{N} \mathbb{E}\left[(\hat{y}_i - y_{i,test})^2\right] ,$$

where

$$\begin{aligned}\mathbb{E}\left[(\hat{y}_i - y_{i,test})^2\right] = \frac{y_{i,train}^2}{(\lambda + P/2)^2} \Bigg[& \frac{P}{2} \frac{1 + K(\gamma\,(x_{i,test} - x_{i,train}))^4}{2} \\ & + \frac{P}{2}\left(\frac{P}{2} - 1\right) K(\gamma\,(x_{i,test} - x_{i,train}))^2 \Bigg] \\ & + y_{i,test}^2 - \frac{2\,y_{i,test}\,y_{i,train}\,P}{2\,\lambda + P} K(\gamma\,(x_{i,test} - x_{i,train}))\,. \end{aligned} \tag{5.4}$$

Here, $K(x) = \exp\big(-||x||^2\big)$ *denotes the Gaussian kernel, and* $||\cdot||$ *is the* l_2 *norm of the vector. In the ridgeless case* $\lambda = 0$, *we have*

$$\begin{aligned}\mathbb{E}[R^2] = 1 - \frac{1}{\sum_{i=1}^{N}(y_{i,test} - \overline{y}_{test})^2} \sum_{i=1}^{N} \Bigg[& \frac{y_{i,train}^2}{P} \big\{1 + K(\gamma\,(x_{i,test} - x_{i,train}))^4 \\ & + (P-2)\,K(\gamma\,(x_{i,test} - x_{i,train}))^2\big\} \\ & + y_{i,test}^2 - 2\,y_{i,test}\,y_{i,train}\,K(\gamma\,(x_{i,test} - x_{i,train}))\Bigg] . \end{aligned}$$

In particular, $\mathbb{E}[R^2]$ *is monotonically increasing in* P *when* $\lambda = 0$.

Proposition 4 shows that when $\lambda = 0$, the expected R^2 is increasing in the number of features. When $\lambda > 0$, the trend may not be monotonic and depends on the fixed dataset. Nevertheless, this dependence is entirely determined by the dataset before the experiment. Next, we characterize the expected aggregated return *Ret*.

Proposition 5. *Under the assumption of Proposition 4, we have*

$$\mathbb{E}[Ret] = \frac{1}{N} \sum_{i=1}^{N} \frac{P/2}{\lambda + P/2} \, y_{i,train} \, y_{i,test} \, K\big(\gamma \left(x_{i,test} - x_{i,train}\right)\big) \, . \tag{5.5}$$

In the ridgeless case $\lambda = 0$, we have

$$\mathbb{E}[Ret] = \frac{1}{N} \sum_{i=1}^{N} y_{i,train} \, y_{i,test} \, K\big(\gamma \left(x_{i,test} - x_{i,train}\right)\big) \, .$$

Although in the ridgeless case the expected return is constant in P, the monotonicity of the expected return under ridge regression is determined entirely by the kernel-weighted covariance term

$$\frac{1}{N} \sum_{i=1}^{N} y_{i,\text{train}} \, y_{i,\text{test}} \, K\big(\gamma \left(x_{i,\text{test}} - x_{i,\text{train}}\right)\big) \, . \tag{5.6}$$

If the value of (5.6) is positive, the investment timing strategy constructed with the RFF estimator yields a positive expected return and equation (5.5) is increasing in P; otherwise, if the value of (5.6) is negative, the corresponding expected return is negative and equation (5.5) is decreasing in P.

We now turn to the Sharpe ratio. Unlike R^2 and the mean return, the Sharpe ratio involves a random denominator that depends nonlinearly on the RFFs, which makes the analysis for finite-P difficult. Nevertheless, as $P \to \infty$, the Sharpe ratio converges to a deterministic limit which we state below.

Proposition 6. *Under the assumptions of Proposition 4, when the number of features $P \to \infty$, we have*

$$SR \xrightarrow{\mathbb{P}} \frac{\frac{1}{\sqrt{N}} \sum_{i=1}^{N} y_{i,train} \, y_{i,test} \, K\big(\gamma \left(x_{i,test} - x_{i,train}\right)\big)}{\sqrt{V_N}} \, ,$$

where

$$V_N = \sum_{i=1}^{N} \Bigg(y_{i,train}\, y_{i,test}\, K\big(\gamma\,(x_{i,test} - x_{i,train})\big) - \frac{1}{N} \sum_{j=1}^{N} y_{j,train}\, y_{j,test}\, K\big(\gamma\,(x_{j,test} - x_{j,train})\big) \Bigg)^2 .$$

This asymptotic expression does not depend on P. Therefore, it does not imply any monotonic pattern for finite P and the behavior of the Sharpe ratio for finite P is analytically unclear. However, it is indeed clear that each Ret_i is linear in the product $y_{i,\text{train}}\, y_{i,\text{test}}$. Therefore, flipping the sign of all such products reverses every Ret_i and consequently reverses the aggregated return and the Sharpe ratio for every P. Thus, in the empirical experiment, the direction of the monotonic relationship between the Sharpe ratio and model complexity can be reversed on the same dataset by a simple sign transformation. We formalize this in the next proposition.

Proposition 7. *Consider two experiments that differ only in that all products $y_{i,train}\, y_{i,test}$ are replaced by $-\,y_{i,train}\, y_{i,test}$. Let Ret_P, Var_P and SR_P denote the return, variance and Sharpe ratio when using a number P of RFFs, and let $\widetilde{Ret}_P$, $\widetilde{Var}_P$, $\widetilde{SR}_P$ denote the corresponding quantities under the sign-flipped experiment. Then, for every P,*

$$\widetilde{Ret}_P = -\,Ret_P\,, \qquad \widetilde{Var}_P = Var_P\,, \qquad \widetilde{SR}_P = -\,SR_P\,.$$

Consequently, any increasing (respectively, decreasing) trend in SR_P as a function of P is reversed when all signs of $y_{i,train}\, y_{i,test}$ are flipped.

These results demonstrate that the empirical patterns documented in Section 5.2.1 arise from mechanical interactions between a fixed financial dataset and high-dimensional random features, rather than from predictive learning.

5.3 Numerical simulation

In this section, we present two numerical simulations, followed by an empirical experiment that uses the same dataset as in KMZ but with the sign of the CRSP return flipped at each odd index. We also consider the case where $T > 1$ and repeat the three experiments.

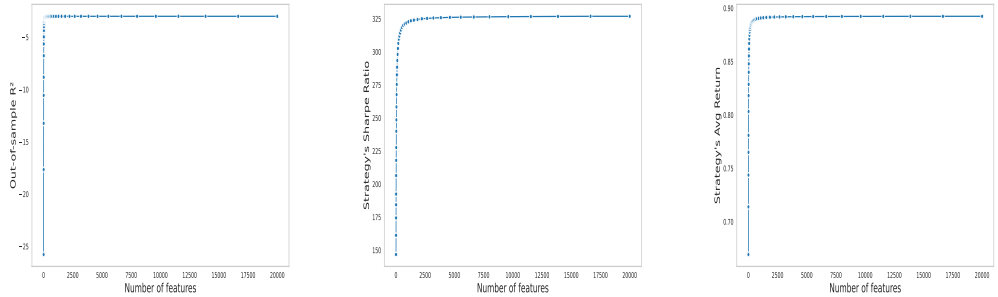


Figure 5.2: Performance metrics for scenario 1 as functions of number of features P, with number of observation $T = 1$, and ridge penalty $\lambda = 1$.

5.3.1 Scenario 1: pure momentum

In this scenario, we consider a pure momentum setting with parameters

$$N = 20\,, \quad d = 10\,, \quad \gamma = 0.1\,, \quad \lambda = 1\,,$$

and covariates generated as $X_t \overset{iid}{\sim} \mathcal{N}(0, I_{10})$. The target variable Y_t is constructed as

$$Y_t = 1 \quad \text{if} \quad t = 2\,i - 1\,, \qquad \text{and} \qquad Y_t = 0.9 \quad \text{if} \quad t = 2\,i\,.$$

We run 500 repetitions of the experiment with X_t and Y_t fixed across repetitions (that is, X_t is generated once and held fixed throughout the RFF experiment). As shown in Figure 5.2, both the average return and the Sharpe ratio of the timing strategy increase monotonically with P, consistent with the theoretical prediction when $Y_t\,Y_{t+1} > 0$.

5.3.2 Scenario 2: pure reversal

Next, consider a pure return-reversal scenario with the same parameter choices as the ones above and the target Y_t alternates in sign:

$$Y_t = 1 \quad \text{if} \quad t = 2\,i - 1\,, \qquad \text{and} \qquad Y_t = -1 \quad \text{if} \quad t = 2\,i\,.$$

Again, we run 500 repetitions with fixed X_t and Y_t. As Figure 5.3 illustrates, both the return and Sharpe ratio of the timing strategy decrease with P, consistent with the theoretical prediction when $Y_t\,Y_{t+1} < 0$.

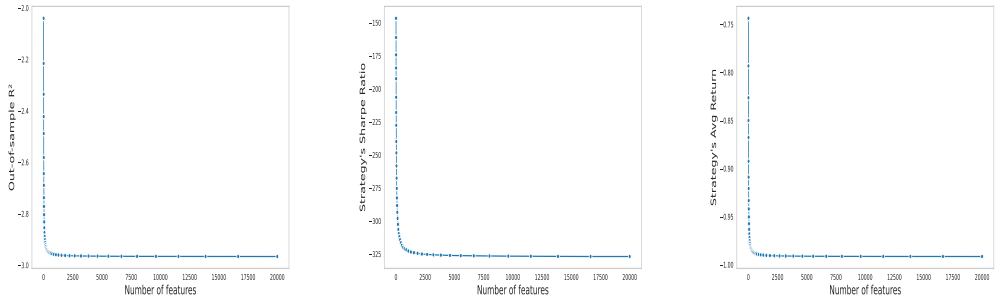


Figure 5.3: Performance metrics for scenario 2 as functions of number of features P, with number of observation $T = 1$, and ridge penalty $\lambda = 1$.

5.3.3 KMZ dataset with sign-flipped CRSP returns

Finally, we revisit the empirical experiment of KMZ. In this modified version, we flip the sign of the CRSP return at each odd index. Let R_t denote the original monthly CRSP return. We construct the modified series $\widetilde{R}_t$ by

$$\widetilde{R}_t = \begin{cases} -R_t\,, & t = 2\,i - 1\,, \\ R_t\,, & t = 2\,i\,. \end{cases}$$

Next, re-run the experiment with $\widetilde{R}_t$ in place of R_t. As Figure 5.4 shows, the trend in the out-of-sample R^2 as a function of the number of features P does not flip. This is consistent with the theory that when the value of ridge penalty λ is small, the relationship between R^2 and P is dominated by the first term in (5.4), which is unaffected by the flip in sign. In contrast, both the return and Sharpe ratio now exhibit a decreasing trend, reversing their behavior in the original KMZ setting, exactly as predicted by the sign reversal of R_t.

5.3.4 Number of observations $T > 1$

Next, we consider the case where the lookback window size $T > 1$ and examine the same three settings as above. For the first two simulation scenarios, the only change is that we increase the sample size to $N = 100$ and for all three scenarios, the lookback window sizes are $T = 3, 6, 12, 36$.

Figures 5.5, 5.6, and 5.7 illustrate that when $T > 1$, the behavior of the timing strategy becomes more complex than in the $T = 1$ case. In particular, the

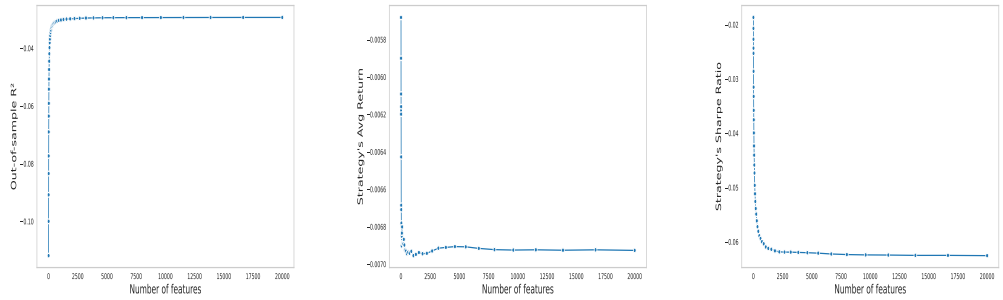


Figure 5.4: Performance of the KMZ experiment with CRSP returns sign-flipped at odd indices, as functions of number of features P, with number of observation $T = 1$, and ridge penalty $\lambda = 1$.

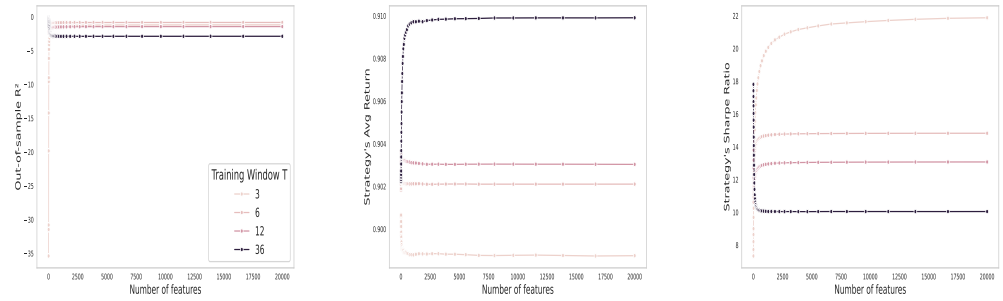


Figure 5.5: Performance metrics for scenario 1 as functions of number of features P, with number of observation $T = 3, 6, 12, 36$, and ridge penalty $\lambda = 1$.

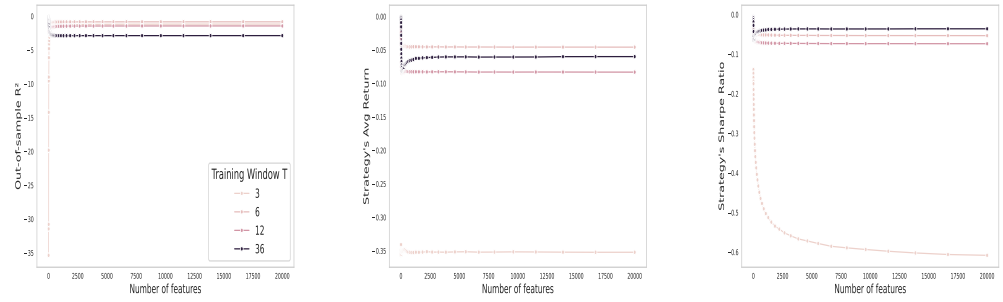


Figure 5.6: Performance metrics for scenario 2 as functions of number of features P, with number of observation $T = 3, 6, 12, 36$, and ridge penalty $\lambda = 1$.

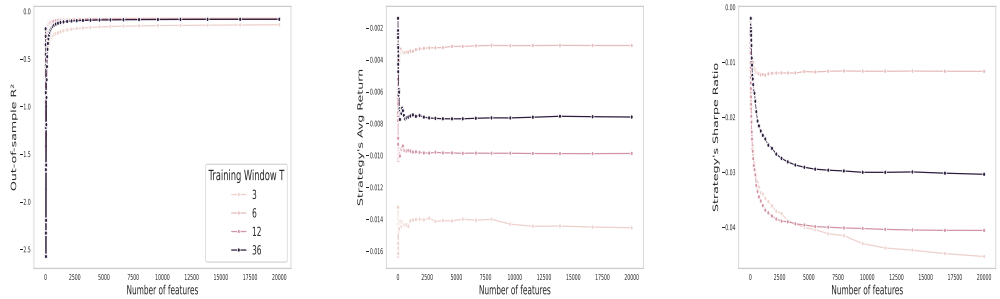


Figure 5.7: Performance of the KMZ experiment with CRSP returns sign-flipped at odd indices, as functions of number of features P, with number of observation $T = 3, 6, 12, 36$, and ridge penalty $\lambda = 1$.

patterns of the return and Sharpe ratio vary substantially across different sizes of the lookback window. For example, the patterns observed for $T = 3$ and $T = 36$ are nearly opposite in the pure momentum and the pure reversal cases. These results indicate that when $T > 1$, the performance of the RFF experiment is not characterized by a simple momentum or reversal structure. Instead, the interaction between the lookback window size and the return dynamics generates more complicated patterns.

In contrast, for the KMZ sign-flip experiment, the behavior remains relatively stable across different values of T. In particular, the decreasing pattern in return and Sharpe ratio persists after the sign flip, suggesting that the reversal of the performance pattern remains robust to changes in the size of the lookback window.

5.4 Conclusion and future work

In this chapter, we showed that the RFF mechanism underlying the results of KMZ continues to appear even in the extreme case of $T = 1$. Empirically, we observe that the increasing trend documented by KMZ persists when the training window consists of a single observation. Theoretically, we derive expressions for the expected return and out-of-sample R^2, which fully characterize their behavior as functions of the number of features P. We further demonstrate that flipping the sign of the underlying returns reverses the trend of Sharpe ratio, so increasing the number of features leads to lower returns and lower Sharpe ratios. Numerical simulations confirm this behavior in a pure reversal setting.

This chapter complements the analysis in Chapters 3 and 4. While those chapters show how imperfect feature quality can limit the benefits of model complexity, the results here show that an apparent improvement in performance with complexity need not necessarily reflect better learning of economic information. Instead, such patterns can also arise mechanically from the construction of high-dimensional features and the resulting estimator. Together, these results highlight that the relationship between model complexity and performance depends both on the quality of the underlying features and on the mechanisms through which complexity is introduced into the model.

An important direction for future research is to extend our analysis to settings where $T > 1$. In this case, deriving closed-form expressions for the expected R^2, return, and Sharpe ratio is more challenging. Our simulation results suggest that when $T > 1$, the behavior of return and Sharpe ratio can vary substantially across different sizes of the lookback window, indicating that the simple patterns observed when $T = 1$ no longer fully characterize the performance of the timing strategy. Nevertheless, it may be possible to establish non-asymptotic bounds for these quantities and to understand how they behave as $P \to \infty$. Such results would provide a more complete picture of the interaction between the number of features, sample size, and performance in RFF-based models.

Bibliography

[1] Yacine Ait-Sahalia, Per A Mykland, and Lan Zhang. How often to sample a continuous-time process in the presence of market microstructure noise. *The Review of Financial Studies*, 18(2):351–416, 2005.

[2] James J Angel, Lawrence E Harris, and Chester S Spatt. Equity trading in the 21st century: An update. *The Quarterly Journal of Finance*, 5(01):1550002, 2015.

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

[4] Vitor Azevedo, Christopher Hoegner, and Mihail Velikov. The expected returns on machine-learning strategies. *Available at SSRN 4702406*, 2023.

[5] Zhidong Bai and Wang Zhou. Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, pages 425–442, 2008.

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

[7] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[8] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2 edition, 1999.

[9] George EP Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

[10] Jonathan Brogaard, Allen Carrion, Thibaut Moyaert, Ryan Riordan,

Andriy Shkilko, and Konstantin Sokolov. High frequency trading and extreme price movements. *Journal of Financial Economics*, 128(2):253–265, 2018.

[11] Jonathan Brogaard and Abalfazl Zareei. Machine learning and the stock market. *Journal of Financial and Quantitative Analysis*, 58(4):1431–1472, 2023.

[12] Eric Budish, Peter Cramton, and John Shim. The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics*, 130(4):1547–1621, 2015.

[13] Daniel Buncic. Simplified: A closer look at the virtue of complexity in return prediction. *Available at SSRN*, 2025.

[14] Donald Lyman Burkholder. Martingale transforms. *The Annals of Mathematical Statistics*, 37(6):1494–1504, 1966.

[15] John Y. Campbell, Andrew W. Lo, and A. Craig MacKinlay. *The Econometrics of Financial Markets*. Princeton University Press, 1997.

[16] Sean Cao, Wei Jiang, Junbo Wang, and Baozhong Yang. From man vs. machine to man+ machine: The art and AI of stock analyses. *Journal of Financial Economics*, 160:103910, 2024.

[17] Álvaro Cartea, Qi Jin, and Yuantao Shi. The limited virtue of complexity in a noisy world. *Available at SSRN*, 2025.

[18] Álvaro Cartea and Yuantao Shi. The mechanical virtue of random Fourier features and model complexity. *Available at SSRN 5930257*, 2025.

[19] Alain P Chaboud, Benjamin Chiquoine, Erik Hjalmarsson, and Clara Vega. Rise of the machines: Algorithmic trading in the foreign exchange market. *The Journal of Finance*, 69(5):2045–2084, 2014.

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

[21] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv preprint arXiv:1511.07289*, 4(5):11, 2015.

[22] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

[23] Hasan Fallahgoul. High-dimensional learning in finance. *arXiv preprint arXiv:2506.03780*, 2025.

[24] Hasan Fallahgoul. Learning from (almost) nothing: An exact theory of projection learning in finance. *Available at SSRN 5577112*, 2025.

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

[26] Guanhao Feng, Jingyu He, Nicholas G Polson, and Jianeng Xu. Deep learning in characteristics-sorted factor models. *Journal of Financial and Quantitative Analysis*, 59(7):3001–3036, 2024.

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

[28] Domenico Giannone, Michele Lenza, and Giorgio E Primiceri. Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437, 2021.

[29] Stefano Giglio, Bryan Kelly, and Dacheng Xiu. Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics*, 14(1):337–368, 2022.

[30] Jeremiah Green, John RM Hand, and X Frank Zhang. The characteristics that provide independent information about average US monthly stock returns. *The Review of Financial Studies*, 30(12):4389–4436, 2017.

[31] Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via

machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

[32] Shihao Gu, Bryan Kelly, and Dacheng Xiu. Autoencoder asset pricing models. *Journal of Econometrics*, 222(1):429–450, 2021.

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

[34] Lars Peter Hansen and Scott F Richard. The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models. *Econometrica*, 55(3):587–613, 1987.

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

[36] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949, 2022.

[37] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[38] JB Heaton, Nicholas G Polson, and Jan Hendrik Witte. Deep learning in finance. *arXiv preprint arXiv:1602.06561*, 2016.

[39] Roy D Henriksson and Robert C Merton. On market timing and investment performance part II: Statistical procedures for evaluating forecasting skills. *Journal of Business*, 54(4):513–534, 1981.

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

[41] Kewei Hou, Chen Xue, and Lu Zhang. Replicating anomalies. *The Review of Financial Studies*, 33(5):2019–2133, 2020.

[42] Jean Jacod, Yingying Li, Per A Mykland, Mark Podolskij, and Mathias

Vetter. Microstructure noise in the continuous case: the pre-averaging approach. *Stochastic Processes and Their Applications*, 119(7):2249–2276, 2009.

[43] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

[44] Jingwen Jiang, Bryan Kelly, and Dacheng Xiu. (Re-)imag(in)ing price trends. *The Journal of Finance*, 78(6):3193–3249, 2023.

[45] Marcin Kacperczyk, Stijn Van Nieuwerburgh, and Laura Veldkamp. Time-varying fund manager skill. *The Journal of Finance*, 69(4):1455–1484, 2014.

[46] Bryan Kelly, Boris Kuznetsov, Semyon Malamud, and Teng Andrea Xu. Large (and deep) factor models. *arXiv preprint arXiv:2402.06635*, 2024.

[47] Bryan Kelly, Semyon Malamud, and Kangying Zhou. The virtue of complexity in return prediction. *The Journal of Finance*, 79(1):459–503, 2024.

[48] Bryan T Kelly and Semyon Malamud. Understanding the virtue of complexity. *Available at SSRN 5346842*, 2025.

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

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

[51] Zhu Li, Weijie J. Su, and Dino Sejdinovic. Benign overfitting and noisy features. *Journal of the American Statistical Association*, 118(544):2876–2888, 2023.

[52] Dennis V Lindley and Adrian FM Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 34(1):1–18, 1972.

[53] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

[54] Benjamin Marrow and Stefan Nagel. Real-time discovery and tracking of return-based anomalies. Technical report, Working Paper, 2024.

[55] Robert C Merton. On market timing and investment performance part I: An equilibrium theory of value for market forecasts. *Journal of Business*, 54(3):363–406, 1981.

[56] Stefan Nagel. Seemingly virtuous complexity in return prediction. *Chicago Booth Research Paper*, (25-10), 2025.

[57] M Hashem Pesaran and Allan Timmermann. Predictability of stock returns: Robustness and economic significance. *The Journal of Finance*, 50(4):1201–1228, 1995.

[58] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 20, 2007.

[59] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in Neural Information Processing Systems*, 21, 2008.

[60] Jack W Silverstein and Zhi Dong Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):175–192, 1995.

[61] Justin A Sirignano. Deep learning for limit order books. *Quantitative Finance*, 19(4):549–570, 2019.

[62] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices., 2012.

[63] Ivo Welch and Amit Goyal. A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508, 2008.

[64] Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

[65] Dacheng Xiu and Zhouyu Shen. Deep autoencoders for nonlinear factor models: Theory and applications. *Available at SSRN*, 2024.

A | Virtue of Complexity with Noisy Features

A.1 Proofs

A.1.1 Proof of Proposition 1.

Throughout the proof, all convergences are in probability unless otherwise stated. In addition, the limits are taken as $T, P \to \infty$ and $P/T \to c$. To prove Proposition 1, we need the following lemmas.

Lemma 3. *Suppose that* $X_t = (X_i)_{i=1}^P$ *with* X_i *independent of* $X_j, \mathbb{E}[X_i] = 0$, $\mathbb{E}[X_i^2] = 1$, *and* $\mathbb{E}[X_i^4] \leq k$, *and* $\hat{\tilde{\Psi}}_{T,P}$ *is such that* $\left\|\hat{\tilde{\Psi}}_{T,P}\right\|_F = o(1)$, *where* $\left\|\hat{\tilde{\Psi}}_{T,P}\right\|_F$ *is the Frobenius norm of* $\hat{\tilde{\Psi}}_{T,P}$. *Then,*

(1) $X_t^\intercal \, \hat{\tilde{\Psi}}_{T,P} \, X_t = \operatorname{tr}\left(\hat{\tilde{\Psi}}_{T,P} \, X_t \, X_t^\intercal\right)$,

(2) $\lim_{P\to\infty} \mathbb{E}\left[\left(X_t^\intercal \, \hat{\tilde{\Psi}}_{T,P} \, X_t - \operatorname{tr}\left(\hat{\tilde{\Psi}}_{T,P}\right)\right)^2\right] = 0.$

In particular, If $\hat{\tilde{\Psi}}_{T,P} = B_P/P$ *where* $\|B_P\| \leq K$, *and* $\|\cdot\|$ *is the Spectral norm operator. We have* $\left\|\hat{\tilde{\Psi}}_{T,P}\right\|_F^2 \leq P\|B_P\|^2/P^2 \leq K/P$ *(see* (A.5) *below), and hence*

$$\lim_{P\to\infty} \mathbb{E}\left[\left(X_t^\intercal \, B_P \, X_t - \operatorname{tr}(B_P)\right)^2\right]/P^2 = 0\,. \tag{A.1}$$

Proof. This Lemma is the same as Lemma IA1 in KMZ. □

The following two lemmas are the same as Lemmas IA4 and IA5 in KMZ.

Lemma 4. *If* A, B *are symmetric positive semi-definite, then*

$$\operatorname{tr}(A\,B) \leq \operatorname{tr}(A)\,||B|| \quad \text{and} \quad \operatorname{tr}\left(A^{1/2}\,B\,A^{1/2}\right) \leq \operatorname{tr}(B)\,||A||\,,$$

where $||A||$ is the Spectral norm of A.

Lemma 5. *Suppose that*

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

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

The proof of the next Lemma is similar to that of Lemma IA3 in KMZ with $\hat{\tilde{\Psi}}_T$ instead of $\hat{\tilde{\Psi}}_T$.

Lemma 6. *We have*

$$P^{-1}\operatorname{tr}\left(Q_P\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\right) - \mathbb{E}\left[P^{-1}\operatorname{tr}\left(Q_P\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\right)\right] \to 0 \quad \text{(A.2)}$$

almost surely as $T, P \to \infty$ and $P/T \to c$ for any sequence of uniformly bounded matrices Q_P, which means $||Q_P|| \leq K$ for all P.

Proof. The proof follows by the same arguments as those in [5]. Let $\hat{\tilde{\Psi}}_{T,t} = \frac{1}{T}\sum_{\tau \neq t} \tilde{F}_\tau \tilde{F}_\tau^\intercal$. By the Sherman–Morrison formula,

$$\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1} = \left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} - \frac{1}{T}\frac{\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t\,\tilde{F}_t^\intercal\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}}{1 + T^{-1}\tilde{F}_t^\intercal\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t}\,.$$

Let $\mathbb{E}_\tau$ denote the conditional expectation given $\tilde{F}_{\tau+1}, \cdots, \tilde{F}_t$. Let also

$$q_T(z) = \frac{1}{P}\operatorname{tr}\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1} Q_P\,.$$

With this notation, we have

$$(\mathbb{E}_{t-1} - \mathbb{E}_t)\left[\frac{1}{P}\operatorname{tr}\left(z\,I + \tilde{\Psi}_{T,t}\right)^{-1} Q_P\right] = 0$$

because $\hat{\tilde{\Psi}}_{T,t}$ is independent of $\tilde{F}_t$, and therefore

$$\begin{aligned}
&\mathbb{E}\left[q_T(z)\right] - q_T(z) = \mathbb{E}_0\left[q_T(z)\right] - \mathbb{E}_T\left[q_T(z)\right] = \sum_{t=1}^{T}\left(\mathbb{E}_{t-1}\left[q_T(z)\right] - \mathbb{E}_t\left[q_T(z)\right]\right) \\
&= \sum_{t=1}^{T}\left(\mathbb{E}_{t-1} - \mathbb{E}_t\right)\left[q_T(z)\right] \\
&= \sum_{t=1}^{T}\left(\mathbb{E}_{t-1} - \mathbb{E}_t\right)\left[q_T(z) - \frac{1}{P}\,\mathrm{tr}\left(z\,I + \tilde{\Psi}_{T,t}\right)^{-1} Q_P\right] \\
&= \frac{1}{P}\sum_{t=1}^{T}\left(\mathbb{E}_{t-1} - \mathbb{E}_t\right)\left[\mathrm{tr}\left(z\,I + \hat{\tilde{\Psi}}_{T}\right)^{-1} Q_P - \mathrm{tr}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} Q_P\right] \\
&= -\frac{1}{P}\sum_{\tau=1}^{T}\left(\mathbb{E}_{t-1} - \mathbb{E}_t\right)\left[l_t\right]\,,
\end{aligned}$$

where we use (A.13) and defined

$$l_t = \mathrm{tr}\left(\frac{1}{T}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t\left(I + \frac{1}{T}\,\tilde{F}_t^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t\right)^{-1}\tilde{F}_t^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} Q_P\right). \tag{A.3}$$

Next, use Lemma 4 to write

$$\begin{aligned}
\|l_t\| &\leq \|Q_P\|\ \mathrm{tr}\left(\frac{1}{T}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t\left(I + \frac{1}{T}\tilde{F}_t^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t\right)^{-1}\tilde{F}_t^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\right) \\
&\leq K\ \mathrm{tr}\left(\frac{1}{T}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1/2}\tilde{F}_t\left(I + \frac{1}{T}\,\tilde{F}_t^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t\right)^{-1}\tilde{F}_t^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1/2}\right) \\
&= K\ \mathrm{tr}\left(B\left(z\,I + B\right)^{-1}\right) \leq K\,,
\end{aligned}$$

where

$$B = \frac{1}{T}\,\tilde{F}_t^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t \in \mathbb{R}\,.$$

Thus,

$$\left(\mathbb{E}_{t-1} - \mathbb{E}_t\right)\left[\mathrm{tr}\left(z\,I + \hat{\tilde{\Psi}}_{T}\right)^{-1} Q_P\right] = \left(\mathbb{E}_{t-1} - \mathbb{E}_t\right)\left[l_t\right]$$

forms a bounded martingale difference sequence. Apply the Burkholder–

Davis–Gundy inequality (see, for example, [14]) to obtain

$$\mathbb{E}\left[|q_T(z) - \mathbb{E}\left[q_T(z)\right]|^q\right] \leq K_q\, P^{-q}\, \mathbb{E}\left(\sum_{t=1}^{T} \left|\left(\mathbb{E}_{t-1} - \mathbb{E}_t\right)\left[l_t\right]\right|^2\right)^{q/2}$$
$$\leq K_q\, (2\,K)^q\, P^{-q/2} \left(\frac{P}{T}\right)^{-q/2} . \tag{A.4}$$

Almost sure convergence follows from Lemma 5. □

Let $\tilde{\Psi}_1 = (\Psi^{1/2}, \Phi^{1/2})$ and $Z_t = (X_t^\intercal, Y_t^\intercal)^\intercal$, then $\tilde{F}_t = \tilde{\Psi}_1 Z_t$ and $\tilde{\Psi}_1 \tilde{\Psi}_1^\intercal = \Psi + \Phi = \tilde{\Psi}$, and with this notation we have the following lemma.

Lemma 7. *We have* $\left\| \tilde{\Psi}_1^\intercal \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \tilde{\Psi}_1 \right\|_F^2 / P^2 \to 0$ *as* $P \to \infty$.

Proof. For any matrix A we have $||A||_F^2 = \mathrm{tr}(A\, A^\intercal)$, thus

$$\left\| \tilde{\Psi}_1^\intercal \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \tilde{\Psi}_1 \right\|_F^2 = \mathrm{tr}\left(\tilde{\Psi}_1^\intercal \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \tilde{\Psi}_1 \tilde{\Psi}_1^\intercal \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \tilde{\Psi}_1 \right)$$
$$= \mathrm{tr}\left(\left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \tilde{\Psi} \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \tilde{\Psi} \right)$$
$$= \mathrm{tr}\left(\left(\left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \tilde{\Psi} \right)^2 \right) .$$

For any symmetric matrix A, B, simple calculations show that

$$\mathrm{tr}\left((A\,B)^2\right) = \mathrm{tr}\left((B\,A)^2\right) = \frac{1}{4}\left(\mathrm{tr}\left((A\,B + B\,A)^2\right) + \mathrm{tr}\left((A\,B - B\,A)^2\right)\right) .$$

The matrix $A\,B + B\,A$ is symmetric so we can write $A\,B + B\,A = Q^\intercal\, \Lambda\, Q$ as its eigenvalue decomposition, where Q is an orthogonal matrix and Λ is a diagonal matrix. Therefore,

$$\mathrm{tr}\left((A\,B + B\,A)^2\right) = \mathrm{tr}\left(\Lambda^2\right) \leq P\, \lambda_{\max}(\Lambda)^2 , \tag{A.5}$$

where $\lambda_{\max}(\Lambda)$ is the largest eigenvalue of Λ, which is also the largest eigen-

value of $A\,B + B\,A$. Basic matrix norm inequality shows that

$$\lambda_{\max}(\Lambda) = \lambda_{\max}(A\,B + B\,A) \leq \lambda_{\max}\,(A\,B) + \lambda_{\max}(B\,A) \leq 2\,\lambda_{\max}\,(A)\,\lambda_{\max}(B)\,.$$

Therefore,

$$\operatorname{tr}\left(\left(\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\,\tilde{\Psi}\right)^2\right) \leq P\,\lambda_{\max}\left(\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\right)\lambda_{\max}(\tilde{\Psi}) \leq K\,P\,z^{-1}\,,$$

and as $P \to \infty$ we have

$$\left\|\tilde{\Psi}_1^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}_1\right\|_F^2 / P^2 \leq K z^{-1}/P \to 0\,.$$

□

Now we are ready to prove Proposition 1.

Proof. Proof of Proposition 1. Let

$$\hat{\tilde{\Psi}}_{T,t} = \frac{1}{T}\sum_{\tau \neq t} \tilde{F}_\tau\,\tilde{F}_\tau^{\intercal}\,.$$

Then, by the Sherman–Morrison formula,

$$\begin{aligned}\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\,\tilde{F}_t &= \left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\,\tilde{F}_t - \frac{1}{T}\,\frac{\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\,\tilde{F}_t\,\tilde{F}_t^{\intercal}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\,\tilde{F}_t}{1 + (T)^{-1}\,\tilde{F}_t^{\intercal}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\,\tilde{F}_t} \\ &= \left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\,\tilde{F}_t\,\frac{1}{1 + (T)^{-1}\,\tilde{F}_t^{\intercal}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\,\tilde{F}_t}\,.\end{aligned} \tag{A.6}$$

By Lemma 3 and Lemma 7 we have

$$P^{-1}\,\tilde{F}_t^{\intercal}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\,\tilde{F}_t = Z_t^{\intercal}\,\tilde{\Psi}_1^{\intercal}\,\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\,\tilde{\Psi}_1\,Z_t \to P^{-1}\,\operatorname{tr}\left(\tilde{\Psi}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\right)\,.$$

Also, by the assumption that $\tilde{\Psi}$ is uniformly bounded in Spectral norm, use Lemma 6 to write

$$P^{-1}\operatorname{tr}\left(\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\right)-\mathbb{E}\left[P^{-1}\operatorname{tr}\left(\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\right)\right]\to 0\,. \quad \text{(A.7)}$$

Lemma 2 implies that

$$P^{-1}\operatorname{tr}\mathbb{E}\left[\left(z\,I+\hat{\tilde{\Psi}}_{T}\right)^{-1}\right]\to m(-z;c)\,,$$

so we have

$$\begin{aligned}
1 &= P^{-1}\operatorname{tr}\left(\mathbb{E}\left[\left(z\,I+\hat{\tilde{\Psi}}_{T}\right)^{-1}\left(z\,I+\hat{\tilde{\Psi}}_{T}\right)\right]\right)\\
&= P^{-1}\,\operatorname{tr}\left(\mathbb{E}\left[\left(z\,I+\hat{\tilde{\Psi}}_{T}\right)^{-1}\right]z\right)+P^{-1}\,\operatorname{tr}\left(\mathbb{E}\left[\left(z\,I+\hat{\tilde{\Psi}}_{T}\right)^{-1}\hat{\tilde{\Psi}}_{T}\right]\right)\\
&= z\,m(-z,c)+P^{-1}\operatorname{tr}\left(\mathbb{E}\left[\left(z\,I+\hat{\tilde{\Psi}}_{T}\right)^{-1}\frac{1}{T}\sum_{t}\tilde{F}_t\tilde{F}_t^{\intercal}\right]\right)\\
&= z\,m(-z,c)+P^{-1}\operatorname{tr}\left(\mathbb{E}\left[\left(z\,I+\hat{\tilde{\Psi}}_{T}\right)^{-1}\tilde{F}_t\tilde{F}_t^{\intercal}\right]\right)\\
&= z\,m(-z,c)+P^{-1}\operatorname{tr}\left(\mathbb{E}\left[\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t\frac{1}{1+T^{-1}\,\tilde{F}_t^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t}\,\tilde{F}_t^{\intercal}\right]\right)\\
&= z\,m(-z,c)+\mathbb{E}\left[\frac{P^{-1}\tilde{F}_t^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t}{1+T^{-1}\,\tilde{F}_t^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t}\right]\,. \qquad \text{(A.8)}
\end{aligned}$$

Now, $\mathbb{E}\left[T^{-1}\,\operatorname{tr}\left(\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\right)\right]\;\leq\;||\tilde{\Psi}||\,(c\,z)^{-1}$ and hence it is uniformly bounded. Let us pick a subsequence of T converging to infinity and such that

$$\mathbb{E}\left[T^{-1}\operatorname{tr}\left(\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\right)\right]\to q$$

for some $q > 0$. By (A.8) and (A.7),

$$\frac{P^{-1}\, F_t^\intercal \left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} F_t}{1 + T^{-1}\, F_t^\intercal \left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} F_t} \to \frac{c^{-1}\, q}{1+q}$$

in probability and this sequence is uniformly bounded. Hence,

$$\mathbb{E}\left[\frac{P^{-1}\, F_t^\intercal \left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} F_t}{1 + (T)^{-1}\, F_t^\intercal \left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} F_t}\right] \to \frac{c^{-1}\, q}{1+q}$$

and we get

$$1 - z\, m(-z, c) = \frac{c^{-1}\, q}{1+q}\,.$$

Thus, the limit of $\xi(z;c) = \mathbb{E}\left[T^{-1}\ \mathrm{tr}\left(\Psi\left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\right)\right]$ as $T, P \to \infty$ and $P/T \to c$ is independent of the subsequence of T and satisfies the required equation. □

A.1.2 Proof of Theorem 1.

Throughout the proof, all convergences are in probability unless otherwise stated. In addition, the limits are taken as $T, P \to \infty$ and $P/T \to c$. Before we go into the proof of Theorem 1, we note that if the matrices A_P, Q_P are random matrices, Lemmas 1, 3 and 6 cannot be used directly because those random matrices are not uniformly bounded. Therefore, we first prove some lemmas similar to Lemma 3 for random matrices. We first introduce a theorem regarding the non-asymptotic bounds for the Spectral norm of random matrix from [62].

Theorem 2. *(Theorem 5.44 in [62]). Let A be an $N \times n$ matrix whose rows A_i are independent random vectors in $\mathbb{R}^n$ with the common second moment matrix $\Sigma = \mathbb{E}[A_i \otimes A_i]$. Let m be a number such that $\|A_i\|_2 \le \sqrt{m}$ almost surely for all i. Then, for every $s \ge 0$, the inequality*

$$\left\| \frac{1}{N}\, A^\intercal\, A - \Sigma \right\| \le \max\left(||\Sigma||^{1/2}\, \delta,\, \delta^2\right), \qquad \textit{where} \qquad \delta = s\sqrt{\frac{m}{N}} \tag{A.9}$$

holds with probability at least $1 - n \cdot \exp\left(-c_1\, s^2\right)$. Here, $c_1 > 0$ is an absolute

constant. In particular, (A.9) *yields*

$$||A|| \leq ||\Sigma||^{1/2} \sqrt{N} + s\sqrt{m}\,.$$

The next lemma shows that the Spectral norm of $\hat{\tilde{\Psi}}_T/P$ converges to 0 in probability.

Lemma 8. *Under Assumptions 1 and 3, we have*

$$\mathbb{P}\left(\lambda_{\max}\left(\hat{\tilde{\Psi}}_T/P\right) > \epsilon\right) \to 0$$

as $P \to \infty$ for all $\epsilon > 0$. In other words, the Spectral norm of $\hat{\tilde{\Psi}}_T/P$ converges to 0 in probability.

Proof. As $\tilde{F}_t = \tilde{\Psi}_1\, Z_t$ and $\hat{\tilde{\Psi}}_T = \frac{1}{T}\sum_t \tilde{F}_t\, \tilde{F}_t^\intercal = \frac{1}{T}\,\tilde{\Psi}_1 \sum_t Z_t\, Z_t^\intercal\, \tilde{\Psi}_1^\intercal$. It suffices to prove the results for $\frac{1}{PT}\sum_t Z_t Z_t^\intercal$. Because $\mathbb{E}[Z_{t,i}^4] < K$ and $Z_{i,t}$ and $Z_{j,t}$ are independent, let $\hat{Z}_t = \sum_i Z_{t,i}^2$, we have

$$\mathbb{E}\left[\hat{Z}_t\right] = P\,, \quad \text{Var}\left[\hat{Z}_t\right] = \mathbb{E}\left[\sum_i \left(Z_{i,t}^2 - 1\right)\right]^2 \leq KP\,.$$

Therefore, let $m_1 = \sqrt{PT}\,\log P$, we can use other rates but we use this for simplicity, we have,

$$\mathbb{P}\left(\exists t, \left|\hat{Z}_t - P\right| > m_1\right) \leq T\,\mathbb{P}\left(\left|\hat{Z}_t - P\right| > m_1\right) \leq \frac{K\,T\,P}{m_1^2} \leq \frac{K}{(\log P)^2}\,.$$

Next, use Theorem 2 with $N = T$, $n = P$, $\Sigma = I$, $m = P + m_1$ and $A_i = Z_t$, so we have with probability at least $1 - P \cdot \exp\left(-c_1\, s^2\right)$ that

$$\left\|\frac{1}{T}\sum_t Z_t\, Z_t^\intercal - I\right\| \leq \max\left(\delta,\, \delta^2\right)\,, \quad \text{where} \quad \delta = s\sqrt{\frac{m}{T}}\,,$$

and hence

$$\left\|\frac{1}{PT}\sum_t Z_t Z_t^\intercal\right\| \leq \max\left(\delta + 1, \delta^2 + 1\right)/P\,, \quad \text{where} \quad \delta = s\sqrt{\frac{m}{T}}\,.$$

Let $s = \sqrt{\frac{1}{c_1} \log \frac{P}{\delta_1}}$, then, with probability at least $1 - \delta_1$, we have that

$$\left\| \frac{1}{PT} \sum_t Z_t \, Z_t^\intercal \right\| \leq \frac{C \, (\log P)^3}{P} \,,$$

where C depends only on δ_1, c_1, G and $c = P/T$. Therefore, when $C \, (\log P)^3 / P < \epsilon$, we have

$$\begin{aligned} \mathbb{P}\left(\lambda_{\max}\left(\hat{\tilde{\Psi}}_T / P \right) > \epsilon \right) &\leq \mathbb{P}\left(\exists i, t, |Z_{t,i}| > \sqrt{m} \right) + \mathbb{P}\left(\left\| \frac{1}{PT} \sum_t Z_t \, Z_t^\intercal \right\| > \frac{C \, (\log P)^3}{P} \right) \\ &\leq \frac{K}{(\log P)^2} + \delta_1 \to \delta_1 \,. \end{aligned} \tag{A.10}$$

Because (A.10) holds for all δ_1, we have

$$\mathbb{P}\left(\lambda_{\max}\left(\hat{\tilde{\Psi}}_T / P \right) > \epsilon \right) \to 0 \,, \quad \text{as} \quad \delta_1 \to 0 \,.$$

□

Now we prove the next lemma, which is similar to Lemma 3.

Lemma 9. *Suppose that $U_t = (U_i)_{i=1}^P$ with U_i independent of U_j, $\mathbb{E}[U_i] = 0$, $\mathbb{E}[U_i^2] = 1$, $\mathbb{E}[U_i^4] \leq k$. Suppose further that U_t is independent from $\hat{\tilde{\Psi}}_T$, then we have*

$$\frac{1}{P} \, U_t^\intercal \, \hat{\tilde{\Psi}}_T \, U_t \to \frac{1}{P} \, \mathrm{tr}\left(\hat{\tilde{\Psi}}_T \right)$$

in probability as $T, P \to \infty$ and $P/T \to c$.

Proof. As shown in the proof of Lemma 8, we have

$$\mathbb{P}\left(\left\| \hat{\tilde{\Psi}}_T \right\| \geq C \, (\log P)^3 \right) \leq \frac{K}{(\log P)^2} + \delta_1$$

for any $\delta_1 > 0$. Now we restrict ourselves on the event $\mathcal{A} = \left\{ \left\| \hat{\tilde{\Psi}}_T \right\| \leq C \, (\log P)^3 \right\}$, under $\mathcal{A}$, we have

$$\left\| \hat{\tilde{\Psi}}_T \right\|_F^2 / P^2 \leq C \, (\log P)^3 / P \to 0 \,.$$

Therefore, because U_t is independent from $\hat{\tilde{\Psi}}_T$, define $V_t = U_t^\intercal \, \hat{\tilde{\Psi}}_T \, U_t / P$ and $\hat{\tilde{\Psi}}_{T,P} = \hat{\tilde{\Psi}}_T / P$. We have

$$\mathbb{E}\left[V_t \mid \hat{\tilde{\Psi}}_T\right] = \operatorname{tr}\left(\hat{\tilde{\Psi}}_{T,P}\right)$$

and hence

$$\mathbb{E}\left[\left(V_t - \operatorname{tr}\left(\hat{\tilde{\Psi}}_{T,P}\right)\right)^2 \mid \hat{\tilde{\Psi}}_T\right] = \mathbb{E}\left[V_t^2 \mid \hat{\tilde{\Psi}}_T\right] - \mathbb{E}\left[V_t \mid \hat{\tilde{\Psi}}_T\right]^2 .$$

It thus suffices to prove that

$$\mathbb{E}\left[V_t^2 \mid \hat{\tilde{\Psi}}_T\right] - \left(\operatorname{tr}\left(\hat{\tilde{\Psi}}_{T,P}\right)\right)^2 \to 0 \,.$$

Now

$$V_t = \sum_{i,j} U_i \, U_j \, \hat{\tilde{\Psi}}_{T,P,i,j}$$

and therefore

$$V_t^2 = \sum_{i_1,j_1,i_2,j_2} U_{i_1} \, U_{j_1} \, \hat{\tilde{\Psi}}_{T,P,i_1,j_1} \, \hat{\tilde{\Psi}}_{T,P,i_2,j_2} \, U_{i_2} U_{j_2} \,.$$

Next, we compute the expectation

$$\begin{aligned}
\mathbb{E}\left[V_t^2 \mid \hat{\tilde{\Psi}}_T\right] &= \sum_{i_1,j_1,i_2,j_2} \hat{\tilde{\Psi}}_{T,P,i_1,j_1} \, \hat{\tilde{\Psi}}_{T,P,i_2,j_2} \, \mathbb{E}\left[U_{i_1} \, U_{j_1} \, U_{i_2} \, U_{j_2}\right] \\
&= \left(\sum_i \hat{\tilde{\Psi}}^2_{T,P,i,i} \, \mathbb{E}\left[U_i^4\right] + \sum_{i,j} \left(\hat{\tilde{\Psi}}^2_{T,P,i,j} + \hat{\tilde{\Psi}}_{T,P,i,i} \, \hat{\tilde{\Psi}}_{T,P,j,j}\right) \mathbb{E}\left[U_i^2 \, U_j^2\right]\right. \\
&= \left(\sum_i k \, \hat{\tilde{\Psi}}^2_{T,P,i,i} + \sum_{i,j} \hat{\tilde{\Psi}}^2_{T,P,i,j} + 2 \, \hat{\tilde{\Psi}}_{T,P,i,i} \, \hat{\tilde{\Psi}}_{T,P,j,j}\right) \\
&= \left((k-1) \sum_i \hat{\tilde{\Psi}}^2_{T,P,i,i} + \sum_{i,j} \hat{\tilde{\Psi}}^2_{T,P,i,j} + \operatorname{tr}(\hat{\tilde{\Psi}}_{T,P})^2\right) ,
\end{aligned}$$

thus

$$\left|\mathbb{E}\left[V_t^2 \mid \hat{\tilde{\Psi}}_T\right] - \operatorname{tr}(\hat{\tilde{\Psi}}_{T,P})^2\right| \leq k \left\|\hat{\tilde{\Psi}}_{T,P}\right\|_F^2 \to 0 \,.$$

Therefore, under $\mathcal{A}$,

$$\frac{1}{P} \, U_t^\intercal \, \hat{\tilde{\Psi}}_T \, U_t \to \frac{1}{P} \, \operatorname{tr}\left(\hat{\tilde{\Psi}}_T\right)$$

in L_2 and thus in probability. As $\mathbb{P}(\mathcal{A}) \leq \delta_1$ for all $\delta_1 > 0$ as $P \to \infty$, we have

$$\frac{1}{P} U_t^\intercal \hat{\tilde{\Psi}}_T U_t \to \frac{1}{P} \operatorname{tr}\left(\hat{\tilde{\Psi}}_T\right)$$

in probability as $T, P \to \infty$. □

Although Lemma 9 only considers $\hat{\tilde{\Psi}}_T$, the lemma still applies as long as the vector U_t and the matrix $\hat{\tilde{\Psi}}_T$ are independent and $\hat{\tilde{\Psi}}_T$ satisfies the norm conditions. Therefore, in the following context, we use Lemma 9 for matrices different from $\hat{\tilde{\Psi}}_T$. We also need a lemma similar to Lemma 6, with $\left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}$ replaced by $\left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-2}$.

Lemma 10. *We have*

$$P^{-1} \operatorname{tr}\left(Q_P \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-2}\right) - \mathbb{E}\left[P^{-1} \operatorname{tr}\left(Q_P \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-2}\right)\right] \to 0 \quad \text{(A.11)}$$

almost surely as $T, P \to \infty$ and $P/T \to c$ for any sequence of uniformly bounded matrices Q_P, which means $||Q_P|| \leq K$ for all P.

Proof. The proof follows by the same arguments as those in [5]. Let $\hat{\tilde{\Psi}}_{T,t} = \frac{1}{T} \sum_{\tau \neq t} \tilde{F}_\tau \tilde{F}_\tau^\intercal$. Let

$$g_t = \frac{1}{T} \left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{F}_t \tilde{F}_t^\intercal \left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \frac{1}{1 + T^{-1} \tilde{F}_t^\intercal \left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{F}_t}. \quad \text{(A.12)}$$

By the Sherman–Morrison formula (A.13),

$$\begin{aligned}\left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-2} = &\left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-2} \\ &- \left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} g_t - g_t \left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} + g_t^2 . \quad \text{(A.13)}\end{aligned}$$

Let $\mathbb{E}_\tau$ denote the conditional expectation given $\tilde{F}_{\tau+1}, \cdots, \tilde{F}_t$. Let also

$$q_T(z) = \frac{1}{P}\,\mathrm{tr}\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-2} Q_P\,.$$

With this notation, because $\hat{\tilde{\Psi}}_{T,t}$ is independent of $\tilde{F}_t$, we have

$$(\mathbb{E}_{t-1} - \mathbb{E}_t)\left[\frac{1}{P}\,\mathrm{tr}\left(z\,I + \tilde{\Psi}_{T,t}\right)^{-2} Q_P\right] = 0\,,$$

and therefore

$$\begin{aligned}
\mathbb{E}\left[q_T(z)\right] - q_T(z) &= \mathbb{E}_0\left[q_T(z)\right] - \mathbb{E}_T\left[q_T(z)\right] = \sum_{t=1}^{T}\left(\mathbb{E}_{t-1}\left[q_T(z)\right] - \mathbb{E}_t\left[q_T(z)\right]\right) \\
&= \sum_{t=1}^{T}\left(\mathbb{E}_{t-1} - \mathbb{E}_t\right)\left[q_T(z)\right] \\
&= \sum_{t=1}^{T}\left(\mathbb{E}_{t-1} - \mathbb{E}_t\right)\left[q_T(z) - \frac{1}{P}\,\mathrm{tr}\left(z\,I + \tilde{\Psi}_{T,t}\right)^{-2} Q_P\right] \\
&= \frac{1}{P}\sum_{t=1}^{T}\left(\mathbb{E}_{t-1} - \mathbb{E}_t\right)\left[\mathrm{tr}\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-2} Q_P - \mathrm{tr}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-2} Q_P\right] \\
&= -\frac{1}{P}\sum_{\tau=1}^{T}\left(\mathbb{E}_{t-1} - \mathbb{E}_t\right)\left[\hat{l}_t\right]\,,
\end{aligned}$$

where we used (A.13) and defined

$$\hat{l}_t = \mathrm{tr}\left(-\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} g_t Q_P - g_t\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} Q_P + g_t^2 Q_P\right)\,. \quad \text{(A.14)}$$

Next, use Lemma 4 and the calculation of l in the proof of Lemma 6 to write

$$\begin{aligned}
\left\|\hat{l}_t\right\| &\leq 2\,z^{-1}\,\|Q_P\|\,\mathrm{tr}\,(g_t) + \|Q_P\|\,\mathrm{tr}(g_t^2) \\
&\leq 2\,z^{-1}\,K + z^{-2}\,K\,\mathrm{tr}(B^2\,(1+B)^{-2}) \\
&\leq (2\,z^{-1}\,K + z^{-2})\,K\,,
\end{aligned}$$

where

$$B = \frac{1}{T}\,\tilde{F}_t^\intercal\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\,\tilde{F}_t \in \mathbb{R}^+\,.$$

Thus,

$$(\mathbb{E}_{t-1} - \mathbb{E}_t) \left[\operatorname{tr} \left(z\, I + \hat{\tilde{\Psi}}_T \right)^{-1} Q_P \right] = (\mathbb{E}_{t-1} - \mathbb{E}_t) \left[\hat{l}_t \right]$$

forms a bounded martingale difference sequence. Applying the Burkholder–Davis–Gundy inequality (see, for example, [14]), we get

$$\begin{aligned} \mathbb{E}\left[|q_T(z) - \mathbb{E}\left[q_T(z)\right]|^q \right] &\leq K_q\, P^{-q}\, \mathbb{E}\left(\sum_{t=1}^{T} \left| (\mathbb{E}_{t-1} - \mathbb{E}_t)\left[l_t\right] \right|^2 \right)^{q/2} \\ &\leq K_q \left((2\, z^{-1} K + z^{-2})\, K \right)^q P^{-q/2}\, (P/T)^{-q/2}\,. \end{aligned} \tag{A.15}$$

Almost sure convergence follows from Lemma 5. □

Now we prove Theorem 1. Note that although not explicitly stated, every expectation we take is conditional expectation on β.

Proof. Proof of Theorem 1. First we calculate $\lim \mathbb{E}\left[\hat{\pi}_t(z)\, R_{t+1} \mid \hat{\beta} \right]$. Note that

$$\mathbb{E}\left[\hat{\pi}_t(z)\, R_{t+1} \mid \hat{\beta} \right] = \mathbb{E}\left[\hat{\beta}^\intercal\, \tilde{F}_t \left(F_t^\intercal\, \beta + \epsilon_{t+1} \right) \mid \hat{\beta} \right] = \hat{\beta}^\intercal\, \Psi\, \beta\,,$$

and write

$$\hat{\Psi}_{\tilde{F},F} = \frac{1}{T} \sum_t \tilde{F}_t\, F_t^\intercal \quad \text{and} \quad q_T = \frac{1}{T} \sum_t \tilde{F}_t\, \epsilon_{t+1}$$

because $\beta, \hat{\beta}$ are independent of $F_t, \tilde{F}_t$ and ϵ_{t+1}. Next, we know that

$$\frac{1}{T} \sum_t \tilde{F}_t\, R_{t+1} = \frac{1}{T} \sum_t \tilde{F}_t\, \left(F_t^\intercal\, \beta + \epsilon_{t+1} \right) = \hat{\Psi}_{\tilde{F},F}\, \beta + q_T\,,$$

and

$$\hat{\beta} = \left(z\, I + \hat{\tilde{\Psi}}_T \right)^{-1} \left(\hat{\Psi}_{\tilde{F},F} \beta + q_T \right)\,, \tag{A.16}$$

$$\hat{\beta}^\intercal\, \Psi\, \beta = \beta^\intercal\, \Psi \left(z\, I + \hat{\tilde{\Psi}}_T \right)^{-1} \hat{\Psi}_{\tilde{F},F}\, \beta + \beta^\intercal\, \Psi \left(z\, I + \hat{\tilde{\Psi}}_T \right)^{-1} q_T\,. \tag{A.17}$$

Because $\mathbb{E}[q_T] = 0$ and $\mathbb{E}[q_T^\intercal\, q_T] = \frac{1}{T^2} \operatorname{tr}(\tilde{\Psi}) \to 0$ we have that q_T converges to 0 in L_2 in probability. Also, as ϵ_{t+1} is independent of $\tilde{F}_t$, we have

$$\mathbb{E}\left[q_T\, q_T^\intercal \mid \{\tilde{F}_t\} \right] = \frac{1}{T}\, \hat{\tilde{\Psi}}_T\,. \tag{A.18}$$

Below, $a \approx b$ denotes that $a - b \to 0$ in L_2. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\beta^\intercal \Psi \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} q_T\, q_T^\intercal \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \Psi\beta\right] \\
&= \mathbb{E}\left[\beta^\intercal \Psi \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \mathbb{E}\left[q_T\, q_T^\intercal \mid \{\tilde{F}_t\}\right] \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \Psi\beta\right] \\
&= \mathbb{E}\left[\beta^\intercal \Psi \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \frac{1}{T}\, \hat{\tilde{\Psi}}_T \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \Psi\, \beta\right] \\
&\approx b_*\, P^{-1}\, \mathrm{tr}\, \mathbb{E}\left[\Psi \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \frac{1}{T}\, \hat{\tilde{\Psi}}_T \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \Psi\right] \\
&\leq z^{-1}\, b_*\, ||\tilde{\Psi}||^2\, P/(P\,T) \to 0
\end{aligned} \tag{A.19}$$

where the third equation follows from Lemma 1 and

$$\left\|\Psi \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} T^{-1}\, \hat{\tilde{\Psi}}_T \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \Psi\right\| \leq (z\,T)^{-1}\, ||\Psi||^2$$

is uniformly bounded; the last inequality follows from Lemma 4. We know from (A.19) that

$$\beta^\intercal\, \Psi \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} q_T \to 0$$

in L_2 thus in probability. Therefore, we only need to calculate the first term in (A.17).

Write $\tilde{F} = (\tilde{F}_t) \in \mathbb{R}^{P\times T}$ and $F = (F_t) \in \mathbb{R}^{P\times T}$, then $\tilde{\Psi}_{\tilde{F},F} = \frac{1}{T}\, \tilde{F} F^\intercal$ and therefore

$$\left\|\tilde{\Psi}_{\tilde{F},F}\right\| \leq \frac{1}{T} \left\|\tilde{F}\right\| ||F||\,.$$

Using Theorem 2 and Lemma 8 we know that

$$\mathbb{P}\left(\left\|\tilde{\Psi}_{\tilde{F},F}\right\| \geq C\,(\log P)^3\right) \leq \frac{K}{(\log P)^2} + \delta_1 \tag{A.20}$$

for any $\delta_1 > 0$. Using the same argument in the proof of Lemma 9 we have that

$$\beta^\intercal\, \Psi \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \hat{\Psi}_{\tilde{F},F}\, \beta \to \mathrm{tr}\left[b_* \Psi \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \hat{\Psi}_{\tilde{F},F}/P\right]$$

in probability. Next, use the Sherman–Morrison formula (A.13) to write

$$
\begin{aligned}
&\frac{1}{P}\,\mathrm{tr}\left[\Psi\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\hat{\Psi}_{\tilde{F},F}\right]\\
&=\frac{1}{P}\,\mathrm{tr}\left[\Psi\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\frac{1}{T}\sum_t \tilde{F}_t\,F_t^{\intercal}\right]\\
&=\frac{1}{P}\,\mathrm{tr}\left[\frac{1}{T}\sum_t \Psi\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{F}_t\,F_t^{\intercal}\right]\\
&=\frac{1}{P}\,\mathrm{tr}\left[\frac{1}{T}\sum_t \Psi\left(\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\right.\right.\\
&\quad\left.\left.-\frac{1}{T}\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t\left(I+\frac{1}{T}\tilde{F}_t^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t\right)^{-1}\tilde{F}_t^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\right)\tilde{F}_tF_t^{\intercal}\right]\\
&=\frac{1}{P}\,\mathrm{tr}\left[\frac{1}{T}\sum_t \Psi\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_tF_t^{\intercal}\right]\\
&\quad-\frac{1}{P}\,\mathrm{tr}\left[\frac{1}{T}\sum_t \Psi\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t\,(I+C_{T,t})^{-1}\,C_{T,t}F_t^{\intercal}\right]\\
&=\frac{1}{P}\,\mathrm{tr}\left[\frac{1}{T}\sum_t F_t^{\intercal}\Psi\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t\right] \qquad \text{(A.21)}\\
&\quad-\frac{1}{P}\,\mathrm{tr}\left[\frac{1}{T}\sum_t F_t^{\intercal}\,\Psi\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t\,(1+C_{T,t})^{-1}\,C_{T,t}\right], \qquad \text{(A.22)}
\end{aligned}
$$

where

$$C_{T,t}=\frac{1}{T}\,\tilde{F}_t^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t\,.$$

Now, because $\tilde{F}_t$ and F_t are independent of $\hat{\tilde{\Psi}}_{T,t}$, by Lemma 3, 9 and Proposition 1 we have

$$C_{T,t}\to\xi(z;c)$$

in probability. For $\frac{1}{P}\,F_t^{\intercal}\,\Psi\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t$, note that

$$F_t^{\intercal}\,\Psi\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t=Z_t^{\intercal}\,(\Psi^{1/2},0)^{\intercal}\,\Psi\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}(\Psi^{1/2},\Phi^{1/2})\,Z_t$$

and

$$\begin{aligned}
&\left\| (\Psi^{1/2}, 0)^\intercal \, \Psi \, \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \, (\Psi^{1/2}, \Phi^{1/2}) \right\| \\
&\leq \left\| \Psi^{1/2} \, \Psi \, \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \, \Psi^{1/2} \, \Psi \right\| + \left\| \Psi^{1/2} \, \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \, \Phi^{1/2} \right\| \\
&\leq 2\, z^{-1} \, ||\tilde{\Psi}||^2
\end{aligned}$$

is uniformly bounded. Hence, using Lemmas 3 and 9 and the same argument as that in the proof of Proposition 1, we know that

$$\frac{1}{P} \, F_t^\intercal \, \Psi \, \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \, \tilde{F}_t \to \frac{1}{P} \, \mathrm{tr} \left(\Psi \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \, \Psi^\intercal \right) \to c^{-1} \, \xi_\Psi(z; c)$$

in probability. Therefore,

$$\frac{1}{P} \left[F_t^\intercal \, \Psi \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \, \tilde{F}_t \, (1 + C_T)^{-1} \, C_T \right] \to c^{-1} \, \xi_\Psi(z; c) \, \xi(z; c) / (1 + \xi(z; c))$$

and hence

$$\begin{aligned}
M_t &= \frac{1}{P} \, F_t^\intercal \, \Psi \, \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \, \tilde{F}_t - \frac{1}{P} \, F_t^\intercal \, \Psi \, \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \, \tilde{F}_t \, (1 + C_{T,t})^{-1} \, C_{T,t} \\
&\to c^{-1} \, \xi_\Psi(z; c) - c^{-1} \, \xi_\Psi(z; c) \, \xi(z; c) / (1 + \xi(z; c)) \\
&= c^{-1} \, \xi_\Psi(z; c) / (1 + \xi(z; c))
\end{aligned}$$

in probability. Note from Lemma 3, equation (A.15) and the proof of Lemma 9 that

$$\frac{1}{P} \, F_t^\intercal \, \Psi \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \, \tilde{F}_t \to \frac{1}{P} \, \mathrm{tr} \left(\Psi \, \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \, \Psi^\intercal \right)$$

in L_2 and that

$$\frac{1}{P} \, \mathrm{tr} \left(\Psi \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \, \Psi^\intercal \right) \to \mathbb{E} \left[\frac{1}{P} \, \mathrm{tr} \left(\Psi \, \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \, \Psi^\intercal \right) \right]$$

in L_2. We have

$$\frac{1}{P} \, F_t^\intercal \, \Psi \, \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \, \tilde{F}_t \to c^{-1} \, \xi_\Psi(z; c)$$

in L_2 because $\mathbb{E}\left[\frac{1}{P}\,\mathrm{tr}\left(\Psi\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\Psi^{\intercal}\right)\right]$ is not random and

$$\mathbb{E}\left[\frac{1}{P}\,\mathrm{tr}\left(\Psi\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\Psi^{\intercal}\right)\right]\to c^{-1}\,\xi_{\Psi}(z;c)\,.$$

Similarly, we have

$$C_{T,t}\to\xi(z;c)$$

in L_2. Therefore, we have

$$\frac{1}{P}\,F_t^{\intercal}\,\Psi\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\,\tilde{F}_t\,(1+C_{T,t})^{-1}\,C_{T,t}\to\frac{1}{P}\,F_t^{\intercal}\,\Psi\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\,\tilde{F}_t\,(1+C_{T,t})^{-1}\,C_{T,t}$$

in L_2 because

$$(X_nY_n-X\,Y)^2=((X_n-X)\,Y_n+X\,(Y_n-Y))^2\le 2\,(X_n-X)^2\,Y_n^2+X^2\,(Y_n-Y)^2$$

and

$$\left(\frac{X_n}{1+X_n}-\frac{X}{1+X}\right)^2=\frac{(X_n-X)^2}{(1+X_n)^2\,(1+X)^2}\le(X_n-X)^2\,.$$

For any $X_n\to X$ and $Y_n\to Y$. Therefore, we have

$$M_t\to c^{-1}\,\xi_{\Psi}(z;c)/(1+\xi(z;c))$$

in L_2. Hence, we write

$$\frac{1}{T^2}\,\mathbb{E}\left(\sum_t(M_t-\lim M_t)\right)^2\le 2\,\mathbb{E}(M_t-\lim M_t)^2\to 0\,.$$

Therefore,

$$\frac{1}{P}\,\mathrm{tr}\left[\Psi\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\hat{\Psi}_{\tilde{F},F}\right]=\frac{1}{T}\,\sum_t M_t\to c^{-1}\,\xi_{\Psi}(z;c)/(1+\xi(z;c))$$

in L_2 and thus in probability. Altogether, we have

$$\begin{aligned}\mathcal{E}(z;c) &= \lim \mathbb{E}\left[\hat{\pi}_t(z)\, R_{t+1} \mid \hat{\beta}\right] \\ &= \lim \left(\hat{\beta}^\intercal\, \Psi\, \beta\right) \\ &= \lim \left(b_*\, \Psi\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\, \hat{\Psi}_{\tilde{F},F}/P\right) \\ &= b_*\, c^{-1}\, \xi_\Psi(z;c)/(1+\xi(z;c))\,. \end{aligned} \tag{A.23}$$

The first part of the proof is completed. Now we turn to the second part of the proof, where we calculate $\lim \mathbb{E}\left[\hat{\pi}_t\,(z)^2 \mid \hat{\beta}\right] = \lim \operatorname{tr}\left(\tilde{\Psi}\, \hat{\beta}\, \hat{\beta}^\intercal\right)$. Substitute in the expression (A.16), and we know that (here we do not have the intersection term because (A.19) tells us that it will go to 0):

$$\begin{aligned}\operatorname{tr}\left(\tilde{\Psi}\, \hat{\beta}\, \hat{\beta}^\intercal\right) &= \operatorname{tr}\left(\tilde{\Psi}\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\, \left(\hat{\Psi}_{\tilde{F},F}\, \beta\, \beta^\intercal\, \hat{\Psi}_{\tilde{F},F}^\intercal + q_T\, q_T^\intercal\right)\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\right) \\ &= \operatorname{tr}\left(\tilde{\Psi}\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\, \hat{\Psi}_{\tilde{F},F}\, \beta\, \beta^\intercal\, \hat{\Psi}_{\tilde{F},F}^\intercal\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\right) \\ &\quad + \operatorname{tr}\left(\tilde{\Psi}\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\, q_T\, q_T^\intercal\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\right) \\ &= \text{ Term 1} + \text{ Term 2}\,. \end{aligned} \tag{A.24}$$

First we deal with Term 2. Note that

$$\operatorname{tr}\left(\tilde{\Psi}\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\, q_T\, q_T^\intercal\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\right) = q_T^\intercal\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\, \tilde{\Psi}\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} q_T\,.$$

Let $A_T = \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\, \tilde{\Psi}\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}$, now we prove the following Lemma:

Lemma 11. *$q_T^\intercal\, A_T\, q_T - T^{-1} \operatorname{tr}\left(A_T\, \hat{\tilde{\Psi}}_T\right) \to 0$ in L_2 and hence in probability.*

Proof. Let $D_T = q_T^\intercal\, A_T\, q_T$. From (A.18) we know that

$$\mathbb{E}\left(D_T \mid \{\tilde{F}_t\}\right) = T^{-1} \operatorname{tr}\left(A_T\, \hat{\tilde{\Psi}}_T\right)\,,$$

where $\{\tilde{F}_t\} = \{\tilde{F}_t, \tilde{F}_{t-1}, \ldots \tilde{F}_{t-T}\}$ are the past noisy features, so it suffices to

prove that

$$\mathbb{E}\left[D_T^2 \mid \{\tilde{F}_t\}\right] - \mathbb{E}\left[D_T \mid \{\tilde{F}_t\}\right]^2 \to 0\,.$$

We have

$$\begin{aligned} q_T^\intercal\, A_T\, q_T &= \sum_{i_1,i_2} A_{T,i_1,i_2}\, q_{i_1}\, q_{i_2} \\ &= T^{-2} \sum_{i_1,i_2} A_{T,i_1,i_2} \sum_{t_1} \tilde{F}_{i_1,t_1}\, \epsilon_{t_1+1} \sum_{t_2} \tilde{F}_{i_2,t_2}\, \epsilon_{t_2+1} \\ &= T^{-1} \sum_{t_1,t_2} \epsilon_{t_1+1}\, \epsilon_{t_2+1} \left(T^{-1} \sum_{i_1,i_2} A_{T,i_1,i_2}\, \tilde{F}_{i_1,t_1} \tilde{F}_{i_2,t_2} \right) \\ &= \epsilon^\intercal\, B_T\, \epsilon\,, \end{aligned}$$

where $\epsilon = (\epsilon_{t-T+1}, ...\epsilon_t)^\intercal$ and

$$B_{T,t_1,t_2} = T^{-2}\, \tilde{F}_{t_1}^\intercal\, A_T\, \tilde{F}_{t_2}\,.$$

By the proof of Lemma 9 we know that

$$\mathbb{E}\left[D_T^2 \mid \{\tilde{F}_t\}\right] - \mathbb{E}\left[D_T \mid \{\tilde{F}_t\}\right]^2 \leq K\, ||B_T||_F^2\,.$$

Therefore,

$$\mathbb{E}\left[\mathbb{E}\left[D_T^2 \mid \{\tilde{F}_t\}\right] - \mathbb{E}\left[D_T \mid \{\tilde{F}_t\}\right]^2\right] \leq K\, E\, ||B_T||_F^2\,.$$

Recall that we write $\tilde{F} = (\tilde{F}_t)$, as $||A_T|| \leq z^{-2}\, ||\tilde{\Psi}||$ is uniformly bounded,

using Lemma 4 we have

$$
\begin{aligned}
\mathbb{E}\left[||B_T||_F^2\right] &= T^{-4}\ \mathrm{tr}\,\mathbb{E}\left[\tilde{F}^\intercal\, A_T\, \tilde{F}\, \tilde{F}^\intercal\, A_T^\intercal\, \tilde{F}\right] \\
&\leq z^{-4}\, ||\tilde{\Psi}||^2\, T^{-4} \sum_{t_1} \mathrm{tr}\,\mathbb{E}\left[\tilde{F}_{t_1}^\intercal\, \tilde{F}\, \tilde{F}^\intercal \tilde{F}_{t_1}\right] \\
&= C\, T^{-4} \sum_{t_1} \mathrm{tr}\,\mathbb{E}\left[\tilde{F}_{t_1}^\intercal \sum_{t_2} \tilde{F}_{t_2}\, \tilde{F}_{t_2}^\intercal\, \tilde{F}_{t_1}\right] \\
&= C\, T^{-4} \sum_{t_1,t_2} \mathbb{E}\left[\left(\tilde{F}_{t_1}^\intercal\, \tilde{F}_{t_2}\right)^2\right] \\
&= T^{-4} \sum_{t_1=t_2} \mathbb{E}\left[\left(\tilde{F}_{t_1}^\intercal\, \tilde{F}_{t_2}\right)^2\right] + T^{-4} \sum_{t_1\neq t_2} \mathbb{E}\left[\left(\tilde{F}_{t_1}^\intercal\, \tilde{F}_{t_2}\right)^2\right] \\
&= C\, T^{-4} T\, \mathbb{E}\left[\left(\tilde{F}_t^\intercal \tilde{F}_t\right)^2\right] + C\, T^{-4} \sum_{t_1\neq t_2} \mathbb{E}\left[\tilde{F}_{t_1}^\intercal\, \tilde{F}_{t_2} \tilde{F}_{t_2}^\intercal \tilde{F}_{t_1}\right] \\
&= C\, T^{-3}\, \mathbb{E}\left[\left(\tilde{F}_t^\intercal \tilde{F}_t\right)^2\right] + C\, T^{-4} \sum_{t_1\neq t_2} \mathbb{E}\left[\tilde{F}_{t_1}^\intercal\, \tilde{\Psi}^\intercal \tilde{F}_{t_1}\right] \\
&= C\, T^{-3}\, \mathbb{E}\left[\left(\tilde{F}_t^\intercal \tilde{F}_t\right)^2\right] + C\, T^{-4} T(T-1)\, \mathrm{tr}\ \mathbb{E}\left[\tilde{\Psi}^\intercal \tilde{\Psi}\right] \\
&\leq C\, T^{-3}\, \mathbb{E}\left[\left(\tilde{F}_t^\intercal\, \tilde{F}_t\right)^2\right] + C\, T^{-2}\, P\, \left\|\tilde{\Psi}^\intercal\, \tilde{\Psi}\right\| \\
&= C\, T^{-3}\, \mathbb{E}\left[\left(\tilde{F}_t^\intercal\, \tilde{F}_t\right)^2\right] + C\, T^{-1}\, c\, \left\|\tilde{\Psi}^\intercal\, \tilde{\Psi}\right\| \\
&\leq C\, ||\tilde{\Psi}||^2\, T^{-3}\, (K\, T + T\, (T-1))) + C\, T^{-1}\, c \left\|\tilde{\Psi}^\intercal\, \tilde{\Psi}\right\| \\
&\to 0\,.
\end{aligned}
$$

Hence, we proved that $q_T^\intercal\, A_T\, q_T - T^{-1}\ \mathrm{tr}\left(A_T\, \hat{\tilde{\Psi}}_T\right) \to 0$ in L_2 and hence in probability. □

Now, for $T^{-1}\,\mathrm{tr}\left(A_T\,\hat{\tilde{\Psi}}_T\right)$ we have

$$\begin{aligned}
T^{-1}\,\mathrm{tr}\left(A_T\,\hat{\tilde{\Psi}}_T\right) &= T^{-1}\,\mathrm{tr}\left(\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\hat{\tilde{\Psi}}_T\right)\\
&= T^{-1}\,\mathrm{tr}\left(\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\left(z\,I+\hat{\tilde{\Psi}}_T-z\,I\right)\right)\\
&= T^{-1}\,\mathrm{tr}\left(\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\right)-z\,T^{-1}\,\mathrm{tr}\left(\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-2}\right).
\end{aligned}$$

By Proposition 1 and Lemma 6 we know that

$$T^{-1}\,\mathrm{tr}\left(\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\right)\to T^{-1}\,\mathbb{E}\left[\mathrm{tr}\left(\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\right)\right]\to\xi(z;c)\,.$$

Note that $\frac{d}{dz}T^{-1}\,\mathrm{tr}\left(\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\right)$ is uniformly bounded, by Lebesgue's dominated convergence theorem we have

$$\begin{aligned}
\frac{d}{dz}T^{-1}\,\mathbb{E}\left[\mathrm{tr}\left(\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\right)\right] &= T^{-1}\,\mathbb{E}\left[\frac{d}{dz}\,\mathrm{tr}\left(\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\right)\right]\\
&= -T^{-1}\,\mathbb{E}\left[\mathrm{tr}\left(\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-2}\tilde{\Psi}\right)\right].
\end{aligned}$$

Also, because for any smooth function f_T,

$$f_T(z+\Delta z)=f_T(z)+f_T'(z)\,\Delta z+\frac{1}{2}\int_z^{z+\Delta z}f_T''(t)\,(z+\Delta z-t)\,\mathrm{dt}\,,\quad\text{(A.25)}$$

if $f_T(z)\to 0$ for all z and $f_T'(z)$ and $f_T''(z)$ are all bounded then we have $f_T'(z)\to 0$. Therefore, we have

$$\frac{d}{dz}T^{-1}\,\mathbb{E}\left[\mathrm{tr}\left(\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\right)\right]\to\frac{d}{dz}\xi(z;c)\,.$$

Also, using Lemma 10 we have

$$\begin{aligned} -z\,T^{-1}\,\mathrm{tr}\left(\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-2}\right) &\to -z\,T^{-1}\,\mathbb{E}\left[\mathrm{tr}\left(\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-2}\tilde{\Psi}\right)\right] \\ &\to z\,\frac{d}{dz}\,T^{-1}\,\mathbb{E}\left[\mathrm{tr}\left(\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\right)\right] \\ &\to z\frac{d}{dz}\xi(z;c)\,. \end{aligned} \tag{A.26}$$

Altogether, we have

$$\text{Term 2} \to \xi(z;c) + z\,\frac{d}{dz}\xi(z;c)\,.$$

Now we deal with Term 1 in (A.24). Note that

$$\text{Term 1} = \beta^\intercal\,\hat{\Psi}^\intercal_{\tilde{F},F}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\hat{\Psi}_{\tilde{F},F}\,\beta\,,$$

and using Lemma 9 we have

$$\begin{aligned} \text{Term 1} &\to P^{-1}\,b_*\,\mathrm{tr}\left(\hat{\Psi}^\intercal_{\tilde{F},F}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\hat{\Psi}_{\tilde{F},F}\right) \\ &= P^{-1}\,b_*\,\mathrm{tr}\left(\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\hat{\Psi}_{\tilde{F},F}\hat{\Psi}^\intercal_{\tilde{F},F}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right) \\ &= P^{-1}\,b_*\,\frac{1}{T^2}\,\mathrm{tr}\left(\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\sum_{t_1,t_2}\tilde{F}_{t_1}\,F^\intercal_{t_1}\,F_{t_2}\,\tilde{F}^\intercal_{t_2}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right) \\ &= P^{-1}\,b_*\,\frac{1}{T^2}\sum_{t_1}\mathrm{tr}\left(\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{F}_{t_1}\,F^\intercal_{t_1}\,F_{t_1}\,\tilde{F}^\intercal_{t_1}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right) \\ &\quad + P^{-1}\,b_*\,\frac{1}{T^2}\sum_{t_1\neq t_2}\mathrm{tr}\left(\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{F}_{t_1}F^\intercal_{t_1}\,F_{t_2}\,\tilde{F}^\intercal_{t_2}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right) \\ &= \text{Term 11} + \text{Term 12}\,. \end{aligned} \tag{A.27}$$

We deal with each term separately. First we deal with

$$M_t = \frac{1}{P\,T}\,\mathrm{tr}\left(\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{F}_t\,F^\intercal_t\,F_t\,\tilde{F}^\intercal_t\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right)\,. \tag{A.28}$$

Next, use the Sherman–Morrison formula (A.13) and define $C_{T,t} = \tilde{F}^\intercal_t\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t$,

and write

$$\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1} \tilde{F}_t = \left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{F}_t \,(1 + C_{T,t})^{-1} \,.$$

Therefore, we have

$$\begin{aligned} M_t &= \frac{1}{PT}\,\mathrm{tr}\left(\tilde{\Psi}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{F}_t \,(1 + C_{T,t})^{-1}\, F_t^\intercal\, F_t \,(1 + C_{T,t})^{-1}\, \tilde{F}_t^\intercal \left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\right) \\ &= \frac{1}{PT}\,\tilde{F}_t^\intercal \left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{\Psi}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{F}_t\, F_t^\intercal\, F_t \,(1 + C_{T,t})^{-2} \,. \end{aligned}$$

Now, use the fact that $\tilde{F}_t$ and F_t are independent from $\hat{\tilde{\Psi}}_{T,t}$ and Lemma 3 and 6 to write

$$\frac{1}{T}\,\tilde{F}_t^\intercal \left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{\Psi}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{F}_t \to \frac{1}{T}\,\mathrm{tr}\left(\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{\Psi}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{\Psi}\right)$$

in L_2 and that

$$\frac{1}{P}\,F_t^\intercal\, F_t \,(1 + C_{T,t})^{-2} \to \frac{1}{P}\,\mathrm{tr}(\Psi)\,(1 + \xi(z;c))^{-2} \to \psi_{*,1}(1 + \xi(z;c))^{-2}$$

in L_2. Similar to the proof of Lemma 10 for $\left(\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{\Psi}\right)^2$ we have

$$\frac{1}{T}\,\mathrm{tr}\left(\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{\Psi}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{\Psi}\right) \to \frac{1}{T}\,\mathbb{E}\left[\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{\Psi}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{\Psi}\right]$$

in L_2. From (A.26) we have that

$$-T^{-1}\,\mathbb{E}\left[\mathrm{tr}\left(\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-2} \tilde{\Psi}\right)\right] \to \frac{d}{dz}\xi(z;c)$$

in L_2. In the next equation we use $X \approx Y$ to denote that $X - Y \to 0$ in L_2.

$$
\begin{aligned}
\xi(z;c) &\approx \frac{1}{T}\,\mathrm{tr}\,\mathbb{E}\left[\tilde{\Psi}\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\right] \\
&= \frac{1}{T}\,\mathrm{tr}\,\mathbb{E}\left[\left(z\,I + \hat{\tilde{\Psi}}_T\right)\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\,\tilde{\Psi}\,\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\right] \\
&\approx -z\frac{d}{dz}\xi(z;c) + \frac{1}{T}\,\mathrm{tr}\,\mathbb{E}\left[\hat{\tilde{\Psi}}_T\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\,\tilde{\Psi}\,\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\right] \\
&= -z\,\frac{d}{dz}\xi(z;c) + \frac{1}{T}\,\mathrm{tr}\,\mathbb{E}\left[\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\,\tilde{F}_t\tilde{F}_t^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\right] \\
&= -z\,\frac{d}{dz}\xi(z;c) + \frac{1}{T}\,\mathrm{tr}\,\mathbb{E}\left[(1+C_T)^{-2}\,\tilde{F}_t^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\,\tilde{\Psi}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\,\tilde{F}_t\right] \\
&\approx -z\,\frac{d}{dz}\xi(z;c) + (1+\xi(z;c))^{-2}\,\frac{1}{T}\,\mathbb{E}\left[\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\,\tilde{\Psi}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\,\tilde{\Psi}\right],
\end{aligned}
$$

where the third equation follows from symmetric and the forth equation follows from the Sherman–Morrison's formula (A.13). Therefore, we have

$$
\frac{1}{T}\,\mathbb{E}\left[\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\,\tilde{\Psi}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\,\tilde{\Psi}\right] \to \left(\xi(z;c) + z\,\frac{d}{dz}\xi(z;c)\right)\,(1+\xi(z;c))^2
$$

and

$$
M_t \to \psi_{*,1}\,\left(\xi(z;c) + z\,\frac{d}{dz}\xi(z;c)\right)
$$

in L_2 and therefore

$$
\text{Term 11} = \frac{b_*}{T}\sum_{t_1} M_t \to b_*\,\psi_{*,1}\left(\xi(z;c) + z\,\frac{d}{dz}\xi(z;c)\right)
$$

in L_2 and thus in probability. Now we deal with Term 12 in (A.27). Consider

$$
M_{t_1,t_2} = P^{-1}\,\mathrm{tr}\left(\tilde{\Psi}\,\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\,\tilde{F}_{t_1}\,F_{t_1}^{\intercal}\,F_{t_2}\,\tilde{F}_{t_2}^{\intercal}\,\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\right),
$$

using the Sherman–Morrison formula (A.13), write $\hat{\tilde{\Psi}}_{T,t_1,t_2} = \frac{1}{T}\sum_{t\neq t_1,t_2}\tilde{F}_t\,\tilde{F}_t^{\intercal}$, note that in the following equation, the C_T we are using might be slightly different from line to line. In each line, C_T can denote one of the following values:

$$
C_{T,t_1} = \frac{1}{T}\,\tilde{F}_{t_1}^{\intercal}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1}\right)^{-1}\,\tilde{F}_{t_1}\,, \qquad C_{T,t_2} = \frac{1}{T}\,\tilde{F}_{t_2}^{\intercal}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_2}\right)^{-1}\,\tilde{F}_{t_2}\,,
$$

$$C_{T,t_1,t_2} = \frac{1}{T}\,\tilde{F}_{t_1}^{\intercal}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,, \qquad C_{T,t_2,t_1} = \frac{1}{T}\,\tilde{F}_{t_2}^{\intercal}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_2}\,.$$

However, as all those values converge to $\xi(z;c)$ in L_2. In the following equation we abuse the notation and treat all values as the same. We have that

$$\begin{aligned}
M_{t_1,t_2} &= P^{-1}\,\mathrm{tr}\left(\tilde{\Psi}\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\,\tilde{F}_{t_1}F_{t_1}^{\intercal}F_{t_2}\tilde{F}_{t_2}^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\right)\\
&= P^{-1}\,\mathrm{tr}\left(\tilde{F}_{t_2}^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\,\tilde{\Psi}\,\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\,\tilde{F}_{t_1}F_{t_1}^{\intercal}F_{t_2}\right)\\
&= P^{-1}\,\mathrm{tr}\left((1+C_T)^{-1}\,\tilde{F}_{t_2}^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t_2}\right)^{-1}\,\tilde{\Psi}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1}\right)^{-1}\,\tilde{F}_{t_1}\,(1+C_T)^{-1}\,F_{t_1}^{\intercal}F_{t_2}\right)\\
&= P^{-1}\,\mathrm{tr}\Big\{(1+C_T)^{-1}\,\tilde{F}_{t_2}^{\intercal}\\
&\times\left(\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1} - \frac{1}{T}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,(1+C_T)^{-1}\,\tilde{F}_{t_1}^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\right)\\
&\times\tilde{\Psi}\left(\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1} - \frac{1}{T}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_2}\,(1+C_T)^{-1}\,\tilde{F}_{t_2}^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\right)\\
&\tilde{F}_{t_1}\,(1+C_T)^{-1}\,F_{t_1}^{\intercal}F_{t_2}\Big\}\\
&= P^{-1}\,\mathrm{tr}\Big\{(1+C_T)^{-1}\,\tilde{F}_{t_2}^{\intercal}\\
&\times\left(\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1} - \frac{1}{T}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,(1+C_T)^{-1}\,\tilde{F}_{t_1}^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\right)\\
&\times\tilde{\Psi}\left(\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1} - \frac{1}{T}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_2}\,(1+C_T)^{-1}\,\tilde{F}_{t_2}^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\right)\\
&\tilde{F}_{t_1}\,(1+C_T)^{-1}\,F_{t_1}^{\intercal}F_{t_2}\Big\}\\
&= P^{-1}\,\mathrm{tr}\Big\{(1+C_T)^{-1}\,\tilde{F}_{t_2}^{\intercal}\\
&\times\Big(\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{\Psi}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\\
&-\frac{1}{T}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,(1+C_T)^{-1}\,\tilde{F}_{t_1}^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{\Psi}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\\
&-\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{\Psi}\frac{1}{T}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_2}\,(1+C_T)^{-1}\,\tilde{F}_{t_2}^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\\
&+\frac{1}{T}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,(1+C_T)^{-1}\,\tilde{F}_{t_1}^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{\Psi}\\
&\times\frac{1}{T}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_2}\,(1+C_T)^{-1}\,\tilde{F}_{t_2}^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\Big)\,\tilde{F}_{t_1}\,(1+C_T)^{-1}\,F_{t_1}^{\intercal}F_{t_2}\Big\}\\
&= M_{t_1,t_2,1} + M_{t_1,t_2,2} + M_{t_1,t_2,3} + M_{t_1,t_2,4}\,. \qquad \text{(A.29)}
\end{aligned}$$

Now we deal with each term separately. First we deal with $M_{t_1,t_2,1}$. We know that

$$M_{t_1,t_2,1} = P^{-1}\left(1+C_T\right)^{-2}\tilde{F}_{t_2}^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{F}_{t_1}\,F_{t_1}^{\intercal}F_{t_2}$$

and

$$\frac{1}{T}\sum_{t_1}M_{t_1,t_2,1} = \frac{\tilde{F}_{t_2}^{\intercal}\left(\frac{1}{T}\sum_{t_1}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{F}_{t_1}\,F_{t_1}^{\intercal}\right)F_{t_2}}{P\left(1+C_T\right)^2}.$$

Consider $\frac{1}{PT}\sum_{t_1}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{F}_{t_1}\,F_{t_1}^{\intercal}$ and let

$$A_{t_1} = \left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1},$$

so we have

$$\sum_{t_1}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{F}_{t_1}\,F_{t_1}^{\intercal} = \left(A_{t-T}\,\tilde{F}_{t-T},...,A_{t-1}\,\tilde{F}_{t-1}\right)F^{\intercal}$$

and therefore

$$\begin{aligned}\left\|\sum_{t_1}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{F}_{t_1}\,F_{t_1}^{\intercal}\right\|_F^2 &= \sum_{t_1}\mathrm{tr}\left(A_{t_1}\tilde{F}_{t_1}F^{\intercal}F\tilde{F}_{t_1}^{\intercal}A_{t_1}^{\intercal}\right)\\ &\leq z^{-4}\left\|\tilde{\Psi}\right\|^2\sum_{t_1}\mathrm{tr}\left(\tilde{F}_{t_1}\,F^{\intercal}\,F\,\tilde{F}_{t_1}^{\intercal}\right).\end{aligned} \tag{A.30}$$

Therefore, as $\tilde{F}_{t_1},\tilde{F}_{t_2},F_{t_1},F_{t_2}$ are all independent from $\hat{\tilde{\Psi}}_{T,t_1,t_2}$ and $\tilde{F}_{t_1},F_{t_2}$ are independent from $\tilde{F}_{t_2},F_{t_2}$, with the same arguments in Lemma 11 we have

$$\frac{1}{T}\sum_{t_1}M_{t_1,t_2,1} \approx \frac{\sum_{t_1}\mathrm{tr}\left(\Psi\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{F}_{t_1}\,F_{t_1}^{\intercal}\right)}{P\,T\left(1+\xi(z;c)\right)^2}.$$

With the same arguments in calculating Term 11 in (A.27), using Lemma 3,

6, and Assumption 3 we have that

$$\begin{aligned}
\frac{1}{T}\sum_{t_1} M_{t_1,t_2,1} &\approx P^{-1}\left(1+\xi(z;c)\right)^{-2}\,\mathrm{tr}\left(\Psi\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{F}_{t_1}F_{t_1}^{\intercal}\right)\\
&= P^{-1}\left(1+\xi(z;c)\right)^{-2}\,\mathrm{tr}\left(F_{t_1}^{\intercal}\Psi\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{F}_{t_1}\right)\\
&\approx P^{-1}\left(1+\xi(z;c)\right)^{-2}\,\mathrm{tr}\left(\Psi^2\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\right)\\
&\approx P^{-1}\left(1+\xi(z;c)\right)^{-2}\mathbb{E}\left[\mathrm{tr}\left(\Psi^2\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\right)\right]\\
&\approx \frac{c^{-1}\,\eta_{\Psi}(z;c)}{\left(1+\xi(z;c)\right)^2}\,,
\end{aligned}$$

where we recall that $X\approx Y$ mean that $X-Y\to 0$ in L_2. Therefore we have

$$\frac{1}{T^2}\sum_{t_1\neq t_2} M_{t_1,t_2,1}\to\frac{c^{-1}\eta_{\Psi}(z;c)}{\left(1+\xi(z;c)\right)^2}$$

in L_2. Now we turn to $M_{t_1,t_2,2}$ in (A.29), we have

$$\begin{aligned}
M_{t_1,t_2,2} =& -P^{-1}\Big((1+C_T)^{-1}\,\tilde{F}_{t_2}^{\intercal}\\
&\times\left(\frac{1}{T}\,z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{F}_{t_1}\left(1+C_T\right)^{-1}\tilde{F}_{t_1}^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{\Psi}\\
&\times\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{F}_{t_1}\left(1+C_T\right)^{-1}F_{t_1}^{\intercal}\Big)F_{t_2}\\
=&\frac{-1}{P\,T}\left(1+C_T\right)^{-3}\tilde{F}_{t_2}^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{F}_{t_1}\tilde{F}_{t_1}^{\intercal}\\
&\times\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{F}_{t_1}F_{t_1}^{\intercal}F_{t_2}\,.
\end{aligned}$$

Let

$$P_{t_1,t_2}=\tilde{F}_{t_1}^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{F}_{t_1}$$

and

$$Q_{t_1,t_2}=\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}=V_{t_1,t_2}^{\intercal}\,\Lambda_{t_1,t_2}\,V_{t_1,t_2}$$

where V_{t_1,t_2} is an orthogonal matrix and $\Lambda_{t_1,t_2}=\mathrm{diag}(\lambda_{1,t_1,t_2},...\lambda_{P,t_1,t_2})$ is the diagonal matrix of the eigenvalues of Q_{t_1,t_2}. Write $V_{t_1,t_2}\,\tilde{F}_{t_1}=(V_{F,t_1,t_2,1},...,V_{F,t_1,t_2,P})$,

we have

$$P_{t_1,t_2} \leq ||\tilde{\Psi}||\,(V_{t_1,t_2}\,\tilde{F}_{t_1})^{\intercal}\,\Lambda_{t_1,t_2}\,V_{t_1,t_2}\,\tilde{F}_{t_1} \leq z^{-1}\,||\tilde{\Psi}||\sum_i \lambda_{t_1,t_2,i}\,V^2_{F,t_1,t_2,i}\,.$$

We know that

$$C_{T,t_1,t_2} = \frac{1}{T}\,\tilde{F}_{t_1}^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1} = \sum_i \lambda_{t_1,t_2,i}\,V^2_{F,t_1,t_2,i}\,.$$

Therefore,

$$P_{t_1,t_2}/(1+C_T) \leq z^{-1}\,||\tilde{\Psi}|| \tag{A.31}$$

is uniformly bounded. Hence

$$\begin{aligned}
&\left\|(1+C_T)^{-1}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,\tilde{F}_{t_1}^{\intercal}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{\Psi}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,F_{t_1}^{\intercal}\right\|_F^2 \\
&\leq z^{-2}\,||\tilde{\Psi}||^2\,\left\|Q_{t_1,t_2}\tilde{F}_{t_1}\,F_{t_1}^{\intercal}\right\|_F^2 \\
&\leq z^{-6}\,\left\|\tilde{\Psi}\right\|^4\,\mathrm{tr}\left(\tilde{F}_{t_1}\,F_{t_1}^{\intercal}\,F_{t_1}\,\tilde{F}_{t_1}^{\intercal}\right)\,.
\end{aligned}$$

Therefore, because $\tilde{F}_{t_1}, \tilde{F}_{t_2}, F_{t_1}, F_{t_2}$ are all independent from $\hat{\tilde{\Psi}}_{T,t_1,t_2}$ and $\tilde{F}_{t_1}, F_{t_2}$ are independent from $\tilde{F}_{t_2}, F_{t_2}$, with the same arguments as those in Lemma 11 we have

$$\begin{aligned}
M_{t_1,t_2,2} \approx \frac{-1}{P\,T}\,(1+\xi(z;c))^{-3}\,\mathrm{tr}\Bigg(&F_{t_1}^{\intercal}\,\Psi\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1} \\
&\times\tilde{F}_{t_1}^{\intercal}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{\Psi}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\Bigg)\,.
\end{aligned}$$

From the calculation for Term 11 in (A.27) we have

$$\frac{1}{T}\,\tilde{F}_{t_1}^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{\Psi}\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1} \to \left(\xi(z;c) + z\,\frac{d}{dz}\xi(z;c)\right)\left(1 + \frac{d}{dz}\xi(z;c)\right)^2$$

in L_2. Also, by Lemma 3 and Assumption 3 we have that

$$\frac{1}{P}\,F_{t_1}^{\intercal}\,\Psi\,\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1} \approx \frac{1}{P}\,\mathbb{E}\left[\left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\Psi^2\right] \to c^{-1}\,\xi_{\Psi}(z;c)$$

in L_2. Therefore, we have that

$$M_{t_1,t_2,2} \to -\frac{c^{-1}\ \left(\xi(z;c) + z\ \frac{d}{dz}\xi(z;c)\right)\ \xi_\Psi(z;c)}{1 + \frac{d}{dz}\xi(z;c)}$$

in L_2 and therefore

$$\frac{1}{T^2}\sum_{t_1 \neq t_2} M_{t_1,t_2,2} \to -\frac{c^{-1}\ \left(\xi(z;c) + z\ \frac{d}{dz}\xi(z;c)\right)\ \xi_\Psi(z;c)}{1 + \frac{d}{dz}\xi(z;c)}$$

in L_2. For $M_{t_1,t_2,3}$ in (A.29), with the same arguments as in the calculation of $M_{t_1,t_2,2}$ we have

$$(1 + C_{T,t_2,t_1})^{-1}\ \tilde{F}_{t_2}^\intercal\ \left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\ \tilde{\Psi}\ \left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\ \tilde{F}_{t_2} \leq z^{-1}\,||\tilde{\Psi}||$$

and hence using the same arguments in the calculation of $M_{t_1,t_2,2}$ we have

$$\begin{aligned}
M_{t_1,t_2,3} =& -P^{-1}\ \Big((1 + C_T)^{-1}\ \tilde{F}_{t_2}^\intercal \\
&\times \frac{1}{T}\ \left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\ \tilde{\Psi}\ \left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\ \tilde{F}_{t_2}\ (1 + C_T)^{-1}\ \tilde{F}_{t_2}^\intercal\ \left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1} \\
&\times \tilde{F}_{t_1}\ (1 + C_T)^{-1}\ F_{t_1}^\intercal\,F_{t_2}\Big) \\
\approx& \frac{1}{P\,T}\,(1 + \xi(z;c))^{-3}\,\tilde{F}_{t_2}^\intercal\ \left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\ \tilde{\Psi} \\
&\times \left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\ \tilde{F}_{t_2}\,\tilde{F}_{t_2}^\intercal\ \left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\ \tilde{F}_{t_1}\,F_{t_1}^\intercal\,F_{t_2} \\
=& \frac{1}{P\,T}\,(1 + \xi(z;c))^{-3}\,F_{t_1}^\intercal\,F_{t_2}\,\tilde{F}_{t_2}^\intercal\ \left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\ \tilde{\Psi} \\
&\times \left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\ \tilde{F}_{t_2}\,\tilde{F}_{t_2}^\intercal\ \left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\ \tilde{F}_{t_1} \\
&\approx \frac{1}{P\,T}\,(1 + \xi(z;c))^{-3}\ \mathrm{tr}\left(\Psi F_{t_2}\tilde{F}_{t_2}^\intercal\ \left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\ \tilde{\Psi}\right. \\
&\left.\times \left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\ \tilde{F}_{t_2}\,\tilde{F}_{t_2}^\intercal\ \left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\right) \\
=& \frac{1}{P\,T}\,(1 + \xi(z;c))^{-3}\ \mathrm{tr}\left(\tilde{F}_{t_2}^\intercal\ \left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\ \tilde{\Psi}\ \left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\ \tilde{F}_{t_2}\right. \\
&\left.\times \tilde{F}_{t_2}^\intercal\ \left(z\,I + \hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\ \Psi\,F_{t_2}\right)\ .
\end{aligned}$$

The same argument as that in the calculation of $M_{t_1,t_2,2}$ leads to

$$\frac{1}{T^2}\sum_{t_1\neq t_2} M_{t_1,t_2,3} \to -\frac{c^{-1}\left(\xi(z;c)+z\,\frac{d}{dz}\xi(z;c)\right)\xi_\Psi(z;c)}{\left(1+\frac{d}{dz}\xi(z;c)\right)}$$

in L_2. Now we deal with the last term in (A.29). We have

$$\begin{aligned}
M_{t_1,t_2,4} =& P^{-1}\left(1+C_T\right)^{-1}\,\tilde{F}_{t_2}^{\intercal} \\
&\times \frac{1}{T}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\left(1+C_T\right)^{-1}\,\tilde{F}_{t_1}^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{\Psi} \\
&\times \frac{1}{T}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_2}\left(1+C_T\right)^{-1}\,\tilde{F}_{t_2}^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1} \\
&\times \tilde{F}_{t_1}\left(1+C_T\right)^{-1}F_{t_1}^{\intercal}F_{t_2} \\
=& \frac{1}{PT^2}\left(1+C_T\right)^{-4}\tilde{F}_{t_2}^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\tilde{F}_{t_1}^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\tilde{\Psi} \\
&\times \left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_2}\tilde{F}_{t_2}^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}F_{t_1}^{\intercal}F_{t_2}\,.
\end{aligned}$$

We will deal with $\tilde{F}_{t_2}^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\tilde{F}_{t_1}^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{\Psi}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_2}$ and $\tilde{F}_{t_2}^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,F_{t_1}^{\intercal}\,F_{t_2}$ separately. First, we have

$$\begin{aligned}
&\left\|\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,\tilde{F}_{t_1}^{\intercal}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{\Psi}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\right\|_F^2 \\
&= \mathrm{tr}\left(Q_{t_1,t_2}\,\tilde{F}_{t_1}\,\tilde{F}_{t_1}^{\intercal}\,Q_{t_1,t_2}\,\tilde{\Psi}\,Q_{t_1,t_2}\,Q_{t_1,t_2}\,\tilde{\Psi}\,Q_{t_1,t_2}\,\tilde{F}_{t_1}\,\tilde{F}_{t_1}^{\intercal}\,Q_{t_1,t_2}\right) \\
&= \tilde{F}_{t_1}^{\intercal}\,Q_{t_1,t_2}\,\tilde{\Psi}\,Q_{t_1,t_2}\,Q_{t_1,t_2}\,\tilde{\Psi}\,Q_{t_1,t_2}\,\tilde{F}_{t_1}\,\tilde{F}_{t_1}^{\intercal}\,Q_{t_1,t_2}\,Q_{t_1,t_2}\,\tilde{F}_{t_1}\,.
\end{aligned}$$

With the same arguments as those in (A.31), we have

$$\begin{aligned}
&\left\|\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,\tilde{F}_{t_1}^{\intercal}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{\Psi}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\right\|_F^2/(T^2\left(1+C_T\right)^2) \\
&\leq z^{-4}\,\left\|\tilde{\Psi}\right\|^2/T^2 \to 0\,.
\end{aligned}$$

Therefore, using Lemma 3 we have

$$\begin{aligned}
&\frac{1}{\sqrt{P}\,T}\,(1+C_T)^{-1}\,\tilde{F}_{t_2}^{\intercal}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,\tilde{F}_{t_1}^{\intercal}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{\Psi}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_2}\\
&\approx \frac{1}{\sqrt{P}\,T}\,(1+C_T)^{-1}\,\mathrm{tr}\left(\tilde{\Psi}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,\tilde{F}_{t_1}^{\intercal}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{\Psi}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\right)\\
&= \frac{1}{\sqrt{P}\,T}\,(1+C_T)^{-1}\,\mathrm{tr}\left(\tilde{F}_{t_1}^{\intercal}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{\Psi}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{\Psi}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\right)\\
&\approx \frac{1}{\sqrt{P}\,T}\,(1+\xi(z;c))^{-1}\,\mathrm{tr}\left(\left(\tilde{\Psi}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\right)^{3}\right)\\
&\leq c\,(1+\xi(z;c))^{-1}\,\left\|\tilde{\Psi}\right\|^{3}\,z^{-3}/\sqrt{P}\\
&\to 0\,.
\end{aligned}\tag{A.32}$$

Next, we deal with $\tilde{F}_{t_2}^{\intercal}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,F_{t_1}^{\intercal}\,F_{t_2}$. We know that

$$\begin{aligned}
\left\|\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,F_{t_1}^{\intercal}\right\|_F^2 &= \mathrm{tr}\left(Q_{t_1,t_2}\,\tilde{F}_{t_1}\,F_{t_1}^{\intercal}\,F_{t_1}\,\tilde{F}_{t_1}^{\intercal}\,Q_{t_1,t_2}\right)\\
&= \tilde{F}_{t_1}^{\intercal}\,Q_{t_1,t_2}^2\,\tilde{F}_{t_1}\,F_{t_1}^{\intercal}\,F_{t_1}\,.
\end{aligned}$$

Therefore, with the same arguments as those in (A.31), we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,F_{t_1}^{\intercal}\right\|_F^2/(T^2(1+C_T)^2)\right] &= \mathbb{E}\left[\tilde{F}_{t_1}^{\intercal}\,Q_{t_1,t_2}^2\,\tilde{F}_{t_1}\,F_{t_1}^{\intercal}\,F_{t_1}/(T^2\,(1+C_T)^2)\right]\\
&\leq z^{-2}\,\mathbb{E}[F_{t_1}^{\intercal}\,F_{t_1}/T^2]\\
&\leq ||\Psi||\,z^{-2}\,c/T\to 0\,.
\end{aligned}$$

Therefore, the same arguments as those in Lemma 11 lead to

$$\frac{1}{T}\,(1+C_T)^{-1}\,\tilde{F}_{t_2}^{\intercal}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,F_{t_1}^{\intercal}\,F_{t_2}\approx\frac{\mathrm{tr}\left(\Psi\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,F_{t_1}^{\intercal}\right)}{T\,(1+C_T)}\,.$$

Using Lemma 3 we have

$$\begin{aligned}\frac{1}{\sqrt{P}\,T}\,(1+C_T)^{-1}\,\tilde{F}_{t_2}^{\intercal}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,F_{t_1}^{\intercal}\,F_{t_2} &\approx \frac{\operatorname{tr}\left(\Psi^2\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\right)}{T\,(1+\xi(z;c))}\\ &\leq c\,(1+\xi(z;c))^{-1}\,||\Psi||^2\,z^{-1}/\sqrt{P}\\ &\to 0\,. \qquad \text{(A.33)}\end{aligned}$$

Combining (A.32) and (A.33) we have

$$M_{t_1,t_2,4}\to 0$$

in L_2 because of Cauchy–Schwartz Inequality. Altogether, we have

$$\begin{aligned}\mathcal{L}(z;c) &= \lim \mathrm{E}\left[\hat{\pi}_t(z)^2\mid\hat{\beta}\right]\\ &= \lim \operatorname{tr}\left(\tilde{\Psi}\,\hat{\beta}\,\hat{\beta}^{\intercal}\right)\\ &= \lim(\text{Term 1})+\lim(\text{Term 2})\\ &= \lim(\text{Term 1})+\xi(z;c)+z\,\frac{d}{dz}\xi(z;c)\\ &= \lim(\text{Term 11})+\lim(\text{Term 12})+\xi(z;c)+z\,\frac{d}{dz}\xi(z;c)\\ &= b_*\psi_{*,1}\left(\xi(z;c)+z\,\frac{d}{dz}\xi(z;c)\right)+\lim(\text{Term 12})+\xi(z;c)+z\,\frac{d}{dz}\xi(z;c)\\ &= (1+b_*\psi_{*,1})\left(\xi(z;c)+z\,\frac{d}{dz}\xi(z;c)\right)+\lim\left(\frac{b_*}{T^2}\sum_{t_1\neq t_2}M_{t_1,t_2}\right)\\ &= (1+b_*\,\psi_{*,1})\,\left(\xi(z;c)+z\,\frac{d}{dz}\xi(z;c)\right)+\frac{b_*\,c^{-1}\,\eta_\Psi(z;c)}{(1+\xi(z;c))^2}\\ &\quad-\frac{2\,b_*\,c^{-1}\,\left(\xi(z;c)+z\,\frac{d}{dz}\xi(z;c)\right)\,\xi_\Psi(z;c)}{1+\xi(z;c)}\,. \qquad \text{(A.34)}\end{aligned}$$

Finally, we calculate $\mathcal{V}(z;c)=\lim \mathbb{E}\left[\left(\hat{\pi}_t(z)\,R_{t+1}\right)^2\mid\hat{\beta}\right]$. As

$$\hat{\beta}(z)=\left(z\,I+T^{-1}\,\sum_{\tau}\tilde{F}_{\tau=t-T}^{t-1}\,\tilde{F}_{\tau}^{\intercal}\right)^{-1}\,\frac{1}{T}\,\sum_{\tau=t-T}^{t-1}\tilde{F}_{\tau}\,R_{\tau+1}$$

and thus is independent from $\tilde{F}_t$ and ϵ_{t+1}. We have

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{\pi}_t(z)\, R_{t+1}\right)^2 \mid \hat{\beta}\right] &= \mathbb{E}\left[\left(\tilde{F}_t^\intercal \hat{\beta}\left(F_t^\intercal\, \beta + \epsilon_{t+1}\right)\right)^2 \mid \hat{\beta}\right] \\
&= \mathbb{E}\left[\left(\hat{\beta}^\intercal\, \tilde{F}_t\, F_t^\intercal\, \beta\right)^2 + \left(\hat{\beta}^\intercal\, \tilde{F}_t\, \epsilon_{t+1}\right)^2 \mid \hat{\beta}\right] \\
&= \hat{\beta}^\intercal\, \mathbb{E}\left[\tilde{F}_t\, F_t^\intercal\, \beta\, \beta^\intercal\, F_t\, \tilde{F}_t^\intercal\right] \hat{\beta} + \hat{\beta}^\intercal\, \mathbb{E}\left[\tilde{F}_t\, \tilde{F}_t^\intercal\right] \hat{\beta} \\
&= \hat{\beta}^\intercal\, \mathbb{E}\left[\tilde{F}_t\, F_t^\intercal\, \beta\, \beta^\intercal\, F_t\, \tilde{F}_t^\intercal\right] \hat{\beta} + \operatorname{tr}\left(\tilde{\Psi}\, \hat{\beta}\, \hat{\beta}^\intercal\right).
\end{aligned}$$

We know from (A.34) that

$$\lim \operatorname{tr}\left(\tilde{\Psi}\, \hat{\beta}\, \hat{\beta}^\intercal\right) = \mathcal{L}(z; c)\,.$$

Now we turn to $\hat{\beta}^\intercal\, \mathbb{E}\left[\tilde{F}_t\, F_t^\intercal\, \beta\, \beta^\intercal\, F_t\, \tilde{F}_t^\intercal\right] \hat{\beta}$. We will need the following three Lemmas, which are the same as Lemma IA15, IA17, and IA18 in KMZ, respectively.

Lemma 12. *(Lemma IA15 in KMZ) Consider the block matrix decomposition*

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

Let $F_{t_1}^\intercal = X_t^\intercal\, Q_1$ and $F_{t_2}^\intercal = X_t^\intercal\, Q_1$, where X_t satisfies the moments conditions in Lemma 3. Then, we have

$$\begin{aligned}
&\mathbb{E}\left[F_{t_2}\, F_{t_1}^\intercal\, Z\, F_{t_1}\, F_{t_1}^\intercal\right] \\
&= \tilde{\Psi}_{2,1}\left(Z + Z^\intercal\right) \tilde{\Psi}_{1,1} + \left(Q_2 \operatorname{diag}\left(\left(\mathbb{E}\left[X^4\right] - 3\right)\, Q_1\, Z\, Q_1\right)\, Q_2 + \operatorname{tr}\left(Z\, \tilde{\Psi}_{1,1}\right) \tilde{\Psi}_{2,1}\right).
\end{aligned}$$

Lemma 13. *(Lemma IA 17 in KMZ) Let $A \circ B = (A_{i,j}\, B_{i,j})_{i,j=1}^P$ denote the Hadamard product of two matrices A, B. Then, $||A \circ B|| \leq ||A||\, ||B||$.*

Lemma 14. *(Lemma IA18 in KMZ) Suppose that $\beta = P^{-1/2}\, \Sigma_\beta^{1/2}\, \tilde{\beta}$ where Σ_β is uniformly bounded, $\mathbb{E}\left[\tilde{\beta}\right] = 0$ and $\mathbb{E}\left[\tilde{\beta}\, \tilde{\beta}^\intercal\right] = I$ and all mixed third-order moments of $\tilde{\beta}$ are zero while $\mathbb{E}\left[\tilde{\beta}_j^4\right]$ are uniformly bounded. Then, for*

any sequence uniformly bounded sequences of matrices A_P, B_P, we have

$$\beta^\intercal A \operatorname{diag}\left(B \beta \beta^\intercal B\right) A \beta \to 0$$

in probability.

Proof. Replace A and B with $A \Sigma_\beta^{1/2}$ and $\Sigma_\beta^{1/2} B$, respectively, to work directly with $\tilde{\beta}$. We abuse notation and denote $A \Sigma_\beta^{1/2}$ and $\Sigma_\beta^{1/2} B$ as A and B. Then,

$$\begin{aligned}
&\tilde{\beta}^\intercal A \operatorname{diag}\left(\tilde{B} \beta \beta^\intercal B\right) A \beta = \sum_{i,j} \tilde{\beta}_i \tilde{\beta}_j \sum_k A_{i,k} A_{k,j} \left(B \tilde{\beta} \tilde{\beta}^\intercal B\right)_{k,k} \\
&= \sum_{i,j} \sum_{j_1,j_2} \tilde{\beta}_i \tilde{\beta}_j \tilde{\beta}_{j_1} \tilde{\beta}_{j_2} \sum_k A_{i,k} A_{k,j} B_{k,j_1} B_{j_2,k} \,.
\end{aligned}$$

Let

$$K = \max_j \mathbb{E}\left[\tilde{\beta}_j^4\right] \,,$$

we have

$$\begin{aligned}
&P^{-2} \mathbb{E}\left[\tilde{\beta}^\intercal A \operatorname{diag}\left(B \tilde{\beta} \tilde{\beta}^\intercal B\right) A \tilde{\beta}\right] \\
&= P^{-2} \mathbb{E}\left[\sum_{i,j} \sum_{j_1,j_2} \tilde{\beta}_i \tilde{\beta}_j \tilde{\beta}_{j_1} \tilde{\beta}_{j_2} \sum_k A_{i,k} A_{k,j} B_{k,j_1} B_{j_2,k}\right] \\
&\leq P^{-2} K \sum_i \sum_k A_{i,k}^2 B_{k,i}^2 + P^{-2} \sum_{i,j,k} A_{i,k} A_{k,i} B_{k,j} B_{j,k} \\
&\quad + P^{-2} \sum_{i,j,k} A_{i,k} B_{k,i} A_{k,j} B_{j,k} + P^{-2} \sum_{i,j,k} A_{i,k} B_{i,k} A_{k,j} B_{k,j} \\
&= P^{-2} K \operatorname{tr}\left((A \circ B)(A \circ B)^\intercal\right) + P^{-2} \operatorname{tr}\left((A A^\intercal) \circ (B B^\intercal)\right) \\
&\quad + P^{-2} \operatorname{tr}\left((A B^\intercal) \circ (A^\intercal B)\right) + P^{-2} \operatorname{tr}\left((A B) \circ (A^\intercal B^\intercal)\right) \,. \qquad \text{(A.35)}
\end{aligned}$$

All terms in the above expressions are $O(P)$ by Lemma 13 because A and B are uniformly bounded. The proof is complete because $\tilde{\beta}^\intercal A \operatorname{diag}\left(\tilde{B} \beta \beta^\intercal B\right) A \beta > 0$. □

Recall that $\tilde{\Psi}_1 = (\Psi^{1/2}, \Phi^{1/2})$ and $Z_t = (X_t^\intercal, Y_t^\intercal)^\intercal$ and let $\tilde{\Psi}_2 = (\Psi^{1/2}, 0)$,

then $\tilde{F}_t = \tilde{\Psi}_1 \, Z_t$ and $F_t = \tilde{\Psi}_2 Z_t$, and we have

$$\mathbb{E}\left[\tilde{F}_t \, F_t^\intercal \, \beta \, \beta^\intercal \, F_t \, \tilde{F}_t^\intercal\right] = \tilde{\Psi}_1 \mathbb{E}\left[Z_t \, Z_t^\intercal \, \tilde{\Psi}_2^\intercal \, \beta \, \beta^\intercal \, \tilde{\Psi}_2 \, Z_t \, Z_t^\intercal\right] \, \tilde{\Psi}_1^\intercal \, .$$

Let $\kappa = \mathbb{E}[x_i^4]$, apply Lemma 12 to $\mathbb{E}\left[Z_t \, Z_t^\intercal \, \tilde{\Psi}_2^\intercal \, \beta \, \beta^\intercal \, \tilde{\Psi}_2 \, Z_t \, Z_t^\intercal\right]$ and write

$$\begin{aligned}\mathbb{E}\left[\tilde{F}_t \, F_t^\intercal \, \beta \, \beta^\intercal \, F_t \, \tilde{F}_t^\intercal\right] &= \tilde{\Psi}_1 \, \left(2 \, \tilde{\Psi}_2^\intercal \, \beta \, \beta^\intercal \, \tilde{\Psi}_2 + \operatorname{diag}\left((\kappa - 3) \, \tilde{\Psi}_2^\intercal \, \beta \, \beta^\intercal \, \tilde{\Psi}_2\right) + \operatorname{tr}\left(\tilde{\Psi}_2^\intercal \, \beta \, \beta^\intercal \, \tilde{\Psi}_2\right)\right) \, \tilde{\Psi}_1^\intercal \\ &= 2 \, \Psi \, \beta \, \beta^\intercal \, \Psi + (\kappa - 3) \, \Psi^{1/2} \, \operatorname{diag}\left(\Psi^{1/2} \, \beta \, \beta^\intercal \, \Psi^{1/2}\right) \, \Psi^{1/2} + \beta^\intercal \, \Psi \, \beta \, \tilde{\Psi} \, .\end{aligned}$$

Therefore,

$$\begin{aligned}\hat{\beta}^\intercal \, \mathbb{E}\left[\tilde{F}_t \, F_t^\intercal \, \beta \, \beta^\intercal \, F_t \, \tilde{F}_t^\intercal\right] \hat{\beta} = 2 \, \hat{\beta}^\intercal \, \Psi \, \beta \, \beta^\intercal \, \Psi \, \hat{\beta} + \hat{\beta}^\intercal \, (\kappa - 3) \, \Psi^{1/2} \, \operatorname{diag}\left(\Psi^{1/2} \, \beta \, \beta^\intercal \, \Psi^{1/2}\right) \, \Psi^{1/2} \, \hat{\beta} \\ + \hat{\beta}^\intercal \, \beta^\intercal \, \Psi \, \beta \, \tilde{\Psi} \, \hat{\beta} \, .\end{aligned}$$

We know from Lemma 1 that

$$\beta^\intercal \, \Psi \, \beta \to b_* \, \psi_{*,1} \, .$$

Combining (A.23) and (A.34) we have

$$\lim 2 \, \hat{\beta}^\intercal \, \Psi \, \beta \, \beta^\intercal \, \Psi \, \hat{\beta} \to 2 \, \mathcal{E}(z; c; a)^2, \quad \lim \hat{\beta}^\intercal \, \beta^\intercal \, \Psi \, \beta \, \tilde{\Psi} \, \hat{\beta} \to b_* \, \psi_{*,1} \, \mathcal{L}(z; c; a) \, . \tag{A.36}$$

Now we deal with

$$\begin{aligned}\hat{\beta}^\intercal \, \Psi^{1/2} \, \operatorname{diag}\left(\Psi^{1/2} \, \beta \, \beta^\intercal \, \Psi^{1/2}\right) \Psi^{1/2} \hat{\beta} &= \operatorname{tr}\left(\Psi^{1/2} \operatorname{diag}\left(\Psi^{1/2} \, \beta \, \beta^\intercal \, \Psi^{1/2}\right) \, \Psi^{1/2} \, \hat{\beta} \, \hat{\beta}^\intercal\right) \\ &= \operatorname{tr}\left(\Psi^{1/2} \operatorname{diag}\left(\Psi^{1/2} \, \beta \, \beta^\intercal \, \Psi^{1/2}\right) \, \Psi^{1/2} \, \hat{\beta} \, \hat{\beta}^\intercal\right) \\ &= \operatorname{tr}\Bigg(\Psi^{1/2} \operatorname{diag}\left(\Psi^{1/2} \, \beta \, \beta^\intercal \, \Psi^{1/2}\right) \Psi^{1/2} \left(z \, I + \hat{\tilde{\Psi}}_T\right)^{-1} \\ &\quad \times \left(\hat{\Psi}_{\tilde{F},F} \, \beta \, \beta^\intercal \, \hat{\Psi}_{\tilde{F},F}^\intercal + q_T \, \beta^\intercal \hat{\Psi}_{\tilde{F},F}^\intercal + \hat{\Psi}_{\tilde{F},F} \, \beta \, q_T + q_T \, q_T^\intercal\right) \\ &\quad \times \left(z \, I + \hat{\tilde{\Psi}}_T\right)^{-1}\Bigg) \, .\end{aligned}$$

Next, we show that

$$\hat{\beta}^\intercal \, \Psi^{1/2} \, \operatorname{diag}\left(\Psi^{1/2} \, \beta \, \beta^\intercal \, \Psi^{1/2}\right) \, \Psi^{1/2} \, \hat{\beta} \to 0 \tag{A.37}$$

in probability. Note that using Cauchy–Schwartz Inequality we only need to

prove that

$$\mathrm{tr}\left(\Psi^{1/2}\,\mathrm{diag}\left(\Psi^{1/2}\,\beta\,\beta^{\intercal}\,\Psi^{1/2}\right)\Psi^{1/2}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\,\hat{\Psi}_{\tilde{F},F}\,\beta\,\beta^{\intercal}\,\hat{\Psi}_{\tilde{F},F}^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right)\to 0 \tag{A.38}$$

in probability and that

$$\mathrm{tr}\left(\Psi^{1/2}\,\mathrm{diag}\left(\Psi^{1/2}\,\beta\,\beta^{\intercal}\,\Psi^{1/2}\right)\Psi^{1/2}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\,q_T\,q_T^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right)\to 0 \tag{A.39}$$

in probability. To prove (A.38), note that from (A.20) we have

$$\mathbb{P}\left(\left\|\tilde{\Psi}_{\tilde{F},F}\right\|\geq C\,(\log P)^3\right)\leq\frac{K}{(\log P)^2}+\delta_1$$

for any $\delta_1>0$. Therefore, using (A.35) in Lemma 14 we have that

$$\mathbb{E}\left[\mathrm{tr}\left(\Psi^{1/2}\,\mathrm{diag}\left(\Psi^{1/2}\,\beta\,\beta^{\intercal}\,\Psi^{1/2}\right)\Psi^{1/2}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\,\hat{\Psi}_{\tilde{F},F}\,\beta\,\beta^{\intercal}\,\hat{\Psi}_{\tilde{F},F}^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right)\right]$$
$$\leq C\log^6 P/P\to 0$$

with probability at least δ_1. Let $\delta_1\to 0$ we have that proved (A.38). For (A.39), we first consider $\mathrm{diag}\left(\Psi^{1/2}\,\beta\,\beta^{\intercal}\,\Psi^{1/2}\right)$, let $\Psi^{1/2}=(a_1,...,a_P)$, as $\Psi^{1/2}$ is symmetric we have

$$(\Psi^{1/2}\,\beta\,\beta^{\intercal}\,\Psi^{1/2})_{k,k}=a_k^{\intercal}\,\beta\,\beta^{\intercal}a_k\,.$$

Therefore, because $\mathbb{E}[\beta_i^2]=b_*/P$, $\mathbb{E}[\beta_i^4]\leq K/P^2$ and β_i,β_j are independent, we have

$$\mathbb{E}\left(\Psi^{1/2}\,\beta\,\beta^{\intercal}\,\Psi^{1/2}\right)_{k,k}=\mathbb{E}[a_k^{\intercal}\,\beta\,\beta^{\intercal}\,a_k]=\mathbb{E}\left[\sum_{i,j}a_{k,i}\beta_i\,\beta_j a_{k,j}\right]=b_*\,a_k^{\intercal}\,a_k/P=b_*\,\Psi_{k,k}/P$$

and

$$\begin{aligned}
\mathbb{E}\left(\Psi^{1/2}\,\beta\,\beta^\intercal\,\Psi^{1/2}\right)^2_{k,k} &= \mathbb{E}\left(\sum_{i,j} a_{k,i}\,\beta_i\,\beta_j\,a_{k,j}\right)^2 \\
&= \mathbb{E}\left[\sum_{i,j,l,m} a_{k,i}\,\beta_i\,\beta_j\,a_{k,j}\,a_{k,l}\,\beta_l\,\beta_m\,a_{k,m}\right] \\
&\leq K\sum_i a_{k,i}^4/P^2 + 6\,b_*^2\sum_{i\neq j} a_{k,i}^2\,a_{k,j}^2/P^2 \\
&\leq (6\,b_*^2+K)\,(a_k^\intercal\,a_k)^2/P^2 \\
&= (6\,b_*^2+K)\,\Psi_{k,k}^2/P^2\,.
\end{aligned}$$

Let $\tilde{m} = 1/\log P$, then we have

$$\begin{aligned}
\mathbb{P}\left(\exists k, |(\Psi^{1/2}\,\beta\,\beta^\intercal\,\Psi^{1/2})_{k,k} - b_*\Psi_{i,i}/P| \geq \tilde{m}\right) &\leq \sum_k \mathbb{P}\left(|(\Psi^{1/2}\,\beta\,\beta^\intercal\,\Psi^{1/2})_{k,k} - b_*\Psi_{i,i}/P| \geq \tilde{m}\right) \\
&\leq \sum_i \frac{\mathrm{Var}\left[(\Psi^{1/2}\,\beta\,\beta^\intercal\,\Psi^{1/2})_{k,k}\right]}{\tilde{m}^2} \\
&\leq \frac{(5\,b_*+K)\,||\Psi||^2\,\log^2 P}{P}\,.
\end{aligned}$$

Therefore, we have

$$\mathbb{P}\left(\left\|\mathrm{diag}\left(\Psi^{1/2}\,\beta\,\beta^\intercal\,\Psi^{1/2}\right)\right\| \geq \tilde{m}\right) \leq \frac{(5b_*+K)||\Psi||^2\log^2 P}{P} \to 0\,.$$

Therefore, using Lemma 11 we have, under the event that $\left\|\mathrm{diag}\left(\Psi^{1/2}\,\beta\,\beta^\intercal\,\Psi^{1/2}\right)\right\| \leq \tilde{m}$,

$$\begin{aligned}
&\mathrm{tr}\left(\Psi^{1/2}\,\mathrm{diag}\left(\Psi^{1/2}\,\beta\,\beta^\intercal\,\Psi^{1/2}\right)\Psi^{1/2}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1} q_T q_T^\intercal \left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right) \\
&\leq \tilde{m}\,\mathrm{tr}\left(\Psi\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1} q_T q_T^\intercal \left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right) \\
&\approx \tilde{m}\,\mathrm{tr}\left(\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\Psi\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\hat{\tilde{\Psi}}_T\right) \\
&\to 0\,.
\end{aligned}$$

Therefore, (A.39) is proven. Altogether, combining (A.36) and (A.37) we

have

$$\begin{aligned}\lim \mathbb{E}\left[(\hat{\pi}_t(z)R_{t+1})^2 \mid \hat{\beta}\right] &= \lim \hat{\beta}^\intercal \, \mathbb{E}\left[\tilde{F}_t \, F_t^\intercal \, \beta \, \beta^\intercal \, F_t \, \tilde{F}_t^\intercal\right] \hat{\beta} + \lim \operatorname{tr}\left(\tilde{\Psi}\hat{\beta}\hat{\beta}^\intercal\right) \\ &= 2\,\mathcal{E}(z;c;a)^2 + (1 + b_* \, \psi_{*,1})\,\mathcal{L}(z;c;a)\,.\end{aligned}$$

Theorem 1 is proven. □

A.1.3 Proof of Proposition 2.

In this section we prove Proposition 2.

Proof. First we prove point (i). In the special case we can explicitly write

$$\mathcal{E}(z;c;a) = \frac{b_* \left(-\sqrt{(a\,c+a+c+z+1)^2 - 4\,(a+1)^2\,c} + a\,c + a + c + z + 1\right)}{2\,(a+1)^2\,c}\,.$$

Next, as $a \geq 0$ and $c \geq 0$, we have

$$\mathcal{E}(z;c;a) \geq 0\,.$$

Taking the derivative of $\mathcal{E}(z;c;a)$ with respect to a, we have

$$\begin{aligned}\frac{\partial}{\partial a}\mathcal{E}(z;c;a) = \big(-b_* \, \big(&-a^2\,c^2 + 2\,a^2\,c - a^2 - 2\,a\,c^2 \\ &- 3\,a\,c\,z + 4\,a\,c - 3\,a\,z - 2\,a - c^2 - 3\,c\,z + 2\,c - 2\,z^2 - 3z - 1 \\ &+ (a\,c + a + c + 2\,z + 1)\sqrt{2\,(a+1)\,(c+1)\,z + (a+1)^2\,(c-1)^2 + z^2}\big)\big) \\ &/ \left(2\,(a+1)^3\,c\,\sqrt{(a\,c+a+c+z+1)^2 - 4\,(a+1)^2\,c}\right) \\ =(-b_*\,(A_1 + B_1))&/ \left(2\,(a+1)^3\,c\,\sqrt{(a\,c+a+c+z+1)^2 - 4\,(a+1)^2 c}\right)\,.\end{aligned}$$

We know that the denominator is greater than 0, on the nominator, we have $B_1 \geq 0$ and

$$A_1^2 - B_1^2 = 4\,(a+1)^3\,c\,\left(-\left((a+1)\,(c-1)^2\right) - 2\,(c+1)\,z\right) < 0\,.$$

Therefore, we have $A_1 + B_1 \geq 0$ and $\frac{\partial}{\partial a}\mathcal{E}(z;c;a) < 0$, thus $\frac{\partial}{\partial a}\mathcal{E}(z;c;a)$ is monotonically decreasing with respect to a. For the convexity, we take second

derivative of $\frac{\partial}{\partial a}\mathcal{E}(z;c;a)$ with respect to a.

$$\begin{aligned}\frac{\partial^2}{\partial a^2}\mathcal{E}(z;c;a) = b_* \Big(4 \Big(&-6\,(a+1)^2 \left(2\,c^2+c+2\right) z^2 - 10\,(a+1)\,(c+1)\,z^3 \\ &-6\,(a+1)^3\,(c+1)\,(c-1)^2\,z - (a+1)^4\,(c-1)^4 - 3\,z^4\Big) \\ &+4\,\Big(7\,(a+1)\,(c+1)\,z^2 + (a+1)^2\,(c\,(5\,c-2)+5)\,z \\ &+(a+1)^3\,(c-1)^2\,(c+1) + 3\,z^3\Big)\,\sqrt{(a\,c+a+c+z+1)^2 - 4\,(a+1)^2\,c}\Big) \\ &/4\,(a+1)^4\,c\,\left((a\,c+a+c+z+1)^2 - 4\,(a+1)^2\,c\right)^{3/2} \\ = b_*\,&(A_2+B_2)/4\,(a+1)^4\,c\,\left((a\,c+a+c+z+1)^2 - 4\,(a+1)^2\,c\right)^{3/2}\,.\end{aligned}$$

We know that the denominator is greater than 0, on the nominator, we have $B_2 \geq 0$ because $c\,(5\,c-2)+5 \geq 0$. Also, we have

$$\begin{aligned}A_2^2 - B_2^2 = -\,&64\,(a+1)^4\,c\,\Big(12\,(a+1)^2\,\left(c^2-1\right)^2\,z^2 + 2\,(a+1)\,(c+1)\,(c\,(5\,c+6)+5)\,z^3 \\ &+6\,(a+1)^3\,(c+1)\,(c-1)^4\,z + (a+1)^4\,(c-1)^6 + 3\,(c\,(c+3)+1)\,z^4\Big) \leq 0\,.\end{aligned}$$

Therefore, we have $A_2 + B_2 \geq 0$ and hence $\frac{\partial^2}{\partial a^2}\mathcal{E}(z;c;a) \geq 0$, the convexity of $\mathcal{E}(z;c;a)$ is proven.

Next we prove point (ii). Recall that as $\mathcal{E}(z;c;a) \geq 0$, we can write the R^2 and Sharpe Ratio as follows:

$$R^2(z;c;a) = \frac{2\,\mathcal{E}(z;c;a) - \mathcal{L}(z;c;a)}{1+b_*}\,,$$

$$SR(z;c;a) = \frac{1}{\sqrt{2 + (1+b_*)\,\mathcal{L}(z;c;a)/\mathcal{E}(z;c;a)^2}}\,.$$

Therefore, for monotonicity and maximum value of R^2 and SR we only need to consider

$$\tilde{R}^2(z;c;a) = 2\,\mathcal{E}(z;c;a) - \mathcal{L}(z;c;a) \tag{A.40}$$

and

$$\widetilde{SR}(z;c;a) = \mathcal{L}(z;c;a)/\mathcal{E}(z;c;a)^2\,. \tag{A.41}$$

In the following context, to simplify notation we let

$$C(z;c;a) = \sqrt{2\,(a+1)\,(c+1)\,z + (a+1)^2\,(c-1)^2 + z^2}\,.$$

First we consider (A.40), we have

$$\begin{aligned}\frac{\partial}{\partial z}\tilde{R}^2(z;c;a) &= 8\left((a+1)\,c\,(ab_*+a+1)-b_*\,z\right)\\ &\quad\times\left(3\,(a+1)\,c\,z^2+3\,(a+1)^2\,c\,(c-1)\,z+(a+1)^3\,(c-1)^3+z^3\right.\\ &\quad\left.+C(z;c;a)\,\left((a+1)\,(2\,c-1)z+(a+1)^2\,(c-1)^2+z^2\right)\right)\\ &\quad/\left((b_*+1)\left((a\,c+a+c+z+1)^2-4\,(a+1)^2\,c\right)^{3/2}\right.\\ &\quad\left.\times\,\left(C(z;c;a)+a\,(c-1)+c+z-1\right)^3\right)\\ &= 8\left((a+1)\,c\,(a\,b_*+a+1)-b_*\,z\right)(A_3+B_3)\\ &\quad/\left((b_*+1)\,\left((a\,c+a+c+z+1)^2-4\,(a+1)^2\,c\right)^{3/2}\right.\\ &\quad\left.\times\,\left(C(z;c;a)+a\,(c-1)+c+z-1\right)^3\right)\,.\end{aligned}$$

We know that

$$2\,(a+1)\,(c+1)\,z+(a+1)^2\,(c-1)^2+z^2\geq\left|(a+1)\,(c-1)+z\right|.$$

Therefore the denominator is greater than 0. Also, as $a,c,z\geq 0$, we have $B_3\geq 0$ and

$$A_3^2-B_3^2=-4\,(1+a)^3\,z^3\leq 0\,.$$

Therefore, $A_3+B_3\geq 0$, when $z<z_*=c\,(1+a)\,(1+a+a\,b_*)/b_*$ we have $\frac{\partial}{\partial z}R^2(z;c;a)\geq 0$ and when $z>z_*$ we have $\frac{\partial}{\partial z}R^2(z;c;a)\leq 0$, and $z=z_*$ is the maximum point of $R^2(z;c;a)$. Next we consider $\widetilde{SR}(z;c;a)$. We have

$$\begin{aligned}\frac{\partial}{\partial z}\widetilde{SR}^2(z;c;a) &= -8\,(a+1)^2\,c\left((a+1)\,c\,(a\,b_*+a+1)-b_*\,z\right)\\ &\quad\times\left(3\,(a+1)^2\,(c-1)\,z+(a+1)^3\,(c-1)^3-3\,(a+1)\,z^2-z^3\right.\\ &\quad\left.+\left(-(a+1)\,(c-2)\,z+(a+1)^2\,(c-1)^2+z^2\right)C(z;c;a)\right)\\ &\quad/\left(b_*^2\,C(z;c;a)^3\,(C(z;c;a)+a\,c-a+c-z-1)^3\right)\\ &= -8\,(a+1)^2\,c\left((a+1)\,c\,(a\,b_*+a+1)-b_*\,z\right)(A_5+B_5)\\ &\quad/\left(b_*^2\,C(z;c;a)^3\,\left(C(z;c;a)+a\,c-a+c-z-1\right)^3\right)\,.\end{aligned}$$

We know that the denominator is greater than 0 and $B_5\geq 0$, also,

$$A_5^2-B_5^2=-4\,(a+1)^3\,c^3\,z^3\leq 0\,.$$

Therefore, because

$$\frac{\partial}{\partial z} SR(z;c;a) = -\frac{1}{2}\, SR(z;c;a)^3\, \frac{\partial}{\partial z} \widetilde{SR}(z;c;a)$$

we have when $z < z_* = c\,(1+a)(1+a+a\,b_*)/b_*$ we have $\frac{\partial}{\partial z} SR(z;c;a) \geq 0$ and when $z > z_*$ we have $\frac{\partial}{\partial z} SR(z;c;a) \leq 0$, and $z = z_*$ is the maximum point of $SR(z;c;a)$. Now we turn the monotonicity and convexity of $R^2(z_*;c;a)$ and $SR(z_*;c;a)$. First, we shall note that, by direct calculation, we have

$$\mathcal{E}(z_*;c;a) = L(z_*;c;a)\,.$$

Therefore, we only need to consider $\mathcal{E}(z_*;c;a)$. First let's consider $\tilde{R}^2(z_*;c;a)$, we have

$$\begin{aligned}\frac{\partial}{\partial a}\tilde{R}^2(z_*;c;a) &= \frac{\partial}{\partial a}\tilde{R}^2(z;c;a)|_{z=z_*} + \frac{\partial}{\partial z}\tilde{R}^2(z;c;a)|_{z=z_*}\frac{\partial}{\partial a}(z_*) \\ &= \frac{\partial}{\partial a}\mathcal{E}(z;c;a)|_{z=z_*} + \frac{\partial}{\partial z}\mathcal{E}(z;c;a)|_{z=z_*}\frac{\partial}{\partial a}(z_*)\,.\end{aligned}$$

From (i) we know that $\frac{\partial}{\partial a}\mathcal{E}(z;c;a) \leq 0$, also, we have

$$\frac{\partial}{\partial z}\mathcal{E}(z;c;a) = \frac{b_*\left(1 - \frac{a\,c+a+c+z+1}{\sqrt{(a\,c+a+c+z+1)^2 - 4\,(a+1)^2\,c}}\right)}{2\,(a+1)^2\,c} \leq 0$$

and

$$\frac{\partial}{\partial a} z_* = c\,(a\,b_* + a + 1)/b_* + (a+1)\,c\,(b_*+1)/b_* \geq 0$$

therefore, we have

$$\frac{\partial}{\partial a} R^2(z_*;c;a) = \frac{\partial}{\partial a}\tilde{R}^2(z_*;c;a)/(1+b_*) = \frac{\partial}{\partial a}\mathcal{E}(z_*;c;a)/(1+b_*) \leq 0\,. \quad \text{(A.42)}$$

For the convexity, let's first consider

$$\frac{1}{\widetilde{SR}(z;c;a)} = \mathcal{E}(z;c;a)^2/\mathcal{L}(z;c;a)\,.$$

We have,

$$\frac{\partial^2}{\partial a^2}\left(\frac{1}{\widetilde{SR}(z;c;a)}\right) = \frac{2\left(\frac{\partial}{\partial a}\widetilde{SR}(z;c;a)\right)^2 - \widetilde{SR}(z;c;a)\frac{\partial^2}{\partial a^2}\widetilde{SR}(z;c;a)}{\widetilde{SR}(z;c;a)^3}\,.$$

Now, we have

$$2\left(\frac{\partial}{\partial a}\widetilde{SR}(z;c;a)\right)^2-\widetilde{SR}(z;c;a)\,\frac{\partial^2}{\partial a^2}\widetilde{SR}(z;c;a)=A_6+B_6\,C(z;c;a)\,,$$

where

$$\begin{aligned}A_6=&2\left(3\,(a+1)^2\,(b_*+1)^2\,c^2+3\,(a+1)\,(b_*+1)\,b_*\,c+b_*^2\right)\\&\times\left(2\,(a+1)\,(c+1)\,z+(a+1)^2\,(c-1)^2+z^2\right)^3\\&+A_{60}+A_{61}\,z+A_{62}\,z^2+A_{63}\,z^3+A_{64}\,z^4+A_{65}\,z^6\,,\end{aligned}$$

where

$$\begin{aligned}A_{60}=&2\,(a+1)^6\,(c-1)^4\,\left(b_*^2\,\left(3\,c\,\left(\left(2\,a^2+a-1\right)\,c^2+\left(a^2+2\right)\,c+(a+1)^2\,c^3+a-1\right)+1\right)\right.\\&\left.+3\,(a+1)\,b_*\,c\,(c+1)\,(c\,(2\,a\,(c+1)+2c-1)+1)+3\,(a+1)^2\,c^2\,(c+1)^2\right)>0\,,\\A_{61}=&4\,(a+1)^5\,(c-1)^2\,\left(b_*^2\,\left(9\,(a+1)^2\,c^5+(a+1)\,(21\,a-8)\,c^4+(7\,a\,(3\,a+1)+13)\,c^3\right.\right.\\&\left.+(a\,(9\,a+5)+5)\,c^2+(9\,a-6)\,c+3\right)+(a+1)\,b_*\,c\,(c\,(6\,a\,(c+1)\,(c\,(3\,c+4)+3)\\&+c\,(c\,(18\,c+13)+7)+5)+9)+3\,(a+1)^2\,c^2\,(c+1)\,(c\,(3c+4)+3))\Big)>0\,,\\A_{62}=&2\,(1+a)^4\,\left(b_*^2\,\left(\left(58\,a^2+44\,a+85\right)\,c^4+45\,(a+1)^2\,c^6+3\,(a+1)\,(22\,a-23)\,c^5\right.\right.\\&\left.+6\,(a\,(11\,a+9)-3)\,c^3+3\,(a\,(15\,a-2)+13)\,c^2+(45\,a-33)\,c+15\right)+b_*\,c\,(1+a)\\&\times\left((90\,a+90)\,c^5+(132\,a-3)\,c^4+(116\,a+44)\,c^3+(132\,a+54)\,c^2\right.\\&\left.\left.+(90\,a-6)\,c+45)\right)>0\,,\\A_{63}=&8\,(a+1)^3\,\left(3\,(a+1)^2\,c^2\,(c+1)\,(c\,(5\,c+4)+5)\right.\\&+b_*^2\,\left(15\,(a+1)^2\,c^5+9\,(a+1)\,(3\,a-1)\,c^4\right.\\&\left.+(9\,a\,(3\,a+2)+7)\,c^3+(3\,a\,(5\,a+3)+2)\,c^2+(15\,a-4)\,c+5\right)+3\,(a+1)\,b_*\,c\\&\times(c\,(\,2a\,(c+1)\,(c\,(5\,c+4)+5)+2\,c\,(c\,(5\,c+3)+3)+3)+5))>0\,,\\A_{64}=&2\left((a+1)^2\,b_*^2\,\left(c\,\left(45\,(a+1)^2\,c^3+(a+1)\,(72\,a+5)\,c^2\right.\right.\right.\\&\left.+(a\,(4\,5a+52)+10)\,c+45\,a+9)+15)+(a+1)^3\,b_*\,c\,(c(18\,a\,(c\,(5\,c+8)+5)\right.\\&\left.\left.+c\,(90\,c+77)+52)+45)+9\,(a+1)^4\,c^2\,(c\,(5\,c+8)+5)\right)>0\,,\\A_{65}=&2\,\left(6\,(a+1)\,b_*^2\,\left(c\,\left(3\,(a+1)^2\,c^2+(a+1)\,(3\,a+2)\,c+3\,a+2\right)+1\right)\right.\\&\left.+6\,(a+1)^2\,b_*\,c\,(c\,(6\,a\,(c+1)+6\,c+5)+3)+18\,(a+1)^3\,c^2(c+1)\right)>0\,,\\A_{66}=&2\left(b_*^2\,(3\,(a+1)\,c\,(a\,c+c+1)+1)+3\,(a+1)\,b_*\,c\,(2\,(a+1)c+1)+3\,(a+1)^2\,c^2\right)>0\,.\end{aligned}$$

Therefore, we have $A_6>0$, for B_6, we have

$$B_6=B_{60}+B_{61}\,z+B_{62}\,z^2+B_{63}\,z^3+B_{64}\,z^4+B_{65}\,z^5\,,$$

where

$$
\begin{aligned}
B_{60} &= (a+1)^5 (c-1)^4 \left(3 (a+1)^2 (b_*+1)^2 c^3 \right. \\
&\quad \left. +3 (a+1) (b_*+1) c^2 (ab_*+a+1) + 3 b_* c (ab_*+a+1) + b_*^2\right) > 0 \,, \\
B_{61} &= 3 (1+a)^6 (c-1)^2 c^2 (5 + c (8+5c)) + (a+1)^4 b_*^2 (c-1)^2 \\
&\quad \times \left(15 (a+1)^2 c^4 + (a+1) (24a+7) c^3 + (a (15a+26) - 4) c^2 + (15a+5) c + 5\right) \\
&\quad + (a+1)^5 b_* (c-1)^2 c (c(6a(c(5c+8)+5) + c(30c+31) + 26) + 15) > 0 \,, \\
B_{62} &= 3 (a+1)^5 c^2 (c+1) (c (10c-1) + 10) + 2 (a+1)^4 b_* c \\
&\quad \times \left(\left(30 (a+1) c^3 + (27a+13) c^2 + 27 (a+1) c + 30a + 17\right) + 15\right) + (a+1)^3 b_*^2 \\
&\quad \times \left(c \left(\left(30a^2 + 34a - 1\right) c + 30 (a+1)^2 c^4 + (a+1) (27a-1) c^3 \right.\right. \\
&\quad \left.\left. +3 (9a (a+2) + 1) c^2 + 30a + 7\right) + 10\right) > 0 \,, \\
B_{63} &= 3 (a+1)^4 c^2 (c (10c+13) + 10) + 2 (a+1)^3 b_* c \left(c \left(30 (a+1) c^2 + (39a+32) c \right.\right. \\
&\quad \left.\left. +30a + 28) + 15\right) + (a+1)^2 b_*^2 \left(c \left(30 (a+1)^2 c^3 + (a+1) (39a+25) c^2 \right.\right.\right. \\
&\quad \left.\left.\left. +2 (a (15a+28) + 6) c + 30a + 19) + 10\right) > 0 \,, \right. \\
B_{64} &= 15 (a+1)^3 c^2 (c+1) \\
&\quad + (a+1) b_*^2 \left(c \left(15 (a+1)^2 c^2 + 3 (a+1) (5a+6) c + 15a + 14\right) + 5\right) \\
&\quad + 3 (a+1)^2 b_* c (c (10a (c+1) + 10c + 11) + 5) > 0 \,, \\
B_{65} &= b_*^2 (3 (a+1) c (ac + c + 1) + 1) + 3 (a+1) b_* c (2 (a+1) c + 1) + 3 (a+1)^2 c^2 > 0 \,.
\end{aligned}
$$

Therefore, we have $B_6 > 0$ and altogether, we have

$$
\frac{\partial^2}{\partial a^2} \left(\frac{1}{\widetilde{SR}(z;c;a)} \right) > 0
$$

and therefore $1/\widetilde{SR}(z;c;a)$ is convex in a. As z_* is the maximum value of $\frac{1}{\widetilde{SR}(z;c;a)}$ with respect to z, using Danskin's theorem (see, e.g., [8]) we have that

$$
\frac{1}{\widetilde{SR}(z_*;c;a)} = \mathcal{E}(z_*;c;a)
$$

is convex in a. Therefore, we proved that $R^2(z_*;c;a)$ is monotonically decreasing and convex in a. Next, consider $SR(z_*;c;a)$. We have

$$
\begin{aligned}
\frac{\partial}{\partial a} SR(z_*;c;a) &= \frac{\partial}{\partial a} \frac{1}{\sqrt{2 + (1+b_*) / \mathcal{E}(z_*;c;a)}} \\
&= \frac{(1+b_*)}{2} SR(z_*;c;a)^3 \frac{\frac{\partial}{\partial a} \mathcal{E}(z_*;c;a)}{\mathcal{E}(z_*;c;a)^2} \,.
\end{aligned}
$$

From (A.42) we know that

$$\frac{\partial}{\partial a} SR(z_*; c; a) \leq 0\,.$$

For the convexity, we have

$$\begin{aligned}\frac{\partial^2}{\partial a^2} SR(z_*; c; a) &= \frac{\partial^2}{\partial a^2} \frac{1}{\sqrt{2 + (1 + b_*)\,/\mathcal{E}(z_*; c; a)}} \\ &= 2\, SR(z_*; c; a)^5\, (b_* + 1)^2\, \left(A_7 + B_7\, C(z_*; c; a)\right)\,,\end{aligned}$$

where

$$A_7 = A_{70} + A_{71}\, b_* + A_{72}\, b_*^2 + A_{73}\, b_*^3 + A_{74}\, b_*^4 + A_{75}\, b_*^5\,,$$

with

$$\begin{aligned}
A_{70} &= 8\,(a+1)^5\, c^5 > 0\,, \\
A_{71} &= 40\,(a+1)^5\, c^5 + 32\, a\,(a+1)^3\, c^4 + 32\,(a+1)^3\, c^4 > 0\,, \\
A_{72} &= 51\, a^2\,(a+1)\, c^3 + 80\,(a+1)^5\, c^5 + 128\, a\,(a+1)^3\, c^4 \\
&\quad + 76\,(a+1)^3\, c^4 + 102\, a\,(a+1)\, c^3 + 51\,(a+1)\, c^3 > 0\,, \\
A_{74} &= 153\, a^2\,(a+1)\, c^3 + 41\, a^2\, c^2 + 80\,(a+1)^5\, c^5 + 192\, a\,(a+1)^3\, c^4 \\
&\quad + 36\,(a+1)^3\, c^4 + 178\, a\,(a+1)\, c^3 + 25\,(a+1)\, c^3 + 82\, a\, c^2 + 41\, c^2 > 0\,, \\
A_{75} &= 40\, a^5\, c^5 + a^4\, \left(200\, c^5 + 128\, c^4\right) + a^3\, \left(400\, c^5 + 356\, c^4 + 153\, c^3\right) \\
&\quad + a^2\, \left(400\, c^5 + 300\, c^4 + 203\, c^3 + 82\, c^2\right) + a\, \left(200\, c^5 + 44\, c^4 + 59\, c^3 + 56\, c^2 + 17\, c\right) \\
&\quad + 40\, c^5 - 28\, c^4 + 9\, c^3 - 26\, c^2 + 17\, c > 0\,, \\
A_{76} &= 8\, a^5\, c^5 + a^4\, \left(40\, c^5 + 32\, c^4\right) + a^3\, \left(80\, c^5 + 76\, c^4 + 51\, c^3\right) \\
&\quad + a^2\, \left(80\, c^5 + 36\, c^4 + 25\, c^3 + 41\, c^2\right) + a\, \left(40\, c^5 - 28\, c^4 + 9\, c^3 - 26\, c^2 + 17\, c\right) \\
&\quad + 8\, c^5 - 20\, c^4 + 35\, c^3 + 13\, c^2 - 15\, c + 3 > 0\,.
\end{aligned}$$

For B_7, we have

$$B_7 = b_*\, \left(B_{70} + B_{71}\, b_* + B_{72}\, b_*^2 + B_{73}\, b_*^3 + B_{74}\, b_*^4\right)\,,$$

where

$$
\begin{aligned}
B_{70} &= 8\,(a+1)^4\,c^4 > 0\,,\\
B_{71} &= 32\,(a+1)^4\,c^4 + 24\,a\,(a+1)^2\,c^3 + 24\,(a+1)^2\,c^3 > 0\,,\\
B_{72} &= 27\,a^2\,c^2 + 48\,(a+1)^4\,c^4 + 72\,a\,(a+1)^2\,c^3 + 36\,(a+1)^2\,c^3 + 54\,a\,c^2 + 27\,c^2 > 0\,,\\
B_{73} &= 14\,c + 14\,a\,c - 6\,c^2 + 48\,a\,c^2 + 54\,a^2\,c^2 + 72\,a\,(1+a)^2\,c^3 + 32\,(1+a)^4\,c^4 > 0\,,\\
B_{74} &= 8\,a^4\,c^4 + a^3\,\left(32\,c^4 + 24\,c^3\right) + a^2\,\left(48\,c^4 + 36\,c^3 + 27\,c^2\right) + a\,\left(32\,c^4 - 6\,c^2 + 14\,c\right)\\
&\quad + 8\,c^4 - 12\,c^3 + 23\,c^2 - 12\,c + 3 > 0\,.
\end{aligned}
$$

Altogether, we proved that $SR(z_*; c; a)$ is monotonically decreasing and convex in a. For the ridgeless case when $z \to 0$, we only need to use L'Hôpital's rule to obtain the results.

□

B | The Limited Virtue of Complexity in a Noisy World

B.1 Proofs

B.1.1 Proof of Corollary 2.

Throughout the proof, all convergences are in probability unless otherwise stated. In addition, the limits are taken as $T, P, P_1 \to \infty$, $P/T \to c$ and $P_1/P = q$.

Proof. In this section we proof Corollary 2, note that

$$\lim_{T\to\infty} T^{-1} \operatorname{tr}\left((z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1})^{-1}\, \tilde{\Psi}_{1,1}\right) = \xi(z; c\, q; a\, q)$$

because of Proposition 1 and $\tilde{\Psi}_{1,1} = (1 + a\, q) I_{P_1}$. Write $\beta_1 = \{\beta_i\}_{i=1}^{P_1}$ and $\beta_2 = \{\beta_i\}_{i=P_1+1}^{P}$, now we calculate

$$\lim \mathbb{E}\left[\hat{\pi}_{1,t}(z)\, R_{t+1} \mid \hat{\tilde{\beta}}_1\right] = \hat{\tilde{\beta}}_1^{\intercal}\, \Psi_{1,1}\, \beta_1\,.$$

Note that

$$\begin{aligned}\hat{\tilde{\beta}}_1^{\intercal}\, \Psi_{1,1}\, \beta_1 &= \beta_1^{\intercal}\, \Psi_{1,1} \left(z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1}\right)^{-1} \left(\hat{\Psi}_{\tilde{F},F,1}\beta_1 + \hat{\Psi}_{\tilde{F},F,2}\, \beta_2 + q_T\right) \\ &= \text{Term 1} + \text{Term 2} + \text{Term 3}\,, \end{aligned} \tag{B.1}$$

where

$$\hat{\Psi}_{\tilde{F},F,1} = \frac{1}{T}\sum_t \tilde{F}_{t,(1)}\, F_{t,(1)}^{\intercal}\,, \qquad \hat{\Psi}_{\tilde{F},F,2} = \frac{1}{T}\sum_t \tilde{F}_{t,(1)}\, F_{t,(2)}^{\intercal}\,, \qquad q_T = \frac{1}{T}\sum_t \tilde{F}_{t,(1)}\, \epsilon_{t+1}\,.$$

With the same arguments as those in the proof of Theorem 1 and $\mathbb{E}(\tilde{F}_{t,(1)} F_{t,(2)}^{\intercal}) = 0$ and $\mathbb{E}(\beta_1 \beta_2^{\intercal}) = 0$, we have

$$\text{Term 2} \to 0\,, \qquad \text{Term 3} \to 0\,, \qquad \text{and} \qquad \text{Term 1} \to \frac{b_*\, c^{-1}\, \xi(z; c\, q; a\, q)}{(1 + \xi(z; c\, q; a\, q))\,(1 + a\, q)}.$$

Therefore, we have

$$\mathcal{E}(z;c;a;q) = \frac{b_*\, c^{-1}\, \xi(z;c\,q;a\,q)}{(1+\xi(z;c\,q;a\,q))\,(1+a\,q)}\,,$$

Now we calculate $\lim \mathbb{E}\left[\hat{\pi}_{1,t}(z)^2 \mid \hat{\tilde{\beta}}_1\right] = \lim \operatorname{tr}\left(\tilde{\Psi}_{1,1}\,\hat{\tilde{\beta}}_1\,\hat{\tilde{\beta}}_1^\intercal\right)$. We know that

$$\begin{aligned}\operatorname{tr}\left(\tilde{\Psi}_{1,1}\,\hat{\tilde{\beta}}_1\,\hat{\tilde{\beta}}_1^\intercal\right) = \operatorname{tr}\Bigg(\tilde{\Psi}_{1,1}\,\left(z\,I+\hat{\tilde{\Psi}}_{1,1}\right)^{-1}\Big(\hat{\Psi}_{\tilde{F},F,1}\,\beta_1\,\beta_1^\intercal\hat{\Psi}^\intercal_{\tilde{F},F,1} \\ +\hat{\Psi}_{\tilde{F},F,2}\beta_2\beta_2^\intercal\hat{\Psi}^\intercal_{\tilde{F},F,2}+q_T q_T^\intercal\Big)\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\Bigg) \qquad \text{(B.2)}\\ = \text{ Term 1} + \text{ Term 2} + \text{Term 3}\,. \qquad \text{(B.3)}\end{aligned}$$

With the same arguments as in the proof of Theorem 1 and note that $\mathbb{E}[\tilde{F}_{t,(1)}\,F^\intercal_{t,(2)}] = 0$ and $\mathbb{E}[\beta_1\,\beta_2^\intercal] = 0$, we have

$$\text{Term 3} \to \xi(z;c\,q;a\,q) + z\,\frac{d}{dz}\xi(z;c\,q;a\,q)\,,$$

$$\begin{aligned}\text{Term 1} \to \frac{b_*\operatorname{tr}(\Psi_{1,1})}{P}\,(\xi(z;\,c\,q;\,a\,q)+z\,\partial\xi(z;\,c\,q;\,a\,q)/\partial z) - \frac{b_*\,c^{-1}\,\partial\xi(z;\,c\,q;\,a\,q)/\partial z}{(1+\xi(z;\,c\,q;\,a\,q))^2} \\ - \frac{2b_*\,c^{-1}\,(\xi(z;\,c\,q;\,a\,q)+z\,\partial\xi(z;\,c\,q;\,a\,q)/\partial z)\,\xi(z;\,c\,q;\,a\,q)}{(1+\xi(z;\,c\,q;\,a\,q))\,(1+a\,q)}\,,\end{aligned}$$

$$\text{Term 2} \to \frac{b_*\operatorname{tr}(\Psi_{2,2})}{P}\,(\xi(z;\,c\,q;\,a\,q)+z\,\partial\xi(z;\,c\,q;\,a\,q)/\partial z)\,.$$

Altogether, we have

$$\begin{aligned}\mathcal{L}(z;c;a;q) = (1+b_*)\,(\xi(z;\,c\,q;\,a\,q)+z\,\partial\xi(z;\,c\,q;\,a\,q)/\partial z) - \frac{b_*\,c^{-1}\,\partial\xi(z;\,c\,q;\,a\,q)/\partial z}{(1+\xi(z;\,c\,q;\,a\,q))^2} \\ - \frac{2b_*\,c^{-1}\,(\xi(z;\,c\,q;\,a\,q)+z\,\partial\xi(z;\,c\,q;\,a\,q)/\partial z)\,\xi(z;\,c\,q;\,a\,q)}{(1+\xi(z;\,c\,q;\,a\,q))\,(1+a\,q)}\,.\end{aligned}$$

Finally, with the same argument as that in the proof of Theorem 1, we have

$$\begin{aligned}\lim \mathbb{E}\left[(\hat{\pi}_{1,t}(z)\,R_{t+1})^2 \mid \hat{\tilde{\beta}}_1\right] &= \lim \hat{\tilde{\beta}}_1^\intercal\,\mathbb{E}\left[\tilde{F}_{t,(1)}F_t^\intercal\,\beta\,\beta^\intercal\,F_t\tilde{F}^\intercal_{t,(1)}\right]\hat{\tilde{\beta}}_1 + \lim \operatorname{tr}\left(\tilde{\Psi}_{1,1}\,\hat{\tilde{\beta}}_1\,\hat{\tilde{\beta}}_1^\intercal\right)\\ &= 2\,\mathcal{E}(z;c;a;q)^2 + (1+b_*)\,\mathcal{L}(z;c;a;q)\,.\end{aligned}$$

□

B.1.2 Proof of Corollary 3.

Throughout the proof, all convergences are in probability unless otherwise stated. In addition, the limits are taken as $T, P, P_1 \to \infty$, $P/T \to c$ and $P_1/P = q$.

Proof. In this section we prove Corollary 3, let's first consider the limit:

$$\lim_{T\to\infty} T^{-1} \operatorname{tr}\left((z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1})^{-1}\, \tilde{\Psi}_{1,1}\right).$$

We know that, by Proposition 1 we have

$$\lim_{T\to\infty} T^{-1} \operatorname{tr}\left((z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1})^{-1}\, \tilde{\Psi}_{1,1}\right) = \frac{1 - z\, m_1(-z;c;q)}{(c\,q)^{-1} - 1 + z\, m_1(-z;c;q)} := \xi_1(z;c;q)\,, \tag{B.4}$$

where $m_1(u;c;q)$ solves

$$m_1(u;c;q) = \frac{1}{1 - c\,q - c\,q\,u\,m_1(u;c;q)}\, m_{\tilde{\Psi}_{1,1}}\left(\frac{u}{1 - c\,q - c\,q\,u\,m_1(u;c;q)}\right), \tag{B.5}$$

where

$$m_{\tilde{\Psi}_{1,1}}(u) = \lim_{P_1\to\infty} P_1^{-1} \operatorname{tr}\left((\tilde{\Psi}_{1,1} - u\, I_{P_1})^{-1}\right)$$

is the limiting Stieltjes transform of the eigenvalue distribution of $\tilde{\Psi}_{1,1}$. As

$$\tilde{\Psi}_{1,1} = \operatorname{diag}\left(1 + f(1/P), 1 + f(2/P), \cdots, 1 + f(P_1/P)\right)$$

and $P_1/P \to q$, we have

$$\begin{aligned}\lim_{P_1\to\infty} P_1^{-1} \operatorname{tr}\left((\tilde{\Psi}_{1,1} - u\, I_{P_1})^{-1}\right) &= \lim_{P_1\to\infty} P_1^{-1} \sum_{i=1}^{P_1} \left(1 + f(i/P) - u\right)^{-1} \\ &= \frac{1}{q}\int_0^q \frac{1}{1 + f(x) - u} dx\,.\end{aligned}$$

Note that $m_{\tilde{\Psi}_{1,1}}(u)$ is a complicated integral form, which means equation

(B.5) do not have a close form solution in general. Now we consider

$$\lim_{T\to\infty} T^{-1} \operatorname{tr}\left(\mathbb{E}\left[\left(z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1}\right)^{-1} \tilde{\Psi}_{1,1} \left(z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1}\right)^{-1} \Psi_{1,1}\, \Psi_{1,1}^{\intercal}\right]\right) .$$

As $\Psi_{1,1} = I_{P_1}$, we have

$$\begin{aligned}
&\lim_{T\to\infty} T^{-1} \operatorname{tr}\left(\mathbb{E}\left[\left(z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1}\right)^{-1} \tilde{\Psi}_{1,1} \left(z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1}\right)^{-1} \Psi_{1,1}\, \Psi_{1,1}^{\intercal}\right]\right) \\
&= \lim_{T\to\infty} T^{-1} \operatorname{tr}\left(\mathbb{E}\left[\left(z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1}\right)^{-1} \tilde{\Psi}_{1,1} \left(z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1}\right)^{-1}\right]\right) \\
&= \lim_{T\to\infty} T^{-1} \operatorname{tr}\left(\mathbb{E}\left[\left(z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1}\right)^{-1} \left(z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1}\right)^{-1} \tilde{\Psi}_{1,1}\right]\right) \\
&= -\frac{\partial \xi_1(z;c;q)}{\partial z} . \qquad \text{(B.6)}
\end{aligned}$$

Next we consider

$$\lim_{T\to\infty} T^{-1} \operatorname{tr}\left(\mathbb{E}\left[\left(z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1}\right)^{-1} \Psi_{1,1}\, \Psi_{1,1}^{\intercal}\right]\right) = \lim_{T\to\infty} T^{-1} \operatorname{tr}\left(\mathbb{E}\left[\left(z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1}\right)^{-1}\right]\right) .$$

We have

$$\begin{aligned}
\lim_{T\to\infty} T^{-1} \operatorname{tr}\left(\mathbb{E}\left[\left(z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1}\right)^{-1}\right]\right) &= \lim_{T\to\infty} c\, q\, P_1^{-1} \operatorname{tr}\left(\mathbb{E}\left[\left(z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1}\right)^{-1}\right]\right) \\
&= c\, q\, m_1(-z;c;q) . \qquad \text{(B.7)}
\end{aligned}$$

Write $\beta_1 = \{\beta_i\}_{i=1}^{P_1}$ and $\beta_2 = \{\beta_i\}_{i=P_1+1}^{P}$, now we calculate

$$\lim \mathbb{E}\left[\hat{\pi}_{1,t}(z)\, R_{t+1} \mid \hat{\tilde{\beta}}_1\right] = \hat{\tilde{\beta}}_1^{\intercal}\, \Psi_{1,1}\, \beta_1 .$$

Note that

$$\begin{aligned}
\hat{\tilde{\beta}}_1^{\intercal}\, \Psi_{1,1}\, \beta_1 &= \beta_1^{\intercal}\, \Psi_{1,1} \left(z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1}\right)^{-1} \left(\hat{\Psi}_{\tilde{F},F,1}\beta_1 + \hat{\Psi}_{\tilde{F},F,2}\, \beta_2 + q_T\right) \\
&= \text{Term 1} + \text{Term 2} + \text{Term 3} ,
\end{aligned}$$

where

$$\hat{\Psi}_{\tilde{F},F,1} = \frac{1}{T}\sum_t \tilde{F}_{t,(1)}\, F_{t,(1)}^{\intercal} , \qquad \hat{\Psi}_{\tilde{F},F,2} = \frac{1}{T}\sum_t \tilde{F}_{t,(1)}\, F_{t,(2)}^{\intercal} , \qquad q_T = \frac{1}{T}\sum_t \tilde{F}_{t,(1)}\, \epsilon_{t+1} .$$

With the same arguments as those in the proof of Theorem 1 and $\mathbb{E}(\tilde{F}_{t,(1)}F_{t,(2)}^{\intercal}) = 0$ and $\mathbb{E}(\beta_1\beta_2^{\intercal}) = 0$, plug in equation (B.4) and (B.7) we have

$$\text{Term } 2 \to 0\,, \qquad \text{Term } 3 \to 0\,, \qquad \text{and} \qquad \text{Term } 1 \to \frac{b_*\, q\, m_1(-z;c;q)}{1+\xi_1(z;c;q)}.$$

Therefore, we have

$$\mathcal{E}(z;c;q) = \frac{b_*\, q\, m_1(-z;c;q)}{1+\xi_1(z;c;q)}\,,$$

Now we calculate $\lim \mathbb{E}\left[\hat{\pi}_{1,t}(z)^2 \mid \hat{\tilde{\beta}}_1\right] = \lim \operatorname{tr}\left(\tilde{\Psi}_{1,1}\,\hat{\tilde{\beta}}_1\,\hat{\tilde{\beta}}_1^{\intercal}\right)$. We know that

$$\begin{aligned}\operatorname{tr}\left(\tilde{\Psi}_{1,1}\,\hat{\tilde{\beta}}_1\,\hat{\tilde{\beta}}_1^{\intercal}\right) &= \operatorname{tr}\left(\tilde{\Psi}_{1,1}\,\left(z\,I + \hat{\tilde{\Psi}}_{1,1}\right)^{-1}\left(\hat{\Psi}_{\tilde{F},F,1}\,\beta_1\,\beta_1^{\intercal}\hat{\Psi}_{\tilde{F},F,1}^{\intercal}\right.\right.\\ &\qquad \left.\left. + \hat{\Psi}_{\tilde{F},F,2}\beta_2\beta_2^{\intercal}\hat{\Psi}_{\tilde{F},F,2}^{\intercal} + q_T q_T^{\intercal}\right)\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\right) \qquad \text{(B.8)}\\ &= \text{ Term } 1 + \text{ Term } 2 + \text{Term } 3\,. \qquad \text{(B.9)}\end{aligned}$$

With the same arguments as in the proof of Theorem 1 and note that $\mathbb{E}[\tilde{F}_{t,(1)}\,F_{t,(2)}^{\intercal}] = 0$ and $\mathbb{E}[\beta_1\,\beta_2^{\intercal}] = 0$, plug in equations (B.4), (B.7) and (B.6) we have

$$\text{Term } 3 \to \xi_1(z;c;q) + z\,\frac{d}{dz}\xi_1(z;c;q)\,,$$

$$\begin{aligned}\text{Term } 1 \to \frac{b_*\operatorname{tr}(\Psi_{1,1})}{P}\,\left(\xi_1(z;c;q) + z\,\partial\xi_1(z;c;q)/\partial z\right) - \frac{b_*\,c^{-1}\,\partial\xi_1(z;c;q)/\partial z}{(1+\xi_1(z;c;q))^2}\\ - \frac{2b_*\,q\,\left(\xi_1(z;c;q) + z\,\partial\xi_1(z;c;q)/\partial z\right)\,m_1(-z;c;q)}{(1+\xi_1(z;c;q))}\,,\end{aligned}$$

$$\text{Term } 2 \to \frac{b_*\operatorname{tr}(\Psi_{2,2})}{P}\,\left(\xi_1(z;c;q) + z\,\partial\xi_1(z;c;q)/\partial z\right)\,.$$

Altogether, we have

$$\begin{aligned}\mathcal{L}(z;c;q) = (1+b_*)\,\left(\xi_1(z;c;q) + z\,\partial\xi_1(z;c;q)/\partial z\right) - \frac{b_*\,c^{-1}\,\partial\xi_1(z;c;q)/\partial z}{(1+\xi_1(z;c;q))^2}\\ - \frac{2b_*\,q\,\left(\xi_1(z;c;q) + z\,\partial\xi_1(z;c;q)/\partial z\right)\,m_1(-z;c;q)}{(1+\xi_1(z;c;q))}\,.\end{aligned}$$

Finally, with the same argument as that in the proof of Theorem 1, we have

$$\begin{aligned}\lim \mathbb{E}\left[\left(\hat{\pi}_{1,t}(z)\, R_{t+1}\right)^2 \mid \hat{\tilde{\beta}}_1\right] &= \lim \hat{\tilde{\beta}}_1^{\intercal}\, \mathbb{E}\left[\tilde{F}_{t,(1)} F_t^{\intercal}\, \beta\, \beta^{\intercal}\, F_t \tilde{F}_{t,(1)}^{\intercal}\right] \hat{\tilde{\beta}}_1 + \lim \operatorname{tr}\left(\tilde{\Psi}_{1,1}\, \hat{\tilde{\beta}}_1\, \hat{\tilde{\beta}}_1^{\intercal}\right) \\ &= 2\, \mathcal{E}(z;c;q)^2 + (1+b_*)\, \mathcal{L}(z;c;q)\,.\end{aligned}$$

□

B.1.3 Proof of Proposition 3.

Proof. Equate derivative

$$\frac{\partial R^2(z;c;q;a)}{\partial z} = \frac{2\,q\,((a+1)\,c\,(ab_*+a-b_*\,q+b_*+1)-b_*\,z)}{(b_*+1)\,\left((a\,c\,q+a+c\,q+z+1)^2-4\,(a+1)^2\,c\,q\right)^{3/2}}$$

to zero shows that the optimal point of $R^2(z;c;a;q)$ is at $z_{*,1} = c\,(a\,q+1)\,(a\,b\,q+a\,q-b\,q+b_*+1)/b_*$. Next, consider

$$SR(z;c;a;q) = \frac{1}{\sqrt{2+(1+b_*)\;\mathcal{L}(z;c;a;q)/\left(\mathcal{E}(z;c;a;q)\right)^2}}\,.$$

We only need to consider

$$\widetilde{SR}(z;c;a;q) = \mathcal{L}(z;c;a;q)/\left(\mathcal{E}(z;c;a;q)\right)^2\,.$$

Note that

$$\frac{\partial \widetilde{SR}(z;c;q;a)}{\partial z} = \frac{2\,(a+1)^2\,c\,(b_*z-(a+1)\,c\,(a\,b_*+a-b_*\,q+b_*+1))}{b_*^2\,\left((a\,c\,q+a+c\,q+z+1)^2-4\,(a+1)^2\,c\,q\right)^{3/2}}\,.$$

Therefore, $z_{*,1} = c\,(a\,q+1)\,(a\,b\,q+a\,q-b\,q+b_*+1)/b_*$ is the optimal point of $SR(z;c;a;q)$, the first part of the proof is complete. Next, we consider the second part. A simple calculation shows that

$$\mathcal{E}(z_{*,1};c;a;q) = \mathcal{L}(z_{*,1};c;a;q)\,.$$

Therefore, we only need to consider $\mathcal{E}(z_{*,1}; c; a; q)$. Next, the expression

$$\frac{\partial \mathcal{E}(z_{*,1}; c; q; a)}{\partial q} = \left(-a\, b_* \sqrt{(a\, q + 1)^2 \left(\frac{(a\,(b_* + 1)\, c\, q + b_*\, c + b_* + c)^2}{b_*^2} - 4\, c\, q \right)} \right.$$
$$+ b_* \,(a\, q + 1)\,((a - 2)\, a\, c\, q + a\, c + a + 2\, c) + a\, c\,(a\, q + 1)^2 \Big)$$
$$\Big/ \left(2\, c\,(a\, q + 1)^2 \sqrt{(a\, q + 1)^2 \left(\frac{(a\,(b_* + 1)\, c\, q + b_*\, c + b_* + c)^2}{b_*^2} - 4\, c\, q \right)} \right)$$

is zero at

$$q_{*,1} = \frac{-\sqrt{a^4\, b_*^2\, c^2 + a^4\, b_*^2\, c + 2\, a^4\, b_*\, c^2 + a^4\, b_*\, c + a^4\, c^2 - a^3\, b_*^2\, c} - a\, b_*\, c}{a^2\, c\,(a + a\, b_* - b_*)},$$

$$q_{*,2} = \frac{\sqrt{a^4\, b_*^2\, c^2 + a^4\, b_*^2\, c + 2\, a^4\, b_*\, c^2 + a^4\, b_*\, c + a^4\, c^2 - a^3\, b_*^2\, c} - a\, b_*\, c}{a^2\, c\,(a + a\, b_* - b_*)}.$$

Now, if $a > 1$ and $c > \frac{a\, b}{(a-1)\,(a^2\,(b_*+1)+a+b_*)}$, we have $q_{*,1} < 0$, $q_{*,2} > 0$, and $\partial \mathcal{E}(z_{*,1}; c; q; a)/\partial q > 0$ for $0 < q < q_{*,2}$ and $\partial \mathcal{E}(z_{*,1}; c; q; a)/\partial q < 0$ for $q_{*,2} < q < 1$. Otherwise, $\partial \mathcal{E}(z_{*,1}; c; q; a)/\partial q > 0$ for $0 < q < 1$.

□

B.2 Motivating examples

Here, we provide the details of how we get the results for Figure 4.1, 4.2 and 4.3 in the introduction.

For Figure 4.1, we follow the same procedure as the Random Fourier Features (RFF) analysis in KMZ, including identical data-processing steps. We vary P from 3 to 10000 and all other parameters and training steps remain as the same of KMZ.

For Figure 4.2, following [31], we collect monthly equity returns from CRSP for all firms listed on the NYSE, AMEX, and Nasdaq from January 1990 to December 2024. We incorporate 94 stock characteristics from [30], 74 industry dummies, and 8 macroeconomic predictors from [63], closely following the procedure in [31].

We employ a two-layer MLP for our model. The first layer has 32 neurons, while the second layer ranges from 5 to 10,000 neurons on a log scale. We use the ELU activation function [21] in each layer for its smoothness. We also apply l_2 regularization during training, with the l_2 parameter varying from 0.001 to 10,000 on a log scale. The data from January 1990 to December 2020 serves as our training set, from January 2021 to December 2022 as our validation set, and from January 2023 to December 2024 as our test set. We train the model using stochastic gradient descent (SGD) for 50 epochs with a batch size of 10,000, and we report averages over 10 repetitions.

For Figure 4.3, we follow the same procedure as the RFF analysis in KMZ, including identical data-processing steps. the only difference is that we deliberately add noise to the RFFs.

For Figure 4.3, in the left-hand panel, the noise is given by

$$\tilde{F}_{i,t} = [\sin(\kappa\,\omega_i^\intercal\,G_t) + a\,\mu_{i,t},\ \cos(\kappa\,\omega_i^\intercal\,G_t) + a\,\eta_{i,t}]^\intercal\ ,\ (\omega_i^\intercal, \mu_{i,t}, \eta_{i,t})^\intercal \sim \text{ i.i.d. } \mathrm{N}(0, I_{17})\,,$$

and in the right-hand panel the noise is given by

$$\tilde{F}_{i,t} = [\sin(\kappa\,\omega_i^\intercal\,G_t) + a\,q\,\mu_{i,t},\ \cos(\kappa\,\omega_i^\intercal\,G_t) + a\,q\,\eta_{i,t}]^\intercal\ ,\ (\omega_i^\intercal, \mu_{i,t}, \eta_{i,t})^\intercal \sim \text{ i.i.d. } \mathrm{N}(0, I_{17})\,,$$

where $q = P_1/P_{max}$ and $P_{max} = 10000$. We vary P_1 from 3 to 10000, and scale the noise by a which is in the range 1 to 10 in both cases.

B.3 Autoencoder details

In this section we provide details of the training process mentioned in section 4.2. First, we use an autoencoder to produce a de-noised version of $F_{i,t}^{(P_1)}$, which we use to estimate a ridge regression for return prediction. Formally, the autoencoder consists of an encoder and a decoder. The encoder is a function $\mathcal{E}_{\theta_1} : \mathbb{R}^{P_1} \to \mathbb{R}^{K_1}$ mapping the input to a latent representation,

$$Z_{i,t} = \mathcal{E}_{\theta_1}(F_{i,t}^{(P_1)})\,,$$

and the decoder $\mathcal{D}_{\theta_2} : \mathbb{R}^{K_1} \to \mathbb{R}^{P_1}$ reconstructs the signal

$$\hat{F}_{i,t}^{(P_1)} = \mathcal{D}_{\theta_2}(Z_{i,t}) = \mathcal{A}_\theta(F_{i,t}^{(P_1)})\,,$$

where $\theta = (\theta_1, \theta_2)$ denotes all the network parameters. In our setting, the encoder and decoder consist of two fully connected layers with ReLU activa-

tions and 40 hidden neurons, with $K_1 = 10$.

We now describe the empirical procedure. For each stock i, let T_i denote the last month of this stock for which data are available. We begin at $t = 120$, use the first 10 years of monthly data to train the initial autoencoder, and roll forward in 12-month steps until $t = T_i - 24$. In each step, we construct a one-year ridge regression training window (from $t+1$ to $t+12$) and evaluate the forecasts in the subsequent 12-month out-of-sample window (from $t+13$ to $t+24$). The procedure is as follows:

1. **Feature selection and autoencoder training**:
   For each level of complexity $P_1 \in \{100, 200, \ldots, 846\}$, we randomly select a fixed subset of P_1 features from the full 846 dimensional vector. Then, for each stock i, and at each decision point t, we use the 10-year window of observations $\{F_{i,s}^{(P_1)}\}_{s=t-119}^{t}$ to train a separate autoencoder. The autoencoder, denoted by $\mathcal{A}_\theta = \mathcal{D}_{\theta_2} \circ \mathcal{E}_{\theta_1}$, is trained to minimize the average squared reconstruction error over this window, that is

$$\min_{\theta} \sum_{s=t-119}^{t} \left\| F_{i,s}^{(P_1)} - \mathcal{A}_\theta(F_{i,s}^{(P_1)}) \right\|^2 .$$

   This training step is repeated independently for each stock i and each time t.

2. **De-noising and ridge regression training**:
   Train the autoencoder at time t, and apply it to the next 12 months of features, $\{F_{i,s}^{(P_1)}\}_{s=t+1}^{t+12}$, to obtain de-noised feature representations

$$\hat{F}_{i,s}^{(P_1)} = \mathcal{A}_\theta(F_{i,s}^{(P_1)}) , \quad s = t+1, \ldots, t+12 .$$

   Next, we fit a ridge regression with these de-noised features to predict one-month-ahead returns over the same period. Specifically, we estimate the regression model

$$R_{s+1,i} = \hat{\beta}_{t,i}^{\top} \hat{F}_{i,s}^{(P_1)} + \varepsilon_{s+1,i} , \quad s = t+1, \ldots, t+12 ,$$

   where $\hat{\beta}_t$ is the coefficient vector estimated with ridge regression. The ridge parameter is selected via cross-validation over the candidate values $\{0.001, 0.1, 1, 10, 100\}$. This regression is also estimated separately for each stock.

3. **Out-of-sample forecasting**:
   Use the ridge coefficients $\hat{\beta}_{t,i}$ from the de-noised features in the previous step to forecast one-month-ahead returns over the subsequent year, i.e., from $t+13$ to $t+24$. The forecasted return at time s is given by

$$\hat{r}_{s,i} = \hat{\beta}_{t,i}^{\top} \hat{F}_{i,s-1}^{(P_1)}, \quad s = t+13, \ldots, t+24\,.$$

   We also record the return of the timing strategy $\hat{r}_{s,i} R_{s,i}$ for each forecasted month.

4. **Rolling procedure**:
   After completing the forecasts for a given time t, we roll the window forward by 12 months, setting $t' = t+12$, and repeat step 1 to 3. That is, we use $\{F_{i,s}^{(P_1)}\}_{s=t'-119}^{t'}$ to retrain a new autoencoder, de-noise the next 12 months, estimate a new ridge regression, and generate forecasts for the following year. This rolling continues until the last possible training point, which is $t = T_i - 24$.

5. **Performance aggregation**:
   After completing the full rolling procedure for all valid time windows and all values of P_1, we aggregate the forecasted returns and realized returns for each stock i. For each complexity level P_1, we compute the out-of-sample R^2 and Sharpe ratio using the aggregated predictions and realized returns. We repeat step 1 to 4 five times and report the average results across repetitions.,

C | The Mechanical Virtue of Random Fourier Features and Model Complexity

C.1 Proofs

C.1.1 Proof of Proposition 4

Let $z_{i,\text{test}}^{\intercal}\, z_{i,\text{train}} = A_i$. Use the Sherman–Morrison formula and $z_{i,\text{train}}^{\intercal}\, z_{i,\text{train}} = P/2$ to write

$$\begin{aligned}
\hat{y}_i &= \hat{\beta}_i^{\intercal}\, z_{i,\text{test}} = y_{i,\text{train}}\, z_{i,\text{test}}^{\intercal}\, (z_{i,\text{train}}\, z_{i,\text{train}}^{\intercal} + \lambda\, I)^{-1}\, z_{i,\text{train}} \\
&= y_{i,\text{train}}\, z_{i,\text{test}}^{\intercal} \left(\lambda^{-1}\, I - \frac{\lambda^{-2}\, z_{i,\text{train}}\, z_{i,\text{train}}^{\intercal}}{1 + \lambda^{-1}\, z_{i,\text{train}}^{\intercal}\, z_{i,\text{train}}} \right) z_{i,\text{train}} \\
&= y_{i,\text{train}} \left[\frac{z_{i,\text{test}}^{\intercal}\, z_{i,\text{train}}}{\lambda} - \frac{\lambda^{-2} \left(z_{i,\text{test}}^{\intercal}\, z_{i,\text{train}} \right) \left(z_{i,\text{train}}^{\intercal}\, z_{i,\text{train}} \right)}{1 + \lambda^{-1}\, z_{i,\text{train}}^{\intercal}\, z_{i,\text{train}}} \right] \\
&= y_{i,\text{train}}\, A_i \left[\frac{1}{\lambda} - \frac{\lambda^{-2}\, P/2}{1 + \lambda^{-1}\, P/2} \right] \\
&= y_{i,\text{train}}\, A_i \left[\frac{1}{\lambda} - \frac{\lambda^{-1}\, P/2}{\lambda + P/2} \right] \\
&= \frac{y_{i,\text{train}}\, A_i}{\lambda + P/2}\,.
\end{aligned}$$

Now, consider $(\hat{y}_i - y_{i,\text{test}})^2$ and write

$$(\hat{y}_i - y_{i,\text{test}})^2 = \frac{y_{i,\text{train}}^2\, A_i^2}{(\lambda + P/2)^2} + y_{i,\text{test}}^2 - \frac{2\, y_{i,\text{test}}\, y_{i,\text{train}}\, A_i}{\lambda + P/2}\,.$$

Next, let $\Delta x_i = x_{i,\text{test}} - x_{i,\text{train}}$. We have

$$\cos\!\big(\gamma\, \omega_j^{\intercal}\, x_{i,\text{test}}\big)\, \cos\!\big(\gamma\, \omega_j^{\intercal}\, x_{i,\text{train}}\big) + \sin\!\big(\gamma\, \omega_j^{\intercal}\, x_{i,\text{test}}\big)\, \sin\!\big(\gamma\, \omega_j^{\intercal}\, x_{i,\text{train}}\big) = \cos\!\big(\gamma\, \omega_j^{\intercal}\, \Delta x_i\big)\,,$$

therefore

$$A_i = \sum_{j=1}^{P/2} \cos\!\big(\gamma\, \omega_j^{\intercal}\, \Delta x_i\big)\,.$$

We know that

$$\mathbb{E}_{\omega_j}\left[\cos\left(\gamma\,\omega_j^\intercal\,\Delta x_i\right)\right] = \exp\left(-\frac{\gamma^2\,||\Delta x_i||^2}{2}\right) = K\left(\gamma\,\Delta x_i\right)$$

because $\omega_j \overset{iid}{\sim} \mathcal{N}(0, I)$, where $K\left(\gamma\,\Delta x_i\right)$ is the Gaussian kernel and $||\cdot||$ is the l_2 norm of a vector. Therefore

$$\mathbb{E}[A_i] = \sum_{j=1}^{P/2} \mathbb{E}\left[\cos\left(\gamma\,\omega_j^\intercal\,\Delta x_i\right)\right] = \frac{P}{2}\,K\left(\gamma\,\Delta x_i\right).$$

To compute $\mathbb{E}[A_i^2]$, write

$$A_i = \sum_{j=1}^{m} u_j\,, \qquad m = \frac{P}{2}\,, \qquad u_j = \cos\left(\gamma\,\omega_j^\intercal\,\Delta x_i\right).$$

We have

$$A_i^2 = \sum_{j=1}^{m} u_j^2 + 2 \sum_{1\le j<k\le m} u_j\,u_k$$

because the u_j are i.i.d., hence

$$\mathbb{E}[A_i^2] = m\,\mathbb{E}[u_1^2] + 2\binom{m}{2}\mathbb{E}[u_1\,u_2]\,.$$

We know that

$$\mathbb{E}[u_1] = K\left(\gamma\,\Delta x_i\right), \qquad \mathbb{E}[u_1\,u_2] = K\left(\gamma\,\Delta x_i\right)^2.$$

Moreover,

$$u_1^2 = \cos^2\left(\gamma\,\omega_j^\intercal\,\Delta x_i\right) = \frac{1+\cos\left(2\,\gamma\,\omega_j^\intercal\,\Delta x_i\right)}{2}\,,$$

and because $2\,\gamma\,\omega_j^\intercal\,\Delta x_i \sim \mathcal{N}\left(0, 4\,\gamma^2\,||\Delta x_i||^2\right)$,

$$\mathbb{E}\left[\cos\left(2\,\gamma\,\omega_j^\intercal\,\Delta x_i\right)\right] = \exp\left(-\frac{(2\,\gamma)^2\,||\Delta x_i||^2}{2}\right) = \exp\left(-2\,\gamma^2\,||\Delta x_i||^2\right) = K\left(\gamma\,\Delta x_i\right)^4.$$

Thus,

$$\mathbb{E}\left[u_1^2\right] = \frac{1}{2}\left(1 + K\left(\gamma\,\Delta x_i\right)^4\right).$$

Put everything together to write

$$\mathbb{E}\left[A_i^2\right] = \frac{P}{2}\,\frac{1+K\big(\gamma\,\Delta x_i\big)^4}{2} + \frac{P}{2}\,\left(\frac{P}{2}-1\right)K\big(\gamma\,\Delta x_i\big)^2\,.$$

Next, substitute $\mathbb{E}[A_i]$ and $\mathbb{E}[A_i^2]$ into $\mathbb{E}\big[(\hat{y}_i - y_{i,\text{test}})^2\big]$ to obtain

$$\begin{aligned}\mathbb{E}\big[(\hat{y}_i - y_{i,\text{test}})^2\big] = \frac{y_{i,\text{train}}^2}{(\lambda+P/2)^2}\Bigg\{&\frac{P}{2}\,\frac{1+K\big(\gamma\,(x_{i,\text{test}}-x_{i,\text{train}})\big)^4}{2}\\ &+\frac{P}{2}\,\left(\frac{P}{2}-1\right)K\big(\gamma\,(x_{i,\text{test}}-x_{i,\text{train}})\big)^2\Bigg\}\\ &+y_{i,\text{test}}^2 - \frac{2\,y_{i,\text{test}}\,y_{i,\text{train}}}{\lambda+P/2}\,\frac{P}{2}\,K\big(\gamma\,(x_{i,\text{test}}-x_{i,\text{train}})\big)\,.\end{aligned}$$

In the ridgeless case $\lambda = 0$, we have

$$\begin{aligned}\mathbb{E}\big[(\hat{y}_i - y_{i,\text{test}})^2\big] = \frac{y_{i,\text{train}}^2}{P}\,&\Big\{1+K\big(\gamma\,(x_{i,\text{test}}-x_{i,\text{train}})\big)^4 + (P-2)\,K\big(\gamma\,(x_{i,\text{test}}-x_{i,\text{train}})\big)^2\Big\}\\ &+y_{i,\text{test}}^2 - 2\,y_{i,\text{test}}\,y_{i,\text{train}}\,K\big(\gamma(x_{i,\text{test}}-x_{i,\text{train}})\big)\,.\end{aligned}$$

Therefore, the expected R^2 is

$$\mathbb{E}\left[R^2\right] = 1 - \frac{1}{\sum_{i=1}^N (y_{i,\text{test}}-\overline{y}_{\text{test}})^2}\,\sum_{i=1}^N \mathbb{E}\left[(\hat{y}_i - y_{i,\text{test}})^2\right]\,,$$

where

$$\begin{aligned}\mathbb{E}\left[(\hat{y}_i - y_{i,\text{test}})^2\right] = \frac{y_{i,\text{train}}^2}{(\lambda+P/2)^2}\Bigg[&\frac{P}{2}\,\frac{1+K(\gamma\,(x_{i,\text{test}}-x_{i,\text{train}}))^4}{2}\\ &+\frac{P}{2}\,\left(\frac{P}{2}-1\right)K(\gamma\,(x_{i,\text{test}}-x_{i,\text{train}}))^2\Bigg]\\ &+y_{i,\text{test}}^2 - \frac{2\,y_{i,\text{test}}\,y_{i,\text{train}}\,P}{2\,\lambda+P}\,K(\gamma\,(x_{i,\text{test}}-x_{i,\text{train}}))\,.\end{aligned}$$

In the ridgeless case $\lambda = 0$, we have

$$\begin{aligned}\mathbb{E}[R^2] = 1 - \frac{1}{\sum_{i=1}^{N}(y_{i,\text{test}} - \overline{y}_{\text{test}})^2} \sum_{i=1}^{N} \Bigg[\frac{y_{i,\text{train}}^2}{P} \Big\{ 1 + K(\gamma\,(x_{i,\text{test}} - x_{i,\text{train}}))^4 \\ + (P-2)\,K(\gamma\,(x_{i,\text{test}} - x_{i,\text{train}}))^2 \Big\} \\ + y_{i,\text{test}}^2 - 2\,y_{i,\text{test}}\,y_{i,\text{train}}\,K(\gamma\,(x_{i,\text{test}} - x_{i,\text{train}})) \Bigg] .\end{aligned}$$

In the ridgeless case $\lambda = 0$, the R^2 is monotonically increasing in P because

$$1 + K\big(\gamma(x_{i,\text{test}} - x_{i,\text{train}})\big)^4 \geq 2\,K\big(\gamma(x_{i,\text{test}} - x_{i,\text{train}})\big)^2 .$$

C.1.2 Proof of Proposition 5

Use the same argument as that in the proof of Proposition 4 and write

$$Ret_i = \hat{y}_i\,y_{i,\text{test}} = \frac{y_{i,\text{test}}\,y_{i,\text{train}}\,A_i}{\lambda + P/2} .$$

Therefore, the expectation of Ret is

$$\begin{aligned}\mathbb{E}[Ret] &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[Ret_i] \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{P/2}{\lambda + P/2}\,y_{i,\text{train}}\,y_{i,\text{test}}\,K\big(\gamma\,(x_{i,\text{test}} - x_{i,\text{train}})\big) .\end{aligned}$$

In particular, in the ridgeless case $\lambda = 0$,

$$\mathbb{E}[Ret] = \frac{1}{N}\sum_{i=1}^{N} y_{i,\text{train}}\,y_{i,\text{test}}\,K\big(\gamma\,(x_{i,\text{test}} - x_{i,\text{train}})\big) .$$

C.1.3 Proof of Proposition 6

By the law of large numbers and the central limit theorem, we have, when $P \to \infty$,

$$\frac{1}{P/2}\, Ret \xrightarrow{\mathbb{P}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\lambda + P/2}\, y_{i,\text{train}}\, y_{i,\text{test}}\, K\big(\gamma\,(x_{i,\text{test}} - x_{i,\text{train}})\big)\,,$$

and

$$\begin{aligned} Ret_i^2 &= \frac{y_{i,\text{train}}^2\, y_{i,\text{test}}^2\, A_i^2}{(\lambda + P/2)^2}, \\ \frac{1}{(P/2)^2}\, Ret_i^2 &= \left(\frac{y_{i,\text{train}}\, y_{i,\text{test}}}{\lambda + P/2}\right)^2 \frac{A_i^2}{(P/2)^2} \\ &\xrightarrow{\mathbb{P}} \left(\frac{y_{i,\text{train}}\, y_{i,\text{test}}}{\lambda + P/2}\right)^2 K\big(\gamma\,(x_{i,\text{test}} - x_{i,\text{train}})\big)^2\,, \qquad P \to \infty\,. \end{aligned}$$

Hence, by the continuous mapping theorem,

$$\begin{aligned} \frac{1}{(P/2)^2} \sum_{i=1}^{N} (Ret_i - Ret)^2 \xrightarrow{\mathbb{P}} \sum_{i=1}^{N} \Bigg(& y_{i,\text{train}}\, y_{i,\text{test}}\, K\big(\gamma(x_{i,\text{test}} - x_{i,\text{train}})\big) \\ & - \frac{1}{N} \sum_{j=1}^{N} y_{j,\text{train}}\, y_{j,\text{test}}\, K\big(\gamma(x_{j,\text{test}} - x_{j,\text{train}})\big) \Bigg)^2. \end{aligned}$$

Therefore, the asymptotic approximation of the Sharpe ratio

$$SR = \frac{Ret}{\sqrt{\frac{1}{N} \sum_{i=1}^{N} (Ret_i - Ret)^2}}$$

is

$$SR \xrightarrow{\mathbb{P}} \frac{\frac{1}{\sqrt{N}} \sum_{i=1}^{N} y_{i,\text{train}}\, y_{i,\text{test}}\, K\big(\gamma\,(x_{i,\text{test}} - x_{i,\text{train}})\big)}{\sqrt{V_N}}\,,$$

where

$$V_N = \sum_{i=1}^{N} \Bigg(y_{i,\text{train}}\, y_{i,\text{test}}\, K\big(\gamma\,(x_{i,\text{test}} - x_{i,\text{train}})\big) - \frac{1}{N} \sum_{j=1}^{N} y_{j,\text{train}}\, y_{j,\text{test}}\, K\big(\gamma\,(x_{j,\text{test}} - x_{j,\text{train}})\big) \Bigg)^2 .$$

C.1.4 Proof of Proposition 7

We know that

$$Ret = \frac{1}{N} \sum_{i=1}^{N} Ret_i = \frac{1}{N} \sum_{i=1}^{N} y_{i,\text{train}}\, y_{i,\text{test}}\, A_i \,,$$

and that

$$Var = \frac{1}{N} \sum_{i=1}^{N} (Ret_i - Ret)^2 \,.$$

Therefore, when each $y_{i,\text{train}}\, y_{i,\text{test}}$ is replaced by $-y_{i,\text{train}}\, y_{i,\text{test}}$, we have

$$\widetilde{Ret_i} = -Ret_i \,, \qquad \widetilde{Ret} = -Ret \,,$$

and

$$\widetilde{Var} = \frac{1}{N} \sum_{i=1}^{N} (\widetilde{Ret_i} - \widetilde{Ret})^2 = \frac{1}{N} \sum_{i=1}^{N} \big(-(Ret_i - Ret)\big)^2 = Var \,.$$

Hence the Sharpe ratio becomes

$$\widetilde{SR} = \frac{\widetilde{Ret}}{\sqrt{\widetilde{Var}}} = \frac{-Ret}{\sqrt{Var}} = -SR \,,$$