The Limited Virtue of Complexity in a Noisy World

Álvaro Cartea [*†] Qi Jin[*‡] Yuantao Shi [*†§]

Click here for latest version

Abstract

Measurement error in features affects the out-of-sample performance of portfolios and of return predictions. In a high-dimensional feature space, increasing model complexity under proper regularization can enhance the predictability of asset returns; however, the Sharpe ratio of a portfolio of assets and the R-squared of the prediction of the asset returns decrease monotonically and are convex as the noise level in features increases. Furthermore, when only a subset of features is observed, there is an optimal level of complexity beyond which incorporating additional features can degrade portfolio performance due to the effect of noise in the features. Thus, the marginal benefits of increasing model complexity will, at some point, start to diminish as the number of observed features increases. Our findings underscore a limited virtue of complexity in financial forecasting, where the performance of portfolios depends on the noise level in features, and where more complex models do not necessarily lead to better performance when features are not perfectly observed.

Keywords: Portfolio choice, machine learning, random matrix theory
JEL classification: G11, G12, G14

[*]Oxford-Man Institute of Quantitative Finance, University of Oxford
[†]Mathematical Institute, University of Oxford
[‡]Department of Statistics, University of Oxford
[§]Corresponding author; Email: yuantao.shi@queens.ox.ac.uk

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 and high-dimensional firm-level characteristics. 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 including more predictive features adversely affects the performance of their models?

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 (Box, 1976). 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 (Belkin et al., 2019; Bartlett et al., 2020; Hastie et al., 2022; Kelly et al., 2024a). These developments give rise to two seemingly opposing views, one of which favors simplicity, and the other one embraces complexity.

This paper 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. Next, we provide three examples to motivate our framework. The first example replicates the results of Kelly et al. (2024a), henceforth KMZ, and the other two examples focus on the relationship between model complexity and predictive performances 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. 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 1 shows that, consistent with the findings of KMZ, increasing model complexity, i.e., including more features, improves the out-of-sample R-square of the predictions.

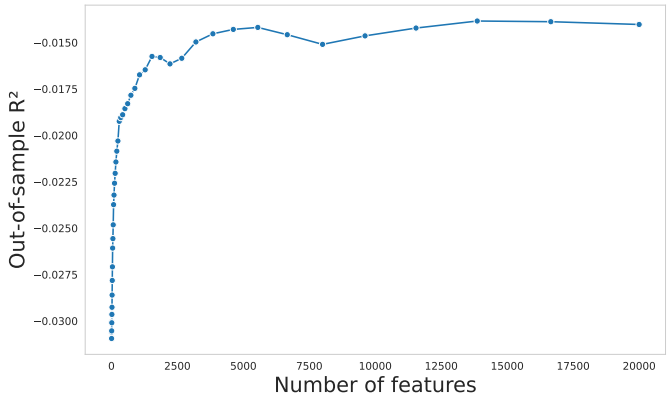


Figure 1: Prediction of stock returns using simulated features that are observed without noise. The x-axis shows the number of features, the y-axis is the out-of-sample R-squared. The results are averaged over 50 repetitions.

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., Gu et al., 2019; Giglio et al., 2022; Jiang et al., 2023; Chen et al., 2024).[1] 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.

Therefore, our second example proceeds as in Gu et al. (2019) where we use market data from NYSE, AMEX, and Nasdaq between 1990 and 2024 to fit a neural network to predict monthly stock returns.[2] Our findings show that the out-of-sample performance of the model does not monotonically increase as the model becomes more complex, see Figure 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 1, where out-of-sample model performance should increase as the model becomes more complex.[3]

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 level of noise in each

[1] Additional examples include Heaton et al. (2016), Sirignano (2019), Gu et al. (2021), Guijarro-Ordonez et al. (2021), Leippold et al. (2022), Avramov et al. (2023), Brogaard and Zareei (2023), Cao et al. (2024), Feng et al. (2024), and Kelly et al. (2024a).

[2] For details of our training process, see Appendix B.

[3] 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.

[4] 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., Gu et al. (2019) and Jiang et al. (2023).

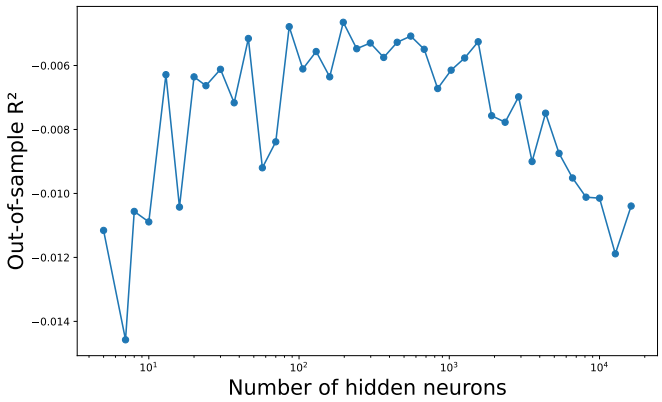


Figure 2: Stock return prediction with a neural network with data from NYSE, AMEX, and Nasdaq from 1990 to 2024. The x-axis shows model complexity (i.e., number of neurons), and the y-axis depicts out-of-sample R-squared. The results are averaged over 20 repetitions.[4]

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, Hastie et al. (2009) and James et al. (2013) show that increasing the complexity of the model in a limited sample setting tends to inflate 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. 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 3 shows how the noise in the artificial features generated with the approach in KMZ affects the link between complexity and model performance. Both panels of the figure show the out-of-sample R-square 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. In the left-hand panel, the noise level of each feature does not change as model complexity 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 increase.

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. 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 2, the results in the right-hand panel of Figure 3 show that as the quality of features decreases more and more, model complexity does not always increase the out-of-sample predictive performance of the model, so the potential gains from adding more features is limited, regardless of how features are learned.

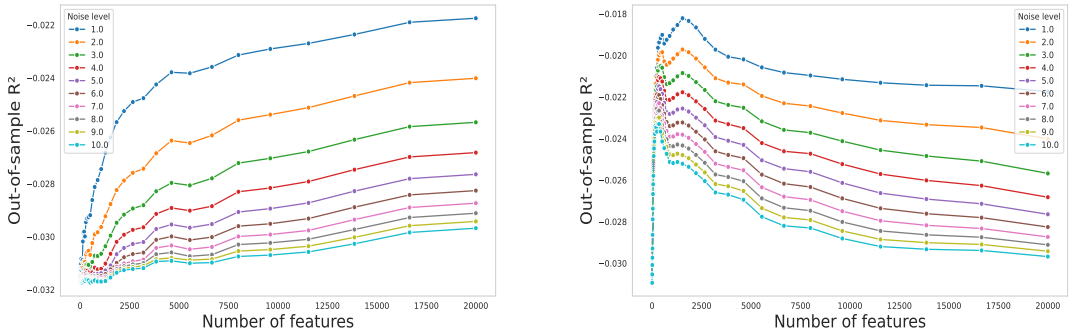


Figure 3: Complexity and model performance with simulated features observed with noise. In the left-hand panel, the noise level in each feature remains fixed as model complexity increases, whereas in the right-hand panel, the noise level for each feature increases linearly as model complexity increases, 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 one use a large set of predictive features when the "true features" are imperfectly observed due to infrastructure limitations?* To answer this question, this paper 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 favours 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.

In contrast to existing studies which assume that features are known with precision or that the dimensionality of the feature model is moderate, we allow both the number of features and the sample size to become large, and observe the features with measurement error. In doing so, we derive how the interaction between noise in the features and the complexity of the model affect the out-of-sample R-squared of the stock return predictions and the Sharpe ratio of a portfolio, where the weight assigned to the risky asset is determined by an investment strategy that uses return predictions.

Our work makes two main contributions. First, we show that the out-of-sample R-squared and Sharpe ratio of a portfolio investment strategy are *monotonically decreasing and convex* as the noise level of the features increases. This theoretical finding suggests that investors with high-quality data and advanced infrastructure can obtain greater benefits from incremental

improvements in their infrastructure to reduce the measurement error in features. In contrast, investors facing higher measurement error due to limited data resources or less sophisticated infrastructure may find that such improvements yield diminishing returns. Importantly, this distinction highlights how the interaction between feature quality and infrastructure impacts the power of predictive models.

Second, we study a *partial model* in high-dimensional settings when not all features are observed. Here, an investor observes a subset of the true features, and features in this subset may be observed with noise. The investor can choose to develop more complex models which requires the investor to learn a larger subset of features. In this case, 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 noisy data degrade feature quality. This result complements the "virtue of complexity" with a pragmatic view of the optimal level of complexity beyond which performance of the predictions and portfolio deteriorates.

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 our paper, 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, Dobriban and Wager (2018), Wu and Xu (2020), and Hastie et al. (2022) detail when, under certain assumptions, the out-of-sample performance of overparameterized or "ridgeless" models is good. Similarly, Gu et al. (2021) and Jiang et al. (2023) show that advanced learning architectures can

significantly improve out-of-sample forecasts. The contribution of this paper 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.

Recent innovations in high-dimensional econometrics and empirical asset pricing have driven a proliferation of candidate features and anomalies (Harvey et al., 2016; Feng et al., 2020). Many of the proposed features rely on data sources subject to reporting lags or other data imperfections, which are studied extensively in macroeconomics and market microstructure (Chaboud et al., 2014). 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 how data imperfections interact with model dimensionality.

Our study also connects to broader questions about partial observability and misspecified feature structures in finance. Classic works, including Pesaran and Timmermann (1995) and Campbell et al. (1998), show that excluding relevant variables can harm predictability and portfolio performance, while more recent studies (e.g., Kacperczyk et al., 2014) 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.

The remainder of the paper is organized as follows. Section 2 presents the data-generating process (DGP) for noisy features, and it details how to introduce noise into the feature space and establishes our main modeling assumptions. Sections 3 and 4 develop the ridge-regularized estimator in high-dimensional environments, characterizing the asymptotic behavior of out-of-sample R-squared and the Sharpe ratio of a dynamic portfolio investment strategy under both correct specification and model misspecification. Section 5 provides an empirical application with U.S. stock market data to illustrate the practical implications of our theoretical results. Section 6 concludes and discusses directions for future research on noisy complexity in asset pricing and beyond.

2 Model

2.1 Asset dynamics

Consider a financial market where the DGP of the **excess** returns of a single asset is given by the predictive model

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

Here, $F_t \in \mathbb{R}^{P\times 1}$ is the true feature matrix, $\beta \in \mathbb{R}^{P\times 1}$ 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.

The investors use historical data to estimate the predictive coefficients β and to construct optimal portfolios. Unlike the literature, where they assume that the investor observes the true feature matrix F_t, here, the investors observe a noisy version $\tilde{F}_t$ of the true feature due

to imperfections in the investor's infrastructure. Therefore, the investors observe the model

$$R_{t+1} = \tilde{F}_t^\intercal \, \beta + \tilde{\epsilon}_{t+1} \,, \tag{2}$$

where $\tilde{\epsilon}_{t+1} = \epsilon_{t+1} + F_t^\intercal \, \beta - \tilde{F}_t^\intercal \, \beta$, so they 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 (1) 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 satistfies

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

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 $\mathbb{E}\left[X_{i,t}\right] = \mathbb{E}\left[X_{i,t}^3\right] = 0$, $\mathbb{E}\left[X_{i,t}^2\right] = 1$, and the fourth moments $\mathbb{E}\left[X_{i,t}^4\right]$ are uniformly bounded.

(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.

(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 $\mathbb{E}[\beta\,\beta^\intercal] = P^{-1}\,b_{*,P}\,I_P$ for some constant $b_{*,P} = \mathbb{E}[||\beta||^2]$, where $b_{*,P} \to b_*$ in probability for some $b_* > 0$, and I_P is the P-dimensional identity matrix. Furthermore, $\mathbb{E}[\beta_i^4] \leq K\,P^{-2}$ for some $K > 0$.

Assumptions 1(a), (b), and (d) are standard in the literature; see, for example, Hastie et al. (2022), Wu and Xu (2020), and Dobriban and Wager (2018)). Assumption 1(d) states that the predictive coefficients β are random and rotationally symmetric with i.i.d. coordinates, ensuring uniform predictability across features, similar to neural network structures.

Assumption 1(c) restricts the covariance structure of the noisy feature matrix $\tilde{F}_t$ to ensure it is an unbiased estimator of F_t. Additionally, $\mathbb{E}\left[\tilde{F}_t\,\tilde{F}_t^\intercal\right] = \Psi + \Phi$ and $\mathbb{E}\left[\tilde{F}_t\,F_t^\intercal\right] = \Psi$ because X_t and Y_t are independent. Independence between X_t and Y_t is not consequential, but we make this assumption for clearer expressions and conciseness.

We make an additional assumption on the distribution of the eigenvalues of the 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 k*th 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-square of portfolios in high-dimensional settings where $P, T \to \infty$ and $P/T \to c > 0$.

The assumptions 1 and 2 lead us to the following lemma (Kelly et al., 2024b).

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

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

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

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.

2.2 Portfolio: timing strategy and performance evaluation

The investors use the noisy features $\tilde{F}_t$ in (2) 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-square 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., Campbell et al. (1998), Pesaran and Timmermann (1995).

From this point onward, all expectations are conditional on the random variable β, without explicitly stated. In particular, under Assumption 1, the future return R_{t+1} is independent of the current return R_t conditional on β.

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., Merton (1981) and Henriksson and Merton (1981).

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

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

and recall that in (1) we work with excess returns.

The goal is to maximize the unconditional Sharpe ratio

$$\mathrm{SR} = \mathbb{E}[R^{\pi}_{t+1}] \Big/ \sqrt{\mathbb{E}[(R^{\pi}_{t+1})^2]} \,. \tag{3}$$

In (3), we use the second moment $\mathbb{E}[(R^{\pi}_{t+1})^2]$ instead of the variance $\text{Var}[R^{\pi}_{t+1}]$. 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, and they argue that using the second moment does not affect the optimal portfolio weight π_t.

Hansen and Richard (1987) show that under the true DGP in (1), the optimal weight π_t that maximizes the unconditional Sharpe ratio (3) is

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

With Assumptions 1 and 2, KMZ show that $(F_t^{\intercal}\,\beta)^2 \to b_*\,\psi_{*,1}$ 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

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

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 $\pi_t^{\mathbf{c}}$ because she only observes the noisy model (2) instead of the true model (1), and hence she does not observe the true predictive coefficients β. Therefore, the investor uses

$$\hat{\pi}_t = \tilde{F}_t^{\intercal}\,\hat{\beta} \tag{4}$$

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.

2.2.2 Out-of-sample R-square

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 (1). 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)

$$\text{MSE}(\hat{\beta}) = \mathbb{E}\left[\left(R_{t+1} - \tilde{F}_t^{\intercal}\,\hat{\beta}\right)^2 \mid \hat{\beta}\right] = \mathbb{E}\left[R_{t+1}^2\right] - 2\,\mathbb{E}\left[\hat{\pi}_t\,R_{t+1} \mid \hat{\beta}\right] + \mathbb{E}\left[\hat{\pi}_t^2 \mid \hat{\beta}\right]. \tag{5}$$

where the expectation in (5) is also conditional on β; therefore, the first term in the decomposition in (5) 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 (5) are the conditional expected return and the conditional squared value of the feasible timing strategy. Therefore, we refer to these two terms 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 exposure to an asset based on timing signals. By varying leverage, the strategy aims to exploit perceived short-term opportunities. The out-of-sample R-square of the prediction in predictive coefficients is given by

$$\mathrm{R}^2(\hat{\beta}) = 1 - \frac{\mathrm{MSE}(\hat{\beta})}{\mathbb{E}\left[R_{t+1}^2\right]}\,. \tag{6}$$

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 (1) is $N(0,1)$. Then, if she were to observe the true DGP in (1), the posterior mean of her estimation of the predictive coefficients would be

$$\mathbb{E}[\beta(z)] = \hat{\beta}(z) = \left(z\,I_P + T^{-1}\sum_{\tau} F_\tau\,F_\tau^\intercal\right)^{-1} T^{-1}\sum_{\tau} F_\tau\,R_{\tau+1}\,, \tag{7}$$

which coincides with the estimator of a ridge-regularized least square estimator, see Lindley and Smith (1972).

However, the investor observes the noisy model in (2), so her estimate of the predictive coefficients is

$$\hat{\beta}(z) = \left(z\, I_P + T^{-1} \sum_{\tau} \tilde{F}_\tau\, \tilde{F}_\tau^\intercal \right)^{-1} T^{-1} \sum_{\tau} \tilde{F}_\tau\, R_{\tau+1}\,. \tag{8}$$

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., Giannone et al. (2021). 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 paper, we use Ridge regression instead of LASSO regression.

The ridge parameter $z_* = c/b_*$ maximizes the out-of-sample R-square and the Sharpe ratio of the timing strategy $\hat{\pi}_t$ in (4) when the investor observes the true DGP in (1), 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 (2), so the optimal ridge parameter z_* takes a more complicated form.

4 Theoretical properties of ridge regression

In this section, we use the ridge estimator in (8) to study both the asymptotic behavior of the unconditional Sharpe ratio in (3) and the out-of-sample R-square in (6) of the feasible timing strategy. Recall that the feasible timing strategy is a function of a noisy version of the DGP in (2).

4.1 Random matrix theory

To understand the limiting behavior of the ridge estimator in (8) 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

$$\begin{aligned} P^{-1} \operatorname{tr}\left((\tilde{\Psi} - u\, I)^{-1}\right) &= P^{-1} \sum_{i=1}^{P} \left(\lambda_i(\tilde{\Psi}) - u\right)^{-1} \\ &= \int \frac{1}{x-u}\, d\boldsymbol{F}^{\tilde{\Psi}}(x)\,, \quad u < 0\,, \end{aligned} \tag{9}$$

where $\lambda_i(\tilde{\Psi})$ are the eigenvalues of $\tilde{\Psi}$, and $\operatorname{tr}(\,\cdot\,)$ is the trace operator. The right-hand side of (9) 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

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

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

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

The following lemma from Bai and Zhou (2008) and Silverstein and Bai (1995) 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*

$$m(u;c) = \frac{1}{1 - c - c\,u\,m(z;c)}\, m_{\tilde{\Psi}}\left(\frac{u}{1 - c - c\,u\,m(u;c)}\right). \tag{10}$$

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 $P, T \to \infty$ and $P/T \to c$, for any $z > 0$, we have*

$$\lim_{T\to\infty} T^{-1}\,\mathrm{tr}\left(\left(z\,I_P + \hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\right) \to \xi(z;c) \tag{11}$$

in probability, where

$$\xi(z;c) = \frac{1 - z\,m(-z;c)}{c^{-1} - 1 + z\,m(-z;c)}. \tag{12}$$

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 (1) is observed and $\xi(z;c)$ is known, one can derive asymptotic results for the R-square 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 (2), 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 (11) insufficient for our results. Below, we derive two useful limits that depend on the covariance between $\tilde{F}_t$ and F_t.

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

$$\xi_\Psi(z;c) = \lim_{T\to\infty} T^{-1}\, \mathrm{tr}\left(\mathbb{E}\left[\left(z\, I_P + \hat{\tilde{\Psi}}_T\right)^{-1} \Psi\, \Psi^\intercal\right]\right) \geq 0\,, \tag{13}$$

$$\eta_\Psi(z;c) = \lim_{T\to\infty} T^{-1}\, \mathrm{tr}\left(\mathbb{E}\left[\left(z\, I_P + \hat{\tilde{\Psi}}_T\right)^{-1} \tilde{\Psi} \left(z\, I_P + \hat{\tilde{\Psi}}_T\right)^{-1} \Psi\, \Psi^\intercal\right]\right) \geq 0\,. \tag{14}$$

The forms of the functions $\xi_\Psi(z;c)$ and $\eta_\Psi(z;c)$ are similar to that of $\xi(z;c)$ because they both involve limits that play a role in the estimation of the covariance matrix between $\tilde{F}_t$ and F_t. In contrast, $\xi(z;c)$ helps to estimate the covariance matrix of $\tilde{F}_t$. However, unlike for $\xi(z;c)$, in general it is very difficult to give explicit expressions for $\xi_\Psi(z;c)$ and $\eta_\Psi(z;c)$. In fact, the convergence of the traces to $\xi_\Psi(z;c)$ and $\eta_\Psi(z;c)$ is not guaranteed, in which case one needs to use a converging subsequence of ξ_Ψ and η_Ψ to obtain meaningful results.

4.2 Asymptotic results

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

Theorem 1. *Let Assumptions 1, 2, and 3 hold. Then, as $P, T \to \infty$ and $P/T \to c$, the*

expected return of the timing strategy is given by

$$\mathcal{E}(z;c) = \lim \mathbb{E}\left[\hat{\pi}_t(z)\, R_{t+1} \mid \hat{\beta}\right] = \frac{b_*\, c^{-1}\, \xi_\Psi(z;c)}{1+\xi(z;c)}\,, \tag{15}$$

and the expected leverage of the timing strategy converges to

$$\begin{aligned}\mathcal{L}(z;c) &= \lim \mathbb{E}\left[\hat{\pi}_t(z)^2 \mid \hat{\beta}\right] \\ &= (1+b_*\,\psi_{*,1})\,(\xi(z;c) + z\,\partial\xi(z;c)/\partial z) + \frac{b_*\, c^{-1}\, \eta_\Psi(z;c)}{(1+\xi(z;c))^2} \\ &\quad - \frac{2\, b_*\, c^{-1}\ (\xi(z;c) + z\,\partial\xi(z;c)/\partial z)\ \xi_\Psi(z;c)}{1+\xi(z;c)}\,,\end{aligned} \tag{16}$$

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

$$R^2(z;c) = \frac{2\,\mathcal{E}(z;c) - \mathcal{L}(z;c)}{1+b_*\,\psi_{*,1}}\,. \tag{17}$$

The conditional second moment of the feasible timing strategy return $\mathbb{E}\left[\left(\hat{\pi}_t(z)\, R_{t+1}\right)^2 \mid \hat{\beta}\right]$ *converges to*

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

and the Sharpe ratio of the feasible timing strategy is

$$SR(z;c) = \frac{\mathcal{E}(z;c)}{\sqrt{2\,\mathcal{E}(z;c)^2 + (1+b_*\,\psi_{*,1})\ \mathcal{L}(z;c)}}\,. \tag{18}$$

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-square and Sharpe ratio of the timing strategy. These results are not easy to interpret. In particular, (17) and (18) 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-square 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 $\eta_\Psi(z; c)$.

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 (Harvey et al., 2016). 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 Marchenko and Pastur (1967). 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

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

Similarly, one can write ξ_Ψ and η_Ψ as

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

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

Substituting the expressions into (15) and (16), 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

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

Next, with

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

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

$$R^2(z;c;a) = \frac{2\,\mathcal{E}(z;c;a) - \mathcal{L}(z;c;a)}{1+b_*}\,, \tag{19}$$

and

$$SR(z;c;a) = \frac{\mathcal{E}(z;c;z)}{\sqrt{2\,\mathcal{E}(z;c;a)^2 + (1+b_*)\;\mathcal{L}(z;c;a)}}\,. \tag{20}$$

Now, with (19) and (20), we derive further results regarding the optimal choice for the ridge parameter z and the properties of the R-square and Sharpe ratio as a function of the noise level a.

Proposition 2. *Let the assumptions in Corollary 1 hold. Then,*

(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-square $R^2(z;c;a)$ 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

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

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.

4.4 A note on convexity

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 the 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 investor's efforts into hiring staff,

acquiring information, and building better models. Therefore, we assume that 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-square, 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 on 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 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.

4.5 Model misspecification

In KMZ, the authors discuss a model misspecification scenario where the investor cannot observe the full set of features and the oberseved ones are without measurement error. Suppose that the DGP is given by (1) with P features. The investor observes a subset P_1 of these features. Denote the observed 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

$$\hat{\pi}_{1,t} = F_{t,(1)}^{\intercal}\, \hat{\beta}_1\,. \tag{21}$$

KMZ prove that, under certain regularity conditions, the out-of-sample R-squared and Sharpe ratio of (21), 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

$$\hat{\tilde{\pi}}_{1,t} = \tilde{F}_{t,(1)}^{\intercal}\, \hat{\tilde{\beta}}_1\,. \tag{22}$$

We analyze the properties of the out-of-sample R-squared and Sharpe ratio of (22) 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.[5] 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*

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

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

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

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

$$R^2(z; c; a; q) = \frac{2\,\mathcal{E}(z; c; a; q) - \mathcal{L}(z; c; a; q)}{1+b_*}\,, \tag{23}$$

and

$$SR(z; c; a; q) = \frac{\mathcal{E}(z; c; a; q)}{\sqrt{2\,\mathcal{E}(z; c; a; q)^2 + (1+b_*)\;\mathcal{L}(z; c; a; q)}}\,. \tag{24}$$

[5] An example of this assumption would be expanding the width of a deep neural network, e.g., Figure 2. When increasing the width of a neural network, one has to re-estimate the weight on all the neurons; hence, when increasing the dimension of a feature model, the aggregate level of noise increases with the number of features. 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-square 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,*

(i) The out-of-sample R-square $R^2(z;c;a;q)$ and Sharpe ratio $SR(z;c;a;q)$ of the timing strategy are both 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-square $R^2(z;c;a;q)$ 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 where the Sharpe ratio of our timing strategy is maximized.

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, the investors are incentivized to discover more features and ultimately approach a full-information model where 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 the investors no longer seek to find more features beyond this level of complexity.[6]

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 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.

5 Empirical analysis

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

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 Welch and Goyal (2008), available monthly from 1926 to 2023.

Following KMZ, we standardize the monthly returns by dividing them by a rolling 12-month standard deviation of the montly returns, where the standard deviation is calculated from the uncentered second moment. To standardize the predictors, we adopt an expanding window approach: at each point in time, all available historical observations up to that date

[6]One can prove that there is an optimal level of complexity also exists 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.

are used to calculate the standard deviation of the predictors. To ensure stability in these calculations, we require at least 36 months of data before starting our analysis, which begins in 1930.

5.2 Noisy features

Empirically, we analyze the noisy DGP (2) in two steps. First, we examine how empirical results (i.e., out-of-sample Sharpe ratio and R-squared) change from low-dimensional cases where $P \ll T$ to high-dimensional cases where $P \gg T$. To investigate this, as in KMZ, we use Random Fourier Features (RFF) (Rahimi and Recht (2007), Rahimi and Recht (2008)). Let G_t represent our 15×1 vector of predictors. The RFF method transforms G_t into a new set of features

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

for $i = 1, \ldots, P/2$. 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 κ 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 $P/2$, so the total number of features equals to P because each i yields two features (the sine and cosine components).[7]

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

[7]Here P is even because we generate a pair of RFFs for each i.

value of P to generate arbitrarily small and large sets of features to explore low-dimensional and high-dimensional settings in our analysis.

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

$$\tilde{G}_t = G_t + a\, Y_t\,, \quad Y_t \sim \mathrm{N}(0, I_{15})\,, \tag{25}$$

so the corresponding noisy features are

$$\tilde{F}_{i,t} = \left[\sin\left(\kappa\, \omega_i^\intercal\, \tilde{G}_t\right),\, \cos\left(\kappa\, \omega_i^\intercal\, \tilde{G}_t\right)\right]^\intercal, \quad \omega_i \sim \text{ i.i.d. } \mathrm{N}(0, I_{15})\,. \tag{26}$$

Here, when $a > 0$, the noise arises from limited access to high-quality data by investors.[8]

The second approach adds noise after generating the features. Here, we first use the noiseless G_t to create the RFF features, $F_{i,t}$. Then, we add noise $\mu_{i,t}$ and $\eta_{i,t}$ to the generated features, so

$$\begin{aligned} &\tilde{F}_{i,t} = \left[\sin\left(\kappa\, \omega_i^\intercal\, G_t\right) + a\, \mu_{i,t},\, \cos\left(\kappa\, \omega_i^\intercal\, G_t\right) + a\, \eta_{i,t}\right]^\intercal, \\ &(\omega_i^\intercal, \mu_{i,t}, \eta_{i,t})^\intercal \sim \text{ i.i.d. } \mathrm{N}(0, I_{17})\,. \end{aligned} \tag{27}$$

As above, a represents the noise level. This approach models a scenario where noisy observations of the features stem from the skills and infrastructure of the investors. Investors with better technical skills (e.g., machine learning expertise and financial knowledge) and superior

[8]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.

infrastructure (e.g., greater computing resources and efficient coding practices) experience less noise in their features, even when using the same data.[9]

We use one-year rolling training windows ($T = 12$ months) and the number of RFFs is $P \in \{2, 100, 500, 1000, 2000, 4000\}$. 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. 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 (25) to add noise directly to the observation G_t and generate $\tilde{G}_t$, then use (26) to create 4000 RFFs. In the second approach, use (27) to generate 4000 RFFs where the noise is added after one generates the features. The bandwidth parameter is $\kappa = 2$, 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 parameter $\log_{10} z \in \{-3, \ldots, 5\}$.[10] The predictors $\tilde{F}_t$ for the regression in (2) are constructed by selecting P features from the set of RFFs generated in step (i).

(iii) **Out-of-sample predictions**: We use the estimates from step (ii) to perform

[9] We introduce noise as discussed in Li et al. (2020), where they suggest that the noise may already exist in the data, or it may be in the features.

[10] Our theoretical findings suggest that the optimal level of shrinkage should be high in settings where the value of $c = P/T$ is large.

recursive out-of-sample predictions and to implement a market timing strategy. For each $t \in \{12, \ldots, 1127\}$months, we estimate (2) with the training data

$$\left\{ R_\tau, \tilde{F}_{\tau-1}, \tau = t, \cdots, t-T+1 \right\}, \quad T = 12\,\text{months}.$$

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-squared, 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).

5.3 Out-of-Sample performance of market timing strategy

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 for shrinkage levels of $\log_{10} z = 0$, $\log_{10} z = 3$, and $\log_{10} z = 4$, see Figures 4, 5, and 6.

The average return and Sharpe ratio of the strategy exhibit nearly convex trends in Figure 4, 5, and 6, 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 feature is compromised when the ridge regularization is insufficient. However, as shrinkage increases to $\log_{10} z = 3$ and $\log_{10} z = 4$, more convex

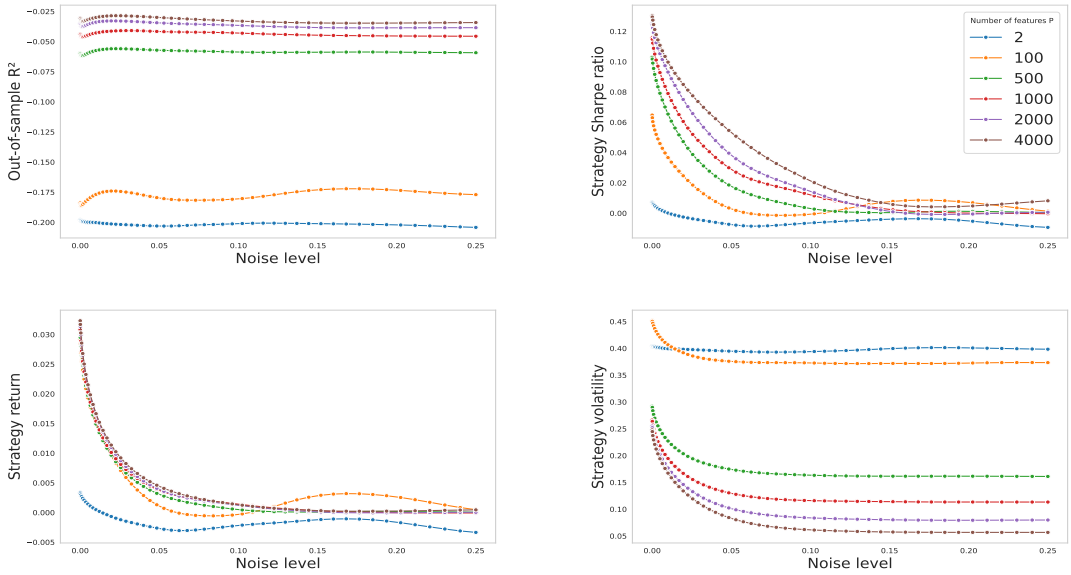


Figure 4: Out-of-sample market timing performance with ridge parameter $\log_{10} z = 0$. Training window is $T = 12$ months and the number of RFFs, P ranges from 2 to 4,000 with $\kappa = 2$, $\log_{10} z = 0$, and noise variance ranging from 0 to 0.25.

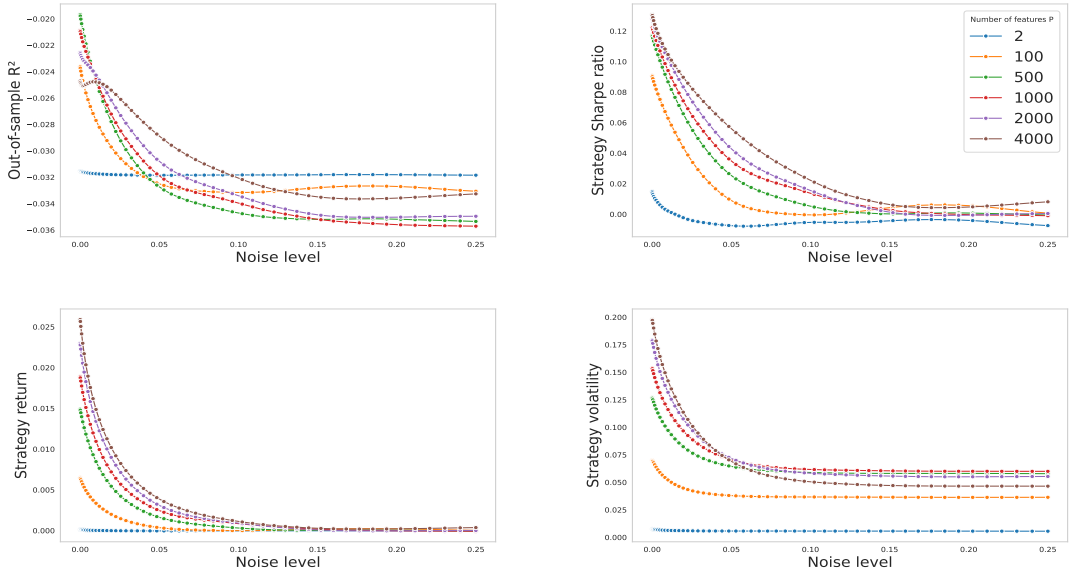


Figure 5: Out-of-sample market timing performance with ridge parameter $\log_{10} z = 3$.

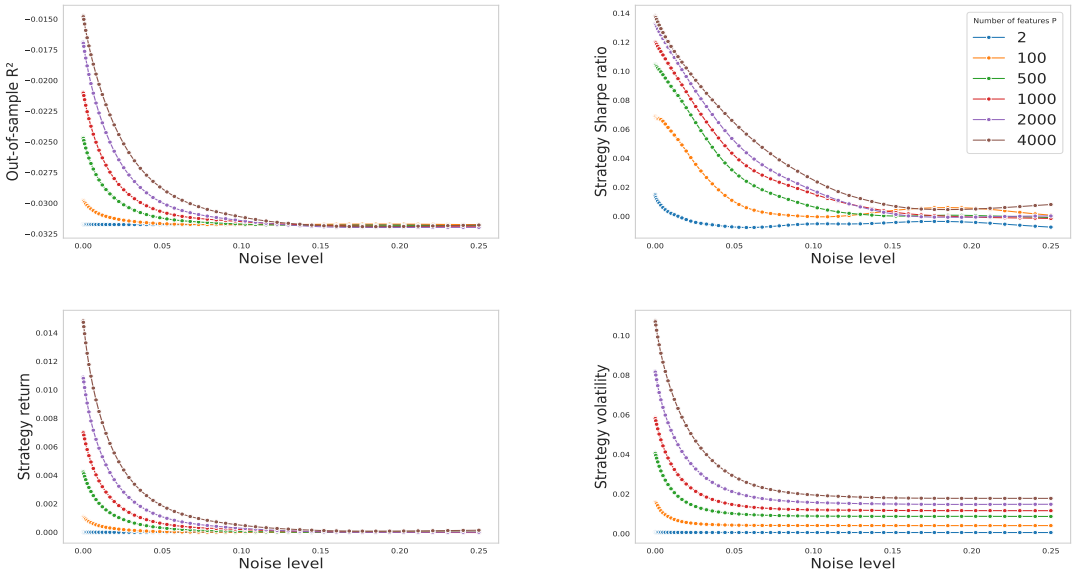


Figure 6: Out-of-sample market timing performance with ridge parameter $\log_{10} z = 4$.

patterns emerge in the R-squared curves. This indicates that regularization helps larger models perform better when dealing with noise, which aligns with our theoretical results that the optimal shrinkage z_* should be large when model complexity c is large.

To further understand the role of shrinkage, we use a three-fold cross-validation procedure to find the optimal ridge parameter $\log_{10} z$. Specifically, for each number P of RFFs and a grid of candidate values for $\log_{10} z$, we compute the cross-validated out-of-sample performance (both R-squared and Sharpe ratio) and select the $\log_{10} z$ that yields the highest average value. Figure 7 shows the optimal R-squared and Sharpe ratio as functions of the noise level. Compared with Figures 4, 5, and 6, we observe a more convex behavior across different P. Moreover, for all P, the maximum R-squared and Sharpe ratio exhibit a monotonically decreasing and convex pattern in the noise level, underscoring the importance of appropriate regularization to mitigate noise while allowing the model to leverage higher complexity.

Next, we focus on the approach where noise is added into the generated features. When

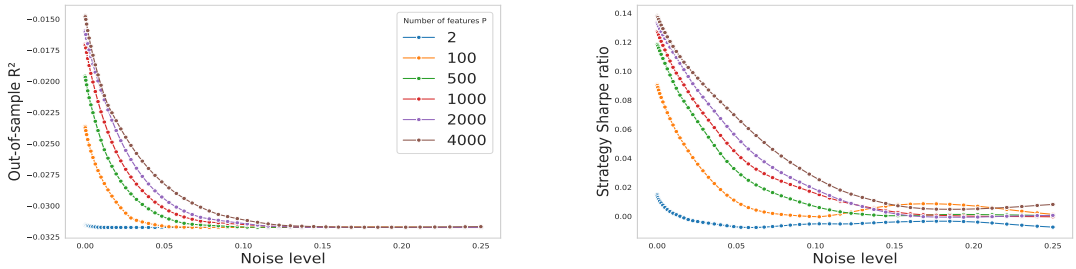


Figure 7: Optimal out-of-sample R-squared and Sharpe ratio for various noise levels added to the data, where the ridge parameter is chosen via three-fold cross-validation for each number P of RFFs.

the value of the ridge shrinkage is 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 cross validated for various shrinkage levels.

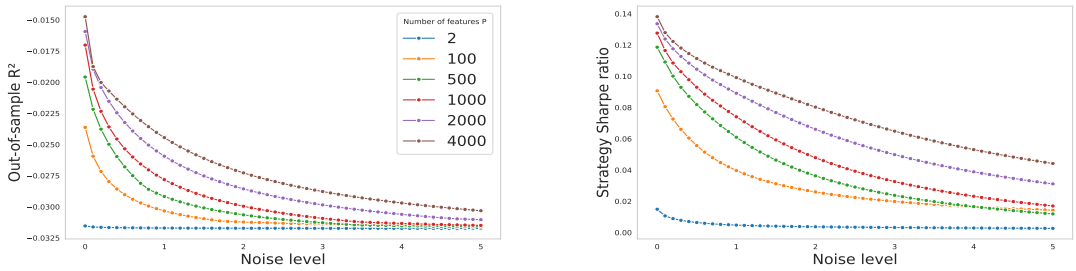


Figure 8: Optimal out-of-sample R-squared and Sharpe ratio for various noise levels added to the feature, where the ridge parameter is chosen via three-fold cross-validation for each number of RFF P.

Figure 8 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 with 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 (2). By adding noise directly to the "true" features, we simulate a scenario where noise affects the underlying feature 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.

6 Conclusion

In this paper, we incorporate noisy features into the analysis of complexity in portfolio optimization. We showed that the performance of portfolio, when features are observed with noise, can be improved through appropriate regularization techniques, such as ridge regression.

We identified a nuanced interplay between noise and performance, which we refer to as the "limited virtue of complexity." While complexity has the potential to enhance portfolio returns, its benefits depend on access to high-quality data and robust infrastructure. Sophisticated investors, equipped with advanced systems and better data resources, gain the most from investing in greater complexity. Conversely, for less sophisticated investors, the returns on such investments diminish as noise overwhelms feature quality.

Additionally, we showed that when investors have access to a partial model only, the relationship between noise and complexity becomes more critical. In such cases, there exists an optimal level of model complexity that balances the trade-off between pursuing more features and mitigating the impact of noise. While discovering additional features can initially improve performance, excessive complexity in the presence of limited data and high noise

levels may lead to feature degradation, ultimately harming portfolio returns. This underscores the importance of carefully calibrating the model's complexity to the quality of available data and computational capacity, particularly for investors constrained by incomplete models or limited resources.

References

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

Bai, Z. and Zhou, W. (2008). Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, pages 425–442.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Bertsekas, D. P. (1997). Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Brogaard, J. and Zareei, A. (2023). Machine learning and the stock market. *Journal of Financial and Quantitative Analysis*, 58(4):1431–1472.

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

Campbell, J. Y., Lo, A. W., MacKinlay, A. C., and Whitelaw, R. F. (1998). The econometrics of financial markets. *Macroeconomic Dynamics*, 2(4):559–562.

Cao, S., Jiang, W., Wang, J., and Yang, B. (2024). From man vs. machine to man+ machine: The art and ai of stock analyses. *Journal of Financial Economics*, 160:103910.

Chaboud, A. P., Chiquoine, B., Hjalmarsson, E., and Vega, C. (2014). Rise of the machines: Algorithmic trading in the foreign exchange market. *The Journal of Finance*, 69(5):2045–2084.

Chen, L., Pelger, M., and Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, 70(2):714–750.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*.

Dobriban, E. and Wager, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279.

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

Feng, G., He, J., Polson, N. G., and Xu, J. (2024). Deep learning in characteristics-sorted factor models. *Journal of Financial and Quantitative Analysis*, 59(7):3001–3036.

Giannone, D., Lenza, M., and Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437.

Giglio, S., Kelly, B., and Xiu, D. (2022). Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics*, 14(1):337–368.

Green, J., Hand, J. R., and Zhang, X. F. (2017). The characteristics that provide independent information about average us monthly stock returns. *The Review of Financial Studies*, 30(12):4389–4436.

Gu, S., Kelly, B., and Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics*, 222(1):429–450.

Gu, S., Kelly, B. T., and Xiu, D. (2019). Empirical asset pricing via machine learning. *Chicago Booth Research Paper*, (18-04):2018–09.

Guijarro-Ordonez, J., Pelger, M., and Zanotti, G. (2021). Deep learning statistical arbitrage. *arXiv preprint arXiv:2106.04028*.

Hansen, L. P. and Richard, S. F. (1987). The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models. *Econometrica*, 55(3):587–613.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). ... and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Heaton, J., Polson, N. G., and Witte, J. H. (2016). Deep learning in finance. *arXiv preprint arXiv:1602.06561*.

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

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jiang, J., Kelly, B., and Xiu, D. (2023). Re - imag (in) ing price trends. *The Journal of Finance*, 78(6):3193–3249.

Kacperczyk, M., Nieuwerburgh, S. V., and Veldkamp, L. (2014). Time-varying fund manager skill. *The Journal of Finance*, 69(4):1455–1484.

Kelly, B., Kuznetsov, B., Malamud, S., and Xu, T. A. (2024a). Large (and deep) factor models. *arXiv preprint arXiv:2402.06635*.

Kelly, B., Malamud, S., and Zhou, K. (2024b). The virtue of complexity in return prediction. *The Journal of Finance*, 79(1):459–503.

Leippold, M., Wang, Q., and Zhou, W. (2022). Machine learning in the chinese stock market. *Journal of Financial Economics*, 145(2):64–82.

Li, Z., Su, W., and Sejdinovic, D. (2020). Benign overfitting and noisy features. *arXiv preprint arXiv:2008.02901*.

Lindley, D. V. and Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 34(1):1–18.

Marchenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536.

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

Pesaran, M. H. and Timmermann, A. (1995). Predictability of stock returns: Robustness and economic significance. *The Journal of Finance*, 50(4):1201–1228.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20.

Rahimi, A. and Recht, B. (2008). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 21.

Silverstein, J. W. and Bai, Z. D. (1995). On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):175–192.

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

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

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

Wu, D. and Xu, J. (2020). On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123.

A Proofs

A.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* $X_j, \mathbb{E}[X_i] = 0, \mathbb{E}[X_i^2] = 1, \mathbb{E}[X_i^4] \leq k$, *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 equation* (32) *below), and hence*

$$\lim_{P\to\infty} \mathbb{E}\left[\left(X_t^\intercal B_P X_t - \operatorname{tr}(B_P)\right)^2\right]/P^2 = 0\,. \tag{28}$$

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*

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

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*

$$P^{-1}\operatorname{tr}\left(Q_P\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\right) - \mathbb{E}\left[P^{-1}\operatorname{tr}\left(Q_P\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\right)\right] \to 0 \tag{29}$$

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 Bai and Zhou (2008). 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

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

where we use (40) and defined

$$l_t = \operatorname{tr}\left(\frac{1}{T}\left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{F}_t \left(I + \frac{1}{T}\,\tilde{F}_t^{\intercal}\left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{F}_t\right)^{-1} \tilde{F}_t^{\intercal}\left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} Q_P\right). \tag{30}$$

Next, use Lemma 4 to write

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

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,

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

forms a bounded martingale difference sequence. Apply the Burkholder–Davis–Gundy inequality (see, for example, Burkholder (1966)) to obtain

$$\mathbb{E} \left[|q_T(z) - \mathbb{E} \left[q_T(z) \right]|^q \right] \leq K_q \, P^{-q} \, \mathbb{E} \left(\sum_{t=1}^{T} \left| \left(\mathbb{E}_{t-1} - \mathbb{E}_t \right) [l_t] \right|^2 \right)^{q/2}$$
$$\leq K_q \, (2\, K)^q \, P^{-q/2} \left(\frac{P}{T} \right)^{-q/2} . \tag{31}$$

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

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

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,

$$\mathrm{tr}\left((A\,B + B\,A)^2\right) = \mathrm{tr}\left(\Lambda^2\right) \leq P\,\lambda_{\max}(\Lambda)^2\,, \tag{32}$$

where $\lambda_{\max}(\Lambda)$ is the largest eigenvalue of Λ, which is also the largest eigenvalue of $A\,B + B\,A$. Basic matrix norm inequality shows that

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

Therefore,

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

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,

$$\begin{aligned} \left(z\, I + \hat{\tilde{\Psi}}_T \right)^{-1} \tilde{F}_t &= \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \tilde{F}_t - \frac{1}{T} \frac{\left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \tilde{F}_t \, \tilde{F}_t^\intercal \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \tilde{F}_t}{1 + (T)^{-1} \, \tilde{F}_t^\intercal \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \tilde{F}_t} \\ &= \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \tilde{F}_t \, \frac{1}{1 + (T)^{-1} \, \tilde{F}_t^\intercal \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \tilde{F}_t} \,. \end{aligned} \tag{33}$$

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

$$P^{-1} \operatorname{tr} \left(\tilde{\Psi} \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \right) - \mathbb{E} \left[P^{-1} \operatorname{tr} \left(\tilde{\Psi} \left(z\, I + \hat{\tilde{\Psi}}_{T,t} \right)^{-1} \right) \right] \to 0\,. \tag{34}$$

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

$$\begin{aligned}
1 &= P^{-1} \operatorname{tr}\left(\mathbb{E}\left[\left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\left(z\, I + \hat{\tilde{\Psi}}_T\right)\right]\right) \\
&= P^{-1}\ \operatorname{tr}\left(\mathbb{E}\left[\left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\right] z\right) + P^{-1}\ \operatorname{tr}\left(\mathbb{E}\left[\left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \hat{\tilde{\Psi}}_T\right]\right) \\
&= z\, m(-z, c) + P^{-1} \operatorname{tr}\left(\mathbb{E}\left[\left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \frac{1}{T} \sum_t \tilde{F}_t \tilde{F}_t^{\intercal}\right]\right) \\
&= z\, m(-z, c) + P^{-1} \operatorname{tr}\left(\mathbb{E}\left[\left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1} \tilde{F}_t \tilde{F}_t^{\intercal}\right]\right) \\
&= z\, m(-z, c) + P^{-1} \operatorname{tr}\left(\mathbb{E}\left[\left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{F}_t \frac{1}{1 + T^{-1}\, \tilde{F}_t^{\intercal}\left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{F}_t} \tilde{F}_t^{\intercal}\right]\right) \\
&= z\, m(-z, c) + \mathbb{E}\left[\frac{P^{-1} \tilde{F}_t^{\intercal}\left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{F}_t}{1 + T^{-1}\, \tilde{F}_t^{\intercal}\left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} \tilde{F}_t}\right]\,. \qquad (35)
\end{aligned}$$

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 (35) and (34),

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

in probability and this sequence is uniformly bounded. Hence,

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

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.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 Vershynin (2010).

Theorem 2. *(Theorem 5.44 in Vershynin (2010)). 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*

$$\left\|\frac{1}{N}\,A^{\intercal}\,A-\Sigma\right\|\le\max\left(\|\Sigma\|^{1/2}\,\delta,\,\delta^2\right),\qquad \textit{where}\qquad \delta=s\,\sqrt{\frac{m}{N}} \tag{36}$$

holds with probability at least $1 - n \cdot \exp(-c_1 s^2)$. *Here,* $c_1 > 0$ *is an absolute constant. In particular,* (36) *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,

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

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

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

and hence

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

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

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

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

$$\begin{aligned} \mathbb{P}\left(\lambda_{\max}\left(\hat{\tilde{\Psi}}_T / P\right) > \epsilon\right) &\le \mathbb{P}\left(\exists i, t, |Z_{t,i}| > \sqrt{m}\right) + \mathbb{P}\left(\left\| \frac{1}{P\,T} \sum_t Z_t\, Z_t^\intercal \right\| > \frac{C\,(\log P)^3}{P}\right) \\ &\le \frac{K}{(\log P)^2} + \delta_1 \to \delta_1 \,. \end{aligned} \tag{37}$$

Because (37) 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, $\mathbb{E}[U_i] = 0$, $\mathbb{E}[U_i^2] = 1$, $\mathbb{E}[U_i^4] \le k$. *Suppose further that* U_t *is independent from* $\hat{\tilde{\Psi}}_T$, *then we have*

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

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

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

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

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

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

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

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

thus

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

Therefore, under $\mathcal{A}$,

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

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

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

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*

$$P^{-1}\,\mathrm{tr}\left(Q_P\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-2}\right) - \mathbb{E}\left[P^{-1}\,\mathrm{tr}\left(Q_P\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-2}\right)\right] \to 0 \tag{38}$$

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 Bai and Zhou (2008). Let $\hat{\tilde{\Psi}}_{T,t} = \frac{1}{T}\sum_{\tau \neq t} \tilde{F}_\tau\,\tilde{F}_\tau^\intercal$. Let

$$g_t = \frac{1}{T}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t\,\tilde{F}_t^\intercal\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\frac{1}{1 + T^{-1}\,\tilde{F}_t^\intercal\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\tilde{F}_t}\,. \tag{39}$$

By the Sherman–Morrison formula (40),

$$\begin{aligned}\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-2} = &\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-2} \\ &- \left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} g_t - g_t\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} + g_t^2\,.\end{aligned} \tag{40}$$

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 (40) and defined

$$\hat{l}_t = \mathrm{tr}\left(-\left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} g_t Q_P - g_t\left(z\, I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1} Q_P + g_t^2 Q_P\right)\,. \tag{41}$$

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

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

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,

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

forms a bounded martingale difference sequence. Applying the Burkholder–Davis–Gundy inequality (see, for example, Burkholder (1966)), we get

$$\begin{aligned}\mathbb{E}\left[|q_T(z) - \mathbb{E}\left[q_T(z)\right]|^q\right] &\leq K_q\, P^{-q}\, \mathbb{E}\left(\sum_{t=1}^{T} \left|(\mathbb{E}_{t-1} - \mathbb{E}_t)\left[l_t\right]\right|^2\right)^{q/2} \\ &\leq K_q\, ((2\, z^{-1} K + z^{-2})\, K)^q\, P^{-q/2}\, (P/T)^{-q/2}\,. \end{aligned} \tag{42}$$

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

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

and write

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

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

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

and

$$\hat{\beta} = \left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\,\left(\hat{\Psi}_{\tilde{F},F}\beta + q_T\right)\,, \tag{43}$$

$$\hat{\beta}^\intercal\,\Psi\,\beta = \beta^\intercal\,\Psi\,\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\,\hat{\Psi}_{\tilde{F},F}\,\beta + \beta^\intercal\,\Psi\,\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\,q_T\,. \tag{44}$$

Because $\mathbb{E}[q_T] = 0$ and $\mathbb{E}[q_T^\intercal\,q_T] = \frac{1}{T^2}\,\mathrm{tr}(\tilde{\Psi}) \to 0$ 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

$$\mathbb{E}\left[q_T\,q_T^\intercal \mid \{\tilde{F}_t\}\right] = \frac{1}{T}\,\hat{\tilde{\Psi}}_T\,. \tag{45}$$

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

$$\begin{aligned}
&\mathbb{E}\left[\beta^\intercal\Psi\,\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\,q_T\,q_T^\intercal\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\Psi\beta\right] \\
&= \mathbb{E}\left[\beta^\intercal\Psi\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\mathbb{E}\left[q_T\,q_T^\intercal \mid \{\tilde{F}_t\}\right]\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\Psi\beta\right] \\
&= \mathbb{E}\left[\beta^\intercal\Psi\,\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\,\frac{1}{T}\,\hat{\tilde{\Psi}}_T\,\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\Psi\,\beta\right] \\
&\approx b_*\,P^{-1}\,\mathrm{tr}\,\mathbb{E}\left[\Psi\,\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\,\frac{1}{T}\,\hat{\tilde{\Psi}}_T\,\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\Psi\right] \\
&\leq z^{-1}\,b_*\,\|\tilde{\Psi}\|^2\,P/(P\,T) \to 0
\end{aligned} \tag{46}$$

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 (46) that $\beta^\intercal\, \Psi \left(z\, I + \hat{\tilde{\Psi}}_T \right)^{-1} q_T$ converges to 0 in L_2 thus in probability. Therefore, we only need to calculate the first term in (44).

Write $\tilde{F} = (\tilde{F}_t) \in \mathbb{R}^{P \times T}$ and $F = (\beta_t) \in \mathbb{R}^{P \times T}$, 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

$$\mathbb{P}\left(\left\| \tilde{\Psi}_{\tilde{F},F} \right\| \geq C\, (\log P)^3 \right) \leq \frac{K}{(\log P)^2} + \delta_1 \tag{47}$$

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

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

in probability. Next, use the Sherman–Morrison formula (40) to write

$$
\begin{aligned}
&\frac{1}{P}\ \mathrm{tr}\left[\Psi\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\ \hat{\Psi}_{\tilde{F},F}\right]\\
&=\frac{1}{P}\ \mathrm{tr}\left[\Psi\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\ \frac{1}{T}\sum_t \tilde{F}_t\,\beta_t^{\intercal}\right]\\
&=\frac{1}{P}\ \mathrm{tr}\left[\frac{1}{T}\sum_t \Psi\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\ \tilde{F}_t\,\beta_t^{\intercal}\right]\\
&=\frac{1}{P}\ \mathrm{tr}\left[\frac{1}{T}\sum_t \Psi\left(\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\right.\right.\\
&\quad\left.\left.-\frac{1}{T}\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\ \tilde{F}_t\ \left(I+\frac{1}{T}\,\tilde{F}_t^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\ \tilde{F}_t\right)^{-1}\ \tilde{F}_t^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\right)\tilde{F}_t\beta_t^{\intercal}\right]\\
&=\frac{1}{P}\ \mathrm{tr}\left[\frac{1}{T}\sum_t \Psi\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\ \tilde{F}_t\beta_t^{\intercal}\right]\\
&\quad-\frac{1}{P}\ \mathrm{tr}\left[\frac{1}{T}\sum_t \Psi\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\ \tilde{F}_t\ (I+C_{T,t})^{-1}\,C_{T,t}\beta_t^{\intercal}\right]\\
&=\frac{1}{P}\ \mathrm{tr}\left[\frac{1}{T}\sum_t \beta_t^{\intercal}\Psi\left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\ \tilde{F}_t\right] \qquad (48)\\
&\quad-\frac{1}{P}\ \mathrm{tr}\left[\frac{1}{T}\sum_t \beta_t^{\intercal}\ \Psi\ \left(z\,I+\hat{\tilde{\Psi}}_{T,t}\right)^{-1}\ \tilde{F}_t\ (1+C_{T,t})^{-1}\,C_{T,t}\right] \qquad (49)
\end{aligned}
$$

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

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

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

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

in probability. Therefore,

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

and hence

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

in probability. Note from Lemma 3, equation (42) and the proof of Lemma 9 that

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

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

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

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

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

in L_2 because

$$(X_nY_n-X\,Y)^2=\left((X_n-X)\,Y_n+X\,(Y_n-Y)\right)^2\leq 2\,(X_n-X)^2\,Y_n^2+X^2\,(Y_n-Y)^2$$

and

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

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

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

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

$$\begin{aligned}
\mathcal{E}(z;c) &= \lim \mathbb{E}\left[\hat{\pi}_t(z)\, R_{t+1} \mid \hat{\beta}\right] \\
&= \lim \left(\hat{\beta}^\intercal\, \Psi\, \beta\right) \\
&= \lim \left(b_*\, \Psi\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\, \hat{\Psi}_{\tilde{F},F}/P\right) \\
&= b_*\, c^{-1}\, \xi_\Psi(z;c)/(1+\xi(z;c))\,. \qquad (50)
\end{aligned}$$

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 (43), and we know that (here we do not have the intersection term because (46) tells us that it will go to 0):

$$\begin{aligned}
\operatorname{tr}\left(\tilde{\Psi}\, \hat{\beta}\, \hat{\beta}^\intercal\right) &= \operatorname{tr}\left(\tilde{\Psi}\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\, \left(\hat{\Psi}_{\tilde{F},F}\, \beta\, \beta^\intercal\, \hat{\Psi}_{\tilde{F},F}^\intercal + q_T\, q_T^\intercal\right)\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\right) \\
&= \operatorname{tr}\left(\tilde{\Psi}\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\, \hat{\Psi}_{\tilde{F},F}\, \beta\, \beta^\intercal\, \hat{\Psi}_{\tilde{F},F}^\intercal\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\right) \\
&\quad + \operatorname{tr}\left(\tilde{\Psi}\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\, q_T\, q_T^\intercal\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\right) \\
&= \text{ Term 1} + \text{ Term 2}\,. \qquad (51)
\end{aligned}$$

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

Proof. Let $D_T = q_T^\intercal\, A_T\, q_T$. From (45) we know that

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

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

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,

$$f_T(z+\Delta z)=f_T(z)+f_T'(z)\,\Delta z+\frac{1}{2}\int_z^{z+\Delta z}f_T''(t)\,(z+\Delta z-t)\,\mathrm{dt}\,, \tag{52}$$

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

$$
\begin{aligned}
-z\,T^{-1}\,\mathrm{tr}\left(\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-2}\right) &\to -z\,T^{-1}\,\mathbb{E}\left[\mathrm{tr}\left(\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-2}\tilde{\Psi}\right)\right] \\
&\to z\,\frac{d}{dz}\,T^{-1}\,\mathbb{E}\left[\mathrm{tr}\left(\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\right)\right] \\
&\to z\frac{d}{dz}\xi(z;c)\,. \qquad (53)
\end{aligned}
$$

Altogether, we have

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

Now we deal with Term 1 in (51). 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

$$
\begin{aligned}
\text{Term 1} &\to P^{-1}\,b_*\,\mathrm{tr}\left(\hat{\Psi}^{\intercal}_{\tilde{F},F}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\hat{\Psi}_{\tilde{F},F}\right) \\
&= P^{-1}\,b_*\,\mathrm{tr}\left(\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\hat{\Psi}_{\tilde{F},F}\hat{\Psi}^{\intercal}_{\tilde{F},F}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right) \\
&= P^{-1}\,b_*\,\frac{1}{T^2}\,\mathrm{tr}\left(\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\sum_{t_1,t_2}\tilde{F}_{t_1}\,\beta^{\intercal}_{t_1}\,\beta_{t_2}\,\tilde{F}^{\intercal}_{t_2}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right) \\
&= P^{-1}\,b_*\,\frac{1}{T^2}\sum_{t_1}\mathrm{tr}\left(\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{F}_{t_1}\,\beta^{\intercal}_{t_1}\,\beta_{t_1}\,\tilde{F}^{\intercal}_{t_1}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right) \\
&\quad + P^{-1}\,b_*\,\frac{1}{T^2}\sum_{t_1\neq t_2}\mathrm{tr}\left(\tilde{\Psi}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\tilde{F}_{t_1}\beta^{\intercal}_{t_1}\,\beta_{t_2}\,\tilde{F}^{\intercal}_{t_2}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right) \\
&= \text{Term 11} + \text{Term 12}\,. \qquad (54)
\end{aligned}
$$

We deal with each term separately. First we deal with

$$M_t = \frac{1}{P\,T}\ \mathrm{tr}\left(\tilde{\Psi}\left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\ \tilde{F}_t\,\beta_t^{\intercal}\,\beta_t\,\tilde{F}_t^{\intercal}\ \left(z\,I + \hat{\tilde{\Psi}}_T\right)^{-1}\right)\ . \tag{55}$$

Next, use the Sherman–Morrison formula (40) and define $C_{T,t} = \tilde{F}_t^{\intercal}\left(z\,I + \hat{\tilde{\Psi}}_{T,t}\right)^{-1}\ \tilde{F}_t$, 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

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

Now, use the fact that $\tilde{F}_t$ and β_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

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

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 (53) 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.

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

where the third equation follows from symmetric and the forth equation follows from the Sherman–Morrison's formula (40). 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 (54). Consider

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

using the Sherman–Morrison formula (40), 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}\beta_{t_1}^{\intercal}\beta_{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}\beta_{t_1}^{\intercal}\beta_{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}\,\beta_{t_1}^{\intercal}\beta_{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}\,\beta_{t_1}^{\intercal}\beta_{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}\,\beta_{t_1}^{\intercal}\beta_{t_2}\Big\}\\
&= P^{-1}\,\mathrm{tr}\Big\{(1+C_T)^{-1}\,\tilde{F}_{t_2}^{\intercal}\\
&\times\Bigg(\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}\Bigg)\,\tilde{F}_{t_1}\,(1+C_T)^{-1}\,\beta_{t_1}^{\intercal}\beta_{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 (56)
\end{aligned}
$$

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

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

and

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

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

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

and therefore

$$\begin{aligned}\left\|\sum_{t_1}\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{\Psi}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,\beta_{t_1}^{\intercal}\right\|_F^2 &= \sum_{t_1}\mathrm{tr}\left(A_{t_1}\tilde{F}_{t_1}F^{\intercal}F\tilde{F}_{t_1}^{\intercal}A_{t_1}^{\intercal}\right)\\ &\leq z^{-4}\left\|\tilde{\Psi}\right\|^2\sum_{t_1}\mathrm{tr}\left(\tilde{F}_{t_1}\,F^{\intercal}\,F\,\tilde{F}_{t_1}^{\intercal}\right)\,.\end{aligned} \tag{57}$$

Therefore, as $\tilde{F}_{t_1},\tilde{F}_{t_2},\beta_{t_1},\beta_{t_2}$ are all independent from $\hat{\tilde{\Psi}}_{T,t_1,t_2}$ and $\tilde{F}_{t_1},\beta_{t_2}$ are independent

from $\tilde{F}_{t_2}, \beta_{t_2}$, with the same arguments in Lemma 11 we have

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

With the same arguments in calculating Term 11 in (54), using Lemma 3, 6, and Assumption 3 we have that

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

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 (56), we have

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

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 $\Lambda_{t_1,t_2} = \text{diag}(\lambda_{1,t_1,t_2}, ...\lambda_{P,t_1,t_2})$ 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

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

We know that

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

Therefore,

$$P_{t_1,t_2}/(1+C_T) \leq z^{-1} \, \|\tilde{\Psi}\| \tag{58}$$

is uniformly bounded. Hence

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

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

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

From the calculation for Term 11 in (54) 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

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

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 (56), 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

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

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 (56). We have

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

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 $\tilde{F}_{t_2}^{\intercal}\,\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,\beta_{t_1}^{\intercal}\,\beta_{t_2}$ 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 (58), we have

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

Therefore, using Lemma 3 we have

$$\begin{aligned}
&\frac{1}{\sqrt{P}\,T}\,(1+C_T)^{-1}\,\tilde{F}_{t_2}^{\intercal} \left(z\, I + \hat{\tilde{\Psi}}_{T,t_1,t_2} \right)^{-1} \tilde{F}_{t_1}\, \tilde{F}_{t_1}^{\intercal} \left(z\, I + \hat{\tilde{\Psi}}_{T,t_1,t_2} \right)^{-1} \tilde{\Psi} \left(z\, I + \hat{\tilde{\Psi}}_{T,t_1,t_2} \right)^{-1} \tilde{F}_{t_2} \\
&\approx \frac{1}{\sqrt{P}\,T}\,(1+C_T)^{-1}\,\mathrm{tr}\left(\tilde{\Psi} \left(z\, I + \hat{\tilde{\Psi}}_{T,t_1,t_2} \right)^{-1} \tilde{F}_{t_1}\, \tilde{F}_{t_1}^{\intercal} \left(z\, I + \hat{\tilde{\Psi}}_{T,t_1,t_2} \right)^{-1} \tilde{\Psi} \left(z\, I + \hat{\tilde{\Psi}}_{T,t_1,t_2} \right)^{-1} \right) \\
&= \frac{1}{\sqrt{P}\,T}\,(1+C_T)^{-1}\,\mathrm{tr}\left(\tilde{F}_{t_1}^{\intercal} \left(z\, I + \hat{\tilde{\Psi}}_{T,t_1,t_2} \right)^{-1} \tilde{\Psi} \left(z\, I + \hat{\tilde{\Psi}}_{T,t_1,t_2} \right)^{-1} \tilde{\Psi} \left(z\, I + \hat{\tilde{\Psi}}_{T,t_1,t_2} \right)^{-1} \tilde{F}_{t_1} \right) \\
&\approx \frac{1}{\sqrt{P}\,T}\,(1+\xi(z;c))^{-1}\,\mathrm{tr}\left(\left(\tilde{\Psi} \left(z\, I + \hat{\tilde{\Psi}}_{T,t_1,t_2} \right)^{-1} \right)^3 \right) \\
&\leq c\,(1+\xi(z;c))^{-1} \left\| \tilde{\Psi} \right\|^3 z^{-3}/\sqrt{P} \\
&\to 0\,. \qquad (59)
\end{aligned}$$

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

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

Therefore, with the same arguments as those in (58), we have

$$\mathbb{E}\left[\left\|\left(z\,I+\hat{\tilde{\Psi}}_{T,t_1,t_2}\right)^{-1}\,\tilde{F}_{t_1}\,\beta_{t_1}^{\intercal}\right\|_F^2/(T^2(1+C_T)^2)\right]=\mathbb{E}\left[\tilde{F}_{t_1}^{\intercal}\,Q_{t_1,t_2}^2\,\tilde{F}_{t_1}\,\beta_{t_1}^{\intercal}\,\beta_{t_1}/(T^2\,(1+C_T)^2)\right]$$

$$\leq z^{-2}\,\mathbb{E}[\beta_{t_1}^{\intercal}\,\beta_{t_1}/T^2]$$

$$\leq \|\Psi\|\,z^{-2}\,c/T\to 0\,.$$

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

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

Using Lemma 3 we have

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

$$\leq c\,(1+\xi(z;c))^{-1}\,\|\Psi\|^2\,z^{-1}/\sqrt{P}$$

$$\to 0\,. \tag{60}$$

Combining (59) and (60) we have

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

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

$$\begin{aligned}
\mathcal{L}(z;c) &= \lim \mathrm{E}\left[\hat{\pi}_t(z)^2 \mid \hat{\beta}\right] \\
&= \lim \mathrm{tr}\left(\tilde{\Psi}\,\hat{\beta}\,\hat{\beta}^\intercal\right) \\
&= \lim(\text{Term 1}) + \lim(\text{Term 2}) \\
&= \lim(\text{Term 1}) + \xi(z;c) + z\,\frac{d}{dz}\xi(z;c) \\
&= \lim(\text{Term 11}) + \lim(\text{Term 12}) + \xi(z;c) + z\,\frac{d}{dz}\xi(z;c) \\
&= b_*\psi_{*,1}\left(\xi(z;c) + z\,\frac{d}{dz}\xi(z;c)\right) + \lim(\text{Term 12}) + \xi(z;c) + z\,\frac{d}{dz}\xi(z;c) \\
&= (1 + b_*\psi_{*,1})\left(\xi(z;c) + z\,\frac{d}{dz}\xi(z;c)\right) + \lim\left(\frac{b_*}{T^2}\sum_{t_1\neq t_2} M_{t_1,t_2}\right) \\
&= (1 + b_*\,\psi_{*,1})\,\left(\xi(z;c) + z\,\frac{d}{dz}\xi(z;c)\right) + \frac{b_*\,c^{-1}\,\eta_\Psi(z;c)}{(1+\xi(z;c))^2} \\
&\quad - \frac{2\,b_*\,c^{-1}\,\left(\xi(z;c) + z\,\frac{d}{dz}\xi(z;c)\right)\,\xi_\Psi(z;c)}{1+\xi(z;c)}\,.
\end{aligned} \tag{61}$$

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

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

We know from (61) that

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

Now we turn to $\hat{\beta}^\intercal\, \mathbb{E}\left[\tilde{F}_t\, \beta_t^\intercal\, \beta\, \beta^\intercal\, \beta_t\, \tilde{F}_t^\intercal\right] \hat{\beta}$. 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

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

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*

$$\beta^\intercal\, A\, \mathrm{diag}\left(B\, \beta\, \beta^\intercal\, B\right)\, A\, \beta \to 0 \tag{IA137}$$

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,

$$\tilde{\beta}^\intercal\, A\, \mathrm{diag}\left(\tilde{B}\, \beta\, \beta^\intercal\, B\right)\, A\, \beta = \sum_{i,j} \tilde{\beta}_i\, \tilde{\beta}_j \sum_k A_{i,k}\, A_{k,j}\, \left(B\, \tilde{\beta}\, \tilde{\beta}^\intercal\, B\right)_{k,k}$$

$$= \sum_{i,j} \sum_{j_1,j_2} \tilde{\beta}_i\, \tilde{\beta}_j\, \tilde{\beta}_{j_1}\, \tilde{\beta}_{j_2} \sum_k A_{i,k}\, A_{k,j}\, B_{k,j_1} B_{j_2,k}\,.$$

Let

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

we have

$$
\begin{aligned}
P^{-2}\,\mathbb{E}\left[\tilde{\beta}^\intercal\, A\ \mathrm{diag}\left(B\,\tilde{\beta}\,\tilde{\beta}^\intercal\, B\right)\ A\,\tilde{\beta}\right] &= P^{-2}\,\mathbb{E}\left[\sum_{i,j}\sum_{j_1,j_2}\tilde{\beta}_i\,\tilde{\beta}_j\,\tilde{\beta}_{j_1}\,\tilde{\beta}_{j_2}\ \sum_k A_{i,k}\,A_{k,j}\,B_{k,j_1}\,B_{j_2,k}\right] \\
&\le P^{-2}\,K\sum_i\sum_k A_{i,k}^2\,B_{k,i}^2 + P^{-2}\sum_{i,j,k} A_{i,k}\,A_{k,i}\,B_{k,j}\,B_{j,k} \\
&\quad + P^{-2}\sum_{i,j,k} A_{i,k}\,B_{k,i}\,A_{k,j}\,B_{j,k} + P^{-2}\sum_{i,j,k} A_{i,k}\,B_{i,k}\,A_{k,j}\,B_{k,j} \\
&= P^{-2}\,K\ \mathrm{tr}\left((A\circ B)\,(A\circ B)^\intercal\right) + P^{-2}\ \mathrm{tr}\left((A\,A^\intercal)\circ(B\,B^\intercal)\right) \\
&\quad + P^{-2}\ \mathrm{tr}\left((A\,B^\intercal)\circ(A^\intercal\,B)\right) + P^{-2}\ \mathrm{tr}\left((A\,B)\circ(A^\intercal\,B^\intercal)\right)\,.
\end{aligned}
\tag{62}
$$

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

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 $\beta_t = \tilde{\Psi}_2 Z_t$, and we have

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

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

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

Therefore,

$$\hat{\beta}^\intercal \, \mathbb{E}\left[\tilde{F}_t \, \beta_t^\intercal \, \beta \, \beta^\intercal \, \beta_t \, \tilde{F}_t^\intercal\right] \hat{\beta} = 2 \, \hat{\beta}^\intercal \, \Psi \, \beta \, \beta^\intercal \, \Psi \, \hat{\beta} + \hat{\beta}^\intercal \, (\kappa - 3) \, \Psi^{1/2} \, \mathrm{diag}\left(\Psi^{1/2} \, \beta \, \beta^\intercal \, \Psi^{1/2}\right) \, \Psi^{1/2} \, \hat{\beta} + \hat{\beta}^\intercal \, \beta^\intercal \, \Psi \, \beta \, \tilde{\Psi} \, \hat{\beta} \, .$$

We know from Lemma 1 that

$$\beta^\intercal \, \Psi \, \beta \to b_* \, \psi_{*,1} \, .$$

Combining (50) and (61) we have

$$\lim 2 \, \hat{\beta}^\intercal \, \Psi \, \beta \, \beta^\intercal \, \Psi \, \hat{\beta} \to 2 \, \mathcal{E}(z; c; a)^2, \quad \lim \hat{\beta}^\intercal \, \beta^\intercal \, \Psi \, \beta \, \tilde{\Psi} \, \hat{\beta} \to b_* \, \psi_{*,1} \, \mathcal{L}(z; c; a) \, . \tag{63}$$

Now we deal with

$$\begin{aligned}
\hat{\beta}^\intercal \, \Psi^{1/2} \, \mathrm{diag}\left(\Psi^{1/2} \, \beta \, \beta^\intercal \, \Psi^{1/2}\right) \Psi^{1/2} \hat{\beta} &= \mathrm{tr}\left(\Psi^{1/2} \, \mathrm{diag}\left(\Psi^{1/2} \, \beta \, \beta^\intercal \, \Psi^{1/2}\right) \, \Psi^{1/2} \, \hat{\beta} \, \hat{\beta}^\intercal\right) \\
&= \mathrm{tr}\left(\Psi^{1/2} \, \mathrm{diag}\left(\Psi^{1/2} \, \beta \, \beta^\intercal \, \Psi^{1/2}\right) \, \Psi^{1/2} \, \hat{\beta} \, \hat{\beta}^\intercal\right) \\
&= \mathrm{tr}\left(\Psi^{1/2} \, \mathrm{diag}\left(\Psi^{1/2} \, \beta \, \beta^\intercal \, \Psi^{1/2}\right) \Psi^{1/2} \left(z \, I + \hat{\tilde{\Psi}}_T\right)^{-1}\right. \\
&\quad \times \left(\hat{\Psi}_{\tilde{F},F} \, \beta \, \beta^\intercal \, \hat{\Psi}_{\tilde{F},F}^\intercal + q_T \, \beta^\intercal \, \hat{\Psi}_{\tilde{F},F}^\intercal + \hat{\Psi}_{\tilde{F},F} \, \beta \, q_T + q_T \, q_T^\intercal\right) \\
&\quad \left. \times \left(z \, I + \hat{\tilde{\Psi}}_T\right)^{-1}\right) \, .
\end{aligned}$$

Next, we show that

$$\hat{\beta}^\intercal \, \Psi^{1/2} \, \mathrm{diag}\left(\Psi^{1/2} \, \beta \, \beta^\intercal \, \Psi^{1/2}\right) \, \Psi^{1/2} \, \hat{\beta} \to 0 \tag{64}$$

in probability. Note that using Cauchy–Schwartz Inequality we only need to prove that

$$\operatorname{tr}\left(\Psi^{1/2}\operatorname{diag}\left(\Psi^{1/2}\,\beta\,\beta^{\intercal}\,\Psi^{1/2}\right)\Psi^{1/2}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\,\hat{\Psi}_{\tilde{F},F}\,\beta\,\beta^{\intercal}\,\hat{\Psi}_{\tilde{F},F}^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right)\to 0 \quad (65)$$

in probability and that

$$\operatorname{tr}\left(\Psi^{1/2}\operatorname{diag}\left(\Psi^{1/2}\,\beta\,\beta^{\intercal}\,\Psi^{1/2}\right)\Psi^{1/2}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}q_T\,q_T^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right)\to 0 \qquad (66)$$

in probability. To prove (65), note that from (47) 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 (62) in Lemma 14 we have that

$$\begin{aligned}&\mathbb{E}\left[\operatorname{tr}\left(\Psi^{1/2}\operatorname{diag}\left(\Psi^{1/2}\,\beta\,\beta^{\intercal}\,\Psi^{1/2}\right)\Psi^{1/2}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\,\hat{\Psi}_{\tilde{F},F}\,\beta\,\beta^{\intercal}\,\hat{\Psi}_{\tilde{F},F}^{\intercal}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right)\right]\\&\leq C\log^6 P/P\to 0\end{aligned}$$

with probability at least δ_1. Let $\delta_1\to 0$ we have that proved (65). For (66), we first consider $\operatorname{diag}\left(\Psi^{1/2}\,\beta\,\beta^{\intercal}\,\Psi^{1/2}\right)$, 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 $\mathbb{E}[\gamma_i^2] = b_*/P$, $\mathbb{E}[\gamma_i^4] \le K/P^2$ and γ_i, γ_j are independent, we have

$$\mathbb{E}\left(\Psi^{1/2}\,\beta\,\beta^\intercal\,\Psi^{1/2}\right)_{k,k} = \mathbb{E}[a_k^\intercal\,\beta\,\beta^\intercal\,a_k] = \mathbb{E}\left[\sum_{i,j} a_{k,i}\gamma_i\,\gamma_j a_{k,j}\right] = b_*\,a_k^\intercal\,a_k/P = b_*\,\Psi_{k,k}/P$$

and

$$\begin{aligned}
\mathbb{E}\left(\Psi^{1/2}\,\beta\,\beta^\intercal\,\Psi^{1/2}\right)_{k,k}^2 &= \mathbb{E}\left(\sum_{i,j} a_{k,i}\,\gamma_i\,\gamma_j\,a_{k,j}\right)^2 \\
&= \mathbb{E}\left[\sum_{i,j,l,m} a_{k,i}\,\gamma_i\,\gamma_j\,a_{k,j}\,a_{k,l}\,\gamma_l\,\gamma_m\,a_{k,m}\right] \\
&\le K\sum_i a_{k,i}^4/P^2 + 6\,b_*^2\sum_{i\neq j} a_{k,i}^2\,a_{k,j}^2/P^2 \\
&\le (6\,b_*^2 + K)\,(a_k^\intercal\,a_k)^2/P^2 \\
&= (6\,b_*^2 + K)\,\Psi_{k,k}^2/P^2\,.
\end{aligned}$$

Let $\tilde{m} = 1/\log P$, then we have

$$\begin{aligned}
\mathbb{P}\left(\exists k, |(\Psi^{1/2}\,\beta\,\beta^\intercal\,\Psi^{1/2})_{k,k} - b_*\Psi_{i,i}/P| \ge \tilde{m}\right) &\le \sum_k \mathbb{P}\left(|(\Psi^{1/2}\,\beta\,\beta^\intercal\,\Psi^{1/2})_{k,k} - b_*\Psi_{i,i}/P| \ge \tilde{m}\right) \\
&\le \sum_i \frac{\mathrm{Var}\left[(\Psi^{1/2}\,\beta\,\beta^\intercal\,\Psi^{1/2})_{k,k}\right]}{\tilde{m}^2} \\
&\le \frac{(5\,b_* + K)\,\|\Psi\|^2\,\log^2 P}{P}\,.
\end{aligned}$$

Therefore, we have

$$\mathbb{P}\left(\left\|\mathrm{diag}\left(\Psi^{1/2}\,\beta\,\beta^\intercal\,\Psi^{1/2}\right)\right\| \ge \tilde{m}\right) \le \frac{(5b_* + K)\|\Psi\|^2\log^2 P}{P} \to 0\,.$$

Therefore, using Lemma 11 we have, under the event that $\left\|\mathrm{diag}\left(\Psi^{1/2}\,\beta\,\beta^\intercal\,\Psi^{1/2}\right)\right\| \le \tilde{m}$,

$$\begin{aligned}
&\mathrm{tr}\left(\Psi^{1/2}\,\mathrm{diag}\left(\Psi^{1/2}\,\beta\,\beta^\intercal\,\Psi^{1/2}\right)\Psi^{1/2}\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1} q_T q_T^\intercal \left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right)\\
&\le \tilde{m}\,\mathrm{tr}\left(\Psi\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1} q_T q_T^\intercal \left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\right)\\
&\approx \tilde{m}\,\mathrm{tr}\left(\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\Psi\left(z\,I+\hat{\tilde{\Psi}}_T\right)^{-1}\hat{\tilde{\Psi}}_T\right)\\
&\to 0\,.
\end{aligned}$$

Therefore, (66) is proven. Altogether, combining (63) and (64) we have

$$\begin{aligned}
\lim \mathbb{E}\left[(\hat{\pi}_t(z)R_{t+1})^2 \mid \hat{\beta}\right] &= \lim \hat{\beta}^\intercal\,\mathbb{E}\left[\tilde{F}_t\,\beta_t^\intercal\,\beta\,\beta^\intercal\,\beta_t\,\tilde{F}_t^\intercal\right]\hat{\beta} + \lim \mathrm{tr}\left(\tilde{\Psi}\hat{\beta}\hat{\beta}^\intercal\right)\\
&= 2\,\mathcal{E}(z;c;a)^2 + (1+b_*\,\psi_{*,1})\,\mathcal{L}(z;c;a)\,.
\end{aligned}$$

Theorem 1 is proven. □

A.3 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

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

because of Proposition 1 and $\tilde{\Psi}_{1,1} = (1+a\,q)I_{P_1}$. Write $\gamma_1 = \{\gamma_i\}_{i=1}^{P_1}$ and $\gamma_2 = \{\gamma_i\}_{i=P_1+1}^{P}$,

now we calculate

$$\lim \mathbb{E}\left[\hat{\pi}_{1,t}(z)\, R_{t+1} \mid \hat{\tilde{\beta}}_1\right] = \hat{\tilde{\beta}}_1^{\intercal}\, \Psi_{1,1}\, \gamma_1\,.$$

Note that

$$\begin{aligned}\hat{\tilde{\beta}}_1^{\intercal}\, \Psi_{1,1}\, \gamma_1 &= \gamma_1^{\intercal}\, \Psi_{1,1}\, \left(z\, I_{P_1} + \hat{\tilde{\Psi}}_{1,1}\right)^{-1}\, \left(\hat{\Psi}_{\tilde{F},F,1}\gamma_1 + \hat{\Psi}_{\tilde{F},F,2}\,\gamma_2 + q_T\right) \\ &= \text{Term 1} + \text{Term 2} + \text{Term 3}\,, \end{aligned} \tag{67}$$

where

$$\hat{\Psi}_{\tilde{F},F,1} = \frac{1}{T}\sum_t \tilde{F}_{t,(1)}\,\beta_{t,(1)}^{\intercal}\,, \qquad \hat{\Psi}_{\tilde{F},F,2} = \frac{1}{T}\sum_t \tilde{F}_{t,(1)}\,\beta_{t,(2)}^{\intercal}\,, \qquad \text{and} \qquad q_T = \frac{1}{T}\sum_t \tilde{F}_{t,(1)}\,\epsilon_{t+1}\,.$$

With the same arguments as those in the proof of Theorem 1 and $\mathbb{E}(\tilde{F}_{t,(1)}\beta_{t,(2)}^{\intercal}) = 0$ and $\mathbb{E}(\gamma_1\gamma_2^{\intercal}) = 0$, 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

$$\begin{aligned}\operatorname{tr}\left(\tilde{\Psi}_{1,1}\,\hat{\tilde{\beta}}_1\,\hat{\tilde{\beta}}_1^{\intercal}\right) &= \operatorname{tr}\left(\tilde{\Psi}_{1,1}\, \left(z\, I + \hat{\tilde{\Psi}}_{1,1}\right)^{-1}\, \left(\hat{\Psi}_{\tilde{F},F,1}\,\gamma_1\,\gamma_1^{\intercal}\hat{\Psi}_{\tilde{F},F,1}^{\intercal} + \hat{\Psi}_{\tilde{F},F,2}\gamma_2\gamma_2^{\intercal}\hat{\Psi}_{\tilde{F},F,2}^{\intercal} + q_T q_T^{\intercal}\right)\, \left(z\, I + \hat{\tilde{\Psi}}_T\right)^{-1}\right) \\ &= \text{ Term 1} + \text{ Term 2} + \text{Term 3}\,. \end{aligned} \tag{68}$$

With the same arguments as in the proof of Theorem 1 and note that $\mathbb{E}[\tilde{F}_{t,(1)}\,\beta_{t,(2)}^{\intercal}] = 0$ and $\mathbb{E}[\gamma_1\,\gamma_2^{\intercal}] = 0$, we have

$$\text{Term 3} \to \xi(z; c\,q; a\,q) + z\,\frac{d}{dz}\xi(z; c\,q; a\,q)\,,$$

$$\text{Term 1} \to \frac{b_*\,\mathrm{tr}(\Psi_{1,1})}{P}\;(\xi(z;\,c\,q;\,a\,q) + z\,\partial\xi(z;\,c\,q;\,a\,q)/\partial z) - \frac{b_*\,c^{-1}\,\partial\xi(z;\,c\,q;\,a\,q)/\partial z}{(1+\xi(z;\,c\,q;\,a\,q))^2}$$
$$- \frac{2b_*\,c^{-1}\;(\xi(z;\,c\,q;\,a\,q) + z\,\partial\xi(z;\,c\,q;\,a\,q)/\partial z)\;\xi(z;\,c\,q;\,a\,q)}{(1+\xi(z;\,c\,q;\,a\,q))\,(1+a\,q)}\,,$$

$$\text{Term 2} \to \frac{b_*\,\mathrm{tr}(\Psi_{2,2})}{P}\;(\xi(z;\,c\,q;\,a\,q) + z\,\partial\xi(z;\,c\,q;\,a\,q)/\partial z)\;.$$

Altogether, we have

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

Finally, with the same argument as that in the proof of Theorem 1, we have

$$\lim \mathbb{E}\left[(\hat{\pi}_{1,t}(z)\,R_{t+1})^2 \mid \hat{\tilde{\beta}}_1\right] = \lim \hat{\tilde{\beta}}_1^{\intercal}\,\mathbb{E}\left[\tilde{F}_{t,(1)}\beta_t^{\intercal}\,\beta\,\beta^{\intercal}\,\beta_t\tilde{F}_{t,(1)}^{\intercal}\right]\hat{\tilde{\beta}}_1 + \lim \mathrm{tr}\left(\tilde{\Psi}_{1,1}\,\hat{\tilde{\beta}}_1\,\hat{\tilde{\beta}}_1^{\intercal}\right)$$
$$= 2\,\mathcal{E}(z;c;a;q)^2 + (1+b_*)\,\mathcal{L}(z;c;a;q)\,.$$

□

A.4 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

$$\begin{aligned}
\frac{\partial}{\partial a}\mathcal{E}(z;c;a) = &\big(-b_* \left(-a^2\,c^2 + 2\,a^2\,c - a^2 - 2\,a\,c^2\right.\\
&- 3\,a\,c\,z + 4\,a\,c - 3\,a\,z - 2\,a - c^2 - 3\,c\,z + 2\,c - 2\,z^2 - 3z - 1\\
&\left.+(a\,c+a+c+2\,z+1)\sqrt{2\,(a+1)\,(c+1)\,z + (a+1)^2\,(c-1)^2 + z^2}\right)\big)\\
&/\left(2\,(a+1)^3\,c\,\sqrt{(a\,c+a+c+z+1)^2 - 4\,(a+1)^2\,c}\right)\\
=&(-b_*\,(A_1+B_1))/\left(2\,(a+1)^3\,c\,\sqrt{(a\,c+a+c+z+1)^2 - 4\,(a+1)^2 c}\right)\,.
\end{aligned}$$

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

$$\tilde{R}^2(z;c;a) = 2\,\mathcal{E}(z;c;a) - \mathcal{L}(z;c;a) \tag{69}$$

and

$$\widetilde{SR}(z;c;a) = \mathcal{L}(z;c;a)/\mathcal{E}(z;c;a)^2\,. \tag{70}$$

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 (69), we have

$$\begin{aligned}
\frac{\partial}{\partial z}\tilde{R}^2(z;c;a) &= 8\,((a+1)\,c\,(ab_*+a+1) - b_*\,z) \\
&\quad \times \left(3\,(a+1)\,c\,z^2 + 3\,(a+1)^2\,c\,(c-1)\,z + (a+1)^3\,(c-1)^3 + z^3\right. \\
&\quad \left. + C(z;c;a)\,\left((a+1)\,(2\,c-1)z + (a+1)^2\,(c-1)^2 + z^2\right)\right) \\
&\quad / \left((b_*+1)\left((a\,c+a+c+z+1)^2 - 4\,(a+1)^2\,c\right)^{3/2}\right. \\
&\quad \left. \times\,\left(C(z;c;a) + a\,(c-1) + c + z - 1\right)^3\right) \\
&= 8\,((a+1)\,c\,(a\,b_*+a+1) - b_*\,z)\,(A_3+B_3) \\
&\quad / \left((b_*+1)\,\left((a\,c+a+c+z+1)^2 - 4\,(a+1)^2\,c\right)^{3/2}\right. \\
&\quad \left. \times\,\left(C(z;c;a) + a\,(c-1) + c + z - 1\right)^3\right)\,.
\end{aligned}$$

We know that

$$2\,(a+1)\,(c+1)\,z+(a+1)^2\,(c-1)^2+z^2\geq|(a+1)\,(c-1)+z|\,.$$

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

$$\begin{aligned}\frac{\partial}{\partial z}\widetilde{SR}^2(z;c;a)=&-8\,(a+1)^2\,c\,((a+1)\,c\,(a\,b_*+a+1)-b_*\,z)\\&\times\left(3\,(a+1)^2\,(c-1)\,z+(a+1)^3\,(c-1)^3-3\,(a+1)\,z^2-z^3\right.\\&\left.+\left(-(a+1)\,(c-2)\,z+(a+1)^2\,(c-1)^2+z^2\right)C(z;c;a)\right)\\&/\left(b_*^2\,C(z;c;a)^3\,(C(z;c;a)+a\,c-a+c-z-1)^3\right)\\=&-8\,(a+1)^2\,c\,((a+1)\,c\,(a\,b_*+a+1)-b_*\,z)\,(A_5+B_5)\\&/\left(b_*^2\,C(z;c;a)^3\;(C(z;c;a)+a\,c-a+c-z-1)^3\right)\,.\end{aligned}$$

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

$$\frac{\partial}{\partial a} R^2(z_*; c; a) = \frac{\partial}{\partial a} \tilde{R}^2(z_*; c; a)/(1 + b_*) = \frac{\partial}{\partial a} \mathcal{E}(z_*; c; a)/(1 + b_*) \leq 0\,. \tag{71}$$

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

$$
\begin{aligned}
A_{60} =& 2\,(a+1)^6\,(c-1)^4\,\left(b_*^2\,\left(3\,c\,\left(\left(2\,a^2+a-1\right)\,c^2+\left(a^2+2\right)\,c+(a+1)^2\,c^3+a-1\right)+1\right)\right.\\
&\left.+3\,(a+1)\,b_*\,c\,(c+1)\,(c\,(2\,a\,(c+1)+2c-1)+1)+3\,(a+1)^2\,c^2\,(c+1)^2\right)>0\,,\\
A_{61} =& 4\,(a+1)^5\,(c-1)^2\,\left(b_*^2\,\left(9\,(a+1)^2\,c^5+(a+1)\,(21\,a-8)\,c^4+(7\,a\,(3\,a+1)+13)\,c^3\right.\right.\\
&\left.+(a\,(9\,a+5)+5)\,c^2+(9\,a-6)\,c+3\right)+(a+1)\,b_*\,c\,(c\,(6\,a\,(c+1)\,(c\,(3\,c+4)+3)\\
&+c\,(c\,(18\,c+13)+7)+5)+9)+3\,(a+1)^2\,c^2\,(c+1)\,(c\,(3c+4)+3))>0\,,\\
A_{62} =& 2\,(1+a)^4\,\left(b_*^2\,\left(\left(58\,a^2+44\,a+85\right)\,c^4+45\,(a+1)^2\,c^6+3\,(a+1)\,(22\,a-23)\,c^5\right.\right.\\
&\left.+6\,(a\,(11\,a+9)-3)\,c^3+3\,(a\,(15\,a-2)+13)\,c^2+(45\,a-33)\,c+15\right)+b_*\,c\,(1+a)\\
&\left.\times\,\left((90\,a+90)\,c^5+(132\,a-3)\,c^4+(116\,a+44)\,c^3+(132\,a+54)\,c^2+(90\,a-6)\,c+45\right)\right)>0\,,\\
A_{63} =& 8\,(a+1)^3\,\left(3\,(a+1)^2\,c^2\,(c+1)\,(c\,(5\,c+4)+5)+b_*^2\,\left(15\,(a+1)^2\,c^5+9\,(a+1)\,(3\,a-1)\,c^4\right.\right.\\
&\left.+(9\,a\,(3\,a+2)+7)\,c^3+(3\,a\,(5\,a+3)+2)\,c^2+(15\,a-4)\,c+5\right)+3\,(a+1)\,b_*\,c\\
&\left.\times(c\,(\,2a\,(c+1)\,(c\,(5\,c+4)+5)+2\,c\,(c\,(5\,c+3)+3)+3)+5))>0\,,\right.\\
A_{64} =& 2\,\left((a+1)^2\,b_*^2\,\left(c\,\left(45\,(a+1)^2\,c^3+(a+1)\,(72\,a+5)\,c^2\right.\right.\right.\\
&\left.\left.+(a\,(4\,5a+52)+10)\,c+45\,a+9)+15)+(a+1)^3\,b_*\,c\,(c(18\,a\,(c\,(5\,c+8)+5)\right.\right.\\
&\left.\left.+c\,(90\,c+77)+52)+45)+9\,(a+1)^4\,c^2\,(c\,(5\,c+8)+5)\right)>0\,,\right.\\
A_{65} =& 2\,\left(6\,(a+1)\,b_*^2\,\left(c\,\left(3\,(a+1)^2\,c^2+(a+1)\,(3\,a+2)\,c+3\,a+2\right)+1\right)\right.\\
&\left.+6\,(a+1)^2\,b_*\,c\,(c\,(6\,a\,(c+1)+6\,c+5)+3)+18\,(a+1)^3\,c^2(c+1)\right)>0\,,\right.\\
A_{66} =& 2\,\left(b_*^2\,(3\,(a+1)\,c\,(a\,c+c+1)+1)+3\,(a+1)\,b_*\,c\,(2\,(a+1)c+1)+3\,(a+1)^2\,c^2\right)>0\,.
\end{aligned}
$$

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

$$\begin{aligned}
B_{60} &= (a+1)^5\,(c-1)^4\,\left(3\,(a+1)^2\,(b_*+1)^2\,c^3\right.\\
&\quad \left.+3\,(a+1)\,(b_*+1)\,c^2\,(ab_*+a+1)+3\,b_*\,c\,(ab_*+a+1)+b_*^2\right) > 0\,,\\
B_{61} &= 3\,(1+a)^6\,(c-1)^2\,c^2\,(5+c\,(8+5\,c))+(a+1)^4\,b_*^2\,(c-1)^2\\
&\quad \times\left(15\,(a+1)^2\,c^4+(a+1)\,(24\,a+7)\,c^3+(a\,(15\,a+26)-4)\,c^2+(15\,a+5)\,c+5\right)\\
&\quad +(a+1)^5 b_*(c-1)^2 c(c(6a(c(5c+8)+5)+c(30c+31)+26)+15) > 0\,,\\
B_{62} &= 3\,(a+1)^5\,c^2\,(c+1)\,(c\,(10c-1)+10)+2\,(a+1)^4\,b_*\,c\\
&\quad \times\left(\left(30\,(a+1)\,c^3+(27\,a+13)\,c^2+27\,(a+1)\,c+30\,a+17\right)+15\right)+(a+1)^3\,b_*^2\\
&\quad \times\left(c\,\left(\left(30\,a^2+34\,a-1\right)c+30\,(a+1)^2\,c^4+(a+1)\,(27\,a-1)\,c^3\right.\right.\\
&\quad \left.\left.+3\,(9\,a\,(a+2)+1)\,c^2+30\,a+7\right)+10\right) > 0\,,\\
B_{63} &= 3\,(a+1)^4\,c^2\,(c\,(10\,c+13)+10)+2\,(a+1)^3\,b_*c\,\left(c\,\left(30\,(a+1)\,c^2+(39\,a+32)\,c\right.\right.\\
&\quad \left.+30\,a+28)+15\right)+(a+1)^2\,b_*^2\,\left(c\,\left(30\,(a+1)^2\,c^3+(a+1)\,(39\,a+25)\,c^2\right.\right.\\
&\quad \left.\left.+2\,(a\,(15\,a+28)+6)\,c+30\,a+19)+10\right) > 0\,,\\
B_{64} &= 15\,(a+1)^3\,c^2\,(c+1)+(a+1)\,b_*^2\,\left(c\left(15\,(a+1)^2\,c^2+3\,(a+1)\,(5\,a+6)\,c+15\,a+14\right)+5\right)\\
&\quad +3\,(a+1)^2\,b_*\,c\,(c\,(10\,a\,(c+1)+10\,c+11)+5) > 0\,,\\
B_{65} &= b_*^2\,(3\,(a+1)\,c\,(a\,c+c+1)+1)+3\,(a+1)b_*\,c\,(2\,(a+1)\,c+1)+3\,(a+1)^2\,c^2 > 0\,.
\end{aligned}$$

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., Bertsekas (1997)) 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 (71) we know that

$$\frac{\partial}{\partial a}SR(z_*;c;a) \le 0\,.$$

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

$$B_{70} = 8\,(a+1)^4\,c^4 > 0\,,$$

$$B_{71} = 32\,(a+1)^4\,c^4 + 24\,a\,(a+1)^2\,c^3 + 24\,(a+1)^2\,c^3 > 0\,,$$

$$B_{72} = 27\,a^2\,c^2 + 48\,(a+1)^4\,c^4 + 72\,a\,(a+1)^2\,c^3 + 36\,(a+1)^2\,c^3 + 54\,a\,c^2 + 27\,c^2 > 0\,,$$

$$B_{73} = 14\,c + 14\,a\,c - 6\,c^2 + 48\,a\,c^2 + 54\,a^2\,c^2 + 72\,a\,(1+a)^2\,c^3 + 32\,(1+a)^4\,c^4 > 0\,,$$

$$B_{74} = 8\,a^4\,c^4 + a^3\,\left(32\,c^4 + 24\,c^3\right) + a^2\,\left(48\,c^4 + 36\,c^3 + 27\,c^2\right) + a\,\left(32\,c^4 - 6\,c^2 + 14\,c\right)$$
$$+ 8\,c^4 - 12\,c^3 + 23\,c^2 - 12\,c + 3 > 0\,.$$

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.

□

A.5 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

$$\begin{aligned}\frac{\partial \mathcal{E}(z_{*,1};c;q;a)}{\partial q} = \Bigg(&-a\,b_*\sqrt{(a\,q+1)^2\,\left(\frac{(a\,(b_*+1)\,c\,q+b_*\,c+b_*+c)^2}{b_*^2}-4\,c\,q\right)}\\ &+b_*\,(a\,q+1)\,((a-2)\,a\,c\,q+a\,c+a+2\,c)+a\,c\,(a\,q+1)^2\Big)\\ &/\left(2\,c\,(a\,q+1)^2\sqrt{(a\,q+1)^2\,\left(\frac{(a\,(b_*+1)\,c\,q+b_*\,c+b_*+c)^2}{b_*^2}-4\,c\,q\right)}\right)\end{aligned}$$

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 Motivating example

Here, we provide the details of how we derived the results for Figure 2 in the introduction. Following Gu et al. (2019), 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 Green et al. (2017), 74 industry dummies, and 8 macroeconomic predictors from Welch and Goyal (2008), closely following the procedure in Gu et al. (2019).

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 (Clevert et al., 2015) 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 1, we follow the same procedure as that in Section 5, including identical data-processing steps. We vary P from 3 to 10000 and all other parameters and training steps remain as described in Section 5.

For Figure 3, in the left-hand panel, the noise is given by

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

and in the right-hand panel the noise is given by

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

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. All other parameters and training steps are as described in Section 5.