s:fi
RESEARCH

# Swiss Finance Institute Research Paper Series N°21-90

## The Virtue of Complexity in Return Prediction


Bryan Kelly
Yale School of Management, AQR Capital Management, and NBER

Semyon Malamud
Ecole Polytechnique Fédérale de Lausanne, Swiss Finance Institute, and CEPR

Kangying Zhou
Yale School of Management

# The Virtue of Complexity in Return Prediction

Bryan Kelly, Semyon Malamud, and Kangying Zhou*

October 7, 2022

**Abstract**

Much of the extant literature predicts market returns with "simple" models that use only a few parameters. Contrary to conventional wisdom, we theoretically prove that simple models severely understate return predictability compared to "complex" models in which the number of parameters *exceeds* the number of observations. We empirically document the virtue of complexity in US equity market return prediction. Our findings establish the rationale for modeling expected returns through machine learning.

**Keywords:** Portfolio choice, machine learning, random matrix theory, benign overfit, overparameterization
**JEL:** C3, C58, C61, G11, G12, G14

*Bryan Kelly is at Yale School of Management, AQR Capital Management, and NBER; www.bryankellyacademic.org. Semyon Malamud is at Swiss Finance Institute, EPFL, and CEPR, and is a consultant to AQR. Kangying Zhou is at Yale School of Management. We are grateful for helpful comments from Cliff Asness, Kobi Boudoukh, James Choi, Frank Diebold, Egemen Eren, Paul Goldsmith-Pinkham, Amit Goyal, Ron Kaniel (discussant), Stefan Nagel (editor), Andreas Neuhierl (discussant), Matthias Pelster (discussant), Olivier Scaillet (discussant), Christian Schlag (discussant), Hui Wang (discussant), Guofu Zhou (discussant), and seminar participants at AQR, Yale, Vienna University of Economics and Business, Philadelphia Fed, Bank for International Settlements, NYU Courant, EPFL, and conference participants at the Macro Finance Society, Adam Smith Asset Pricing Conference, SFS Cavalcade North America Conference, Hong Kong Conference for Fintech, AI, and Big Data in Business, Wharton Jacobs-Levy Conference, Research Symposium on Finance and Economics, China International Risk Forum, Stanford SITE New Frontiers in Asset Pricing, and XXI Symposium. AQR Capital Management is a global investment management firm, which may or may not apply similar investment techniques or methods of analysis as described herein. The views expressed here are those of the authors and not necessarily those of AQR. Semyon Malamud gratefully acknowledges support from the Swiss Finance Institute and the Swiss National Science Foundation.

# 1 Introduction

The finance literature has recently seen rapid advances in return prediction methods borrowing from the machine learning canon. The primary economic use case of these predictions has been portfolio construction. While a number of papers have documented significant empirical gains in portfolio performance through the use of machine learning, there is little theoretical understanding of return forecasts and portfolios formed from heavily parameterized models.

We provide a theoretical analysis of such "machine learning portfolios." Our analysis can be summarized in the following thought experiment. Imagine there is a true predictive model of the form

$$R_{t+1} = f(G_t) + \epsilon_{t+1} \tag{1}$$

where $R$ is an asset return, $G$ is a fixed set of predictive signals, and $f$ is a smooth function. The predictors $G$ may be known to the analyst, but the prediction function $f$ is unknown. Rather than futilely guessing the functional form, the analyst relies on the universal approximation rationale of, e.g., Hornik et al. (1990), that $f$ can be approximated with a sufficiently wide neural network,

$$f(G_t) \approx \sum_{i=1}^{P} S_{i,t}\beta_i,$$

where $S_{i,t} = \tilde{f}(w_i' G_t)$ is a known nonlinear activation function with known weights $w_i$ and $P$ is sufficiently large.[1] As a result, (1) takes the form

$$R_{t+1} = \sum_{i=1}^{P} S_{i,t}\beta_i + \tilde{\epsilon}_{t+1}. \tag{2}$$

The training sample for this regression has a fixed number of data points, $T$, and the

[1] Assuming known weights $w_i$ is innocuous, as the universal approximation result applies even if weights are randomly generated (Rahimi and Recht, 2007). In fact, our empirical analysis uses the Rahimi and Recht (2007) random Fourier feature method to generate features as in (2).

analyst must decide on the "complexity," or the number of features $P$, to use in their approximating model. A simple model, one with $P << T$, will have low variance thanks to parsimonious parameterization but will be a coarse approximator of $f$. On the other hand, a high-complexity model ($P > T$) has better approximation potential but may be poorly behaved and will require shrinkage/bias. Our central research question is: What level of model complexity (which $P$) should the analyst opt for? Does the approximation improvement from large $P$ justify the statistical costs (higher variance and/or higher bias)?

Answer: We prove that, in the high-complexity regime ($P > T$), expected out-of-sample forecast accuracy and portfolio performance are *strictly increasing* in model complexity. The analyst should always use the largest approximating model that she can compute. Applying optimal shrinkage to this large $P$ model enhances performance further (indeed, we derive the choice of shrinkage that maximizes expected out-of-sample model performance). In other words, when the true data-generating process (DGP) is unknown, the approximation gains achieved through model complexity dominate the statistical costs of heavy parameterization. The interpretation is not necessarily that asset returns are subject to a large number of fundamental driving forces. Even when the driving variables ($G_t$) are low-dimensional, complex models better leverage the information content of $G_t$ by more accurately approximating the unknown and likely nonlinear prediction function.

To provide intuitive characterizations of forecast and portfolio behavior in complex models, our theoretical environment has two simplifying aspects. First, the machine learning models we study are restricted to high-dimensional linear models. As suggested by equation (2), this sacrifices little generality as a number of recent papers have established an equivalence between high-dimensional linear models and more sophisticated models such as deep neural networks (Jacot et al., 2018; Hastie et al., 2019; Allen-Zhu et al., 2019). Second, we focus on a single risky asset. Thus prediction is isolated to the time-series dimension, and

the portfolio optimization problem reduces to market timing.[2] These two simplifications make our key findings more accessible, yet neither is critical for our conclusions.

To provide a baseline for our findings, consider the well-known deficiency of ordinary least squares (OLS) prediction in high dimensions. As the number of regressors, $P$, approaches the number of data points, $T$, the expected out-of-sample $R^2$ tends to negative infinity. An immediate implication is that a portfolio strategy attempting to use OLS return forecasts in such a setting will have divergent variance. In turn, its expected out-of-sample Sharpe ratio collapses to zero. The intuition behind this is simple: When the number of regressors is similar to the number of data points, the regressor covariance matrix is unstable, and its inversion induces wild variation in coefficient estimates and forecasts. This is commonly interpreted as overfitting: With $P = T$, the regression exactly fits the training data and performs poorly out-of-sample.

We are particularly interested in the behavior of portfolios in the *high model complexity* regime, where the number of predictors *exceeds* the number of observations ($P > T$).[3] In this case, standard regression logic no longer holds because the regressor inverse covariance matrix is not defined. However, the pseudo-inverse is defined, and it corresponds to a limiting ridge regression with infinitesimal shrinkage, or the "ridgeless" limit. An emergent statistics and machine learning literature shows that, in the high-complexity regime, ridgeless regression can achieve accurate out-of-sample forecasts despite fitting the training data perfectly.[4]

We analyze related phenomena in the context of return prediction and portfolio optimization. We establish the striking theoretical result that market timing strategies based on

[2]The single asset time series case is economically important in its own right. It coincides with predictive regression for the market return, which has been the primary method for investigating a central organizing question of asset pricing: How much do discount rates vary over time? While our analysis can be applied to a panel of many assets, the roles of covariances in asset returns and signals across stocks complicate the theory.

[3]The statistics and machine learning community often refer to $P > T$ as the "high-dimensional" or "over-parameterized" regime. We avoid terminology like "over-parameterized" and "overfit" as it suggests the model uses too many parameters, which is not necessarily the case. For example, the true data-generating process may be highly complex (i.e., $P$ is large relative to $T$); thus, a correctly specified model would require $P > T$. When an empirical model has the same specification as the true model, we would prefer to call it correctly parameterized as opposed to over-parameterized.

[4]This seemingly counterintuitive phenomenon is sometimes called "benign overfit" (Bartlett et al., 2020; Tsigler and Bartlett, 2020).

ridgeless least squares predictions generate positive Sharpe ratio improvements for arbitrarily high levels of model complexity. Stated more plainly, when the true DGP is highly complex—i.e., it has many more parameters than there are training data observations—one might think that a timing strategy based on ridgeless regression is bound to fail. After all, it *exactly* fits the training data with zero error. Surprisingly, this intuition is wrong. We prove that strategies based on extremely high-dimensional models can thrive out-of-sample and outperform strategies based on simpler models under fairly general conditions.

Our theoretical analysis delivers a number of additional conclusions. First, it shows that the out-of-sample $R^2$ from a prediction model is an incomplete measure of its economic value. A market timer can generate significant economic profits even when the predictive $R^2$ is negative. The reason is that the $R^2$ is heavily influenced by the variance of forecasts.[5] A very low out-of-sample $R^2$ indicates a highly volatile timing strategy. But the properties of least squares imply that the expected out-of-sample return of a timing strategy is always positive. So, as long as the timing variance is not too high (the $R^2$ not too negative), the timing Sharpe ratio can be substantial.

Second, we study two theoretical cases, one for correctly specified models and one for mis-specified models. The correctly specified case develops the behavior of timing portfolios when the true DGP varies from simple to complex, holding the data size fixed. This is valuable for developing a general understanding of machine learning portfolios for various DGPs. But the correct model specification is unrealistic—it is unlikely that we ever have a predictor data set that nests all relevant conditioning information, and it is also unlikely that we use information in the proper functional form. Our main theoretical results pertain to mis-specified models, and this analysis coincides with the thought experiment above. In

[5]That is, $R^2$ is not just about predictive correlation. Consider a simple model with a single predictor and a coefficient estimate many times larger than the true value. This scale error will tend to drive the $R^2$ negative, but it won't affect the correlation between the model fits and the true conditional expectation. The $R^2$ is negative only because the variance of the fits is off. Relatedly, Rapach et al. (2010) show that MSE decomposes into a scale-free (correlation) component and a scale-dependent component. It is the scale-free component that is important for trading strategy performance. Leitch and Tanner (1991), Cenesizoglu and Timmermann (2012), and Rapach and Zhou (2013) also emphasize the importance of evaluating return prediction models based on their economic value in terms of trading strategy performance.

practice, when we vary the empirical model specification from simple to complex, we change how accurately the model approximates a fixed DGP.

Third, while the results discussed so far refer primarily to the case of ridgeless regression, we show that machine learning portfolios tend to incrementally benefit from moving away from the ridgeless limit by introducing non-trivial shrinkage. The bias induced by heavier ridge shrinkage lowers the expected returns to market timing, but the associated variance reduction reins in the volatility of the strategy. The Sharpe ratio tends to benefit from higher shrinkage because the variance reduction overwhelms the deterioration in expected timing returns. This is especially true when $P \approx T$, where the behavior of ridgeless regression is most vulnerable.

From a technical standpoint, we characterize the behavior of portfolios in the high-complexity regime using asymptotic analysis as the model's size grows with the number of observations at a fixed rate ($T \to \infty$ and $P/T \to c > 0$). When $P \to \infty$, the regular asymptotic results, such as laws of large numbers and central limit theorems, do not hold. Such analysis requires the apparatus of random matrix theory, on which we draw heavily to derive our results. Conceptually, this delivers an approximation of how a machine learning model behaves as we gradually increase the number of parameters holding the amount of data fixed.

We conduct an extensive empirical analysis that demonstrates the virtues of model complexity in a canonical asset pricing problem: predicting the aggregate US equity market return.[6] In particular, we study market timing strategies based on predictions from very simple models with a single parameter to extremely complex models with over 10,000 parameters (applied to training samples with as few as 12 monthly observations). The data inputs to our models are 15 standard predictor variables from the finance literature compiled by Goyal and Welch (2008). To map our data analysis to the theory, we require a method that smoothly transitions from low to high-complexity models while holding the

[6]Surveys of this large literature include Koijen and Van Nieuwerburgh (2011), Cochrane (2011), and Rapach and Zhou (2022). For early machine learning approaches to market return prediction, see Rapach et al. (2010) and Kelly and Pruitt (2013).

underlying information set fixed. The random feature method of Rahimi and Recht (2007) is ideal for this. We use it to construct expanding neural network architectures that take the Goyal and Welch (2008) predictors as inputs and maintain the core ridge regression structure of our theory.

We find extraordinary agreement between empirical patterns and our theoretical predictions. Over the standard CRSP sample 1926–2020, out-of-sample market timing Sharpe ratio improvements (relative to market buy-and-hold) reach roughly 0.47 per annum with $t$-statistics near 3.0. This is despite the fact that the out-of-sample predictive $R^2$ is substantially negative for the vast majority of models, consistent with the theoretical argument that predictive $R^2$ is inappropriate for judging the economic benefit of a machine learning model.

Timing positions from high-complexity models are remarkable. They behave similarly to long-only strategies, following the Campbell and Thompson (2008) recommendation to impose a non-negativity constraint on expected market returns. But our models learn this behavior as opposed to being handed a constraint. Moreover, machine learning strategies learn to divest leading up to NBER recessions, successfully doing so in 14 out of 15 recessions in our test sample on a purely out-of-sample basis.

This paper relates most closely to emergent literature that studies the theoretical properties of machine learning models. A number of recent papers show that linear models combined with random matrix theory help characterize the behavior of neural networks trained by gradient descent.[7] In particular, wide neural networks (many nodes in each layer) are effectively kernel regressions, and "early stopping" in neural network training is closely related to ridge regularization (Ali et al., 2019). Recent research also emphasizes the phenomenon of benign overfit and "double descent," in which expected forecast error drops in the high-complexity regime.[8]

In this literature, the closest paper to ours is Hastie et al. (2019), who derive nearly optimal error bounds in finite samples for bias and risk in the ridge(less) regression under

[7]See, for example, Jacot et al. (2018); Hastie et al. (2019); Du et al. (2018, 2019); Allen-Zhu et al. (2019).
[8]See, for example, Spigler et al. (2019); Belkin et al. (2019b,a, 2020); Bartlett et al. (2020).

very general conditions.[9] They are also the first to introduce mis-specified models where some of the signals may be unobservable. In this paper, we focus on the (easier) asymptotic regime. We use a different method of proof and relax some of the technical conditions on the distributions of signals, using recent results of Yaskov (2016). In particular, we allow for non-uniformly positive definite covariance matrices. Most importantly, instead of focusing on the prediction model forecast error variance, we characterize expected out-of-sample expected returns, volatility, and Sharpe ratios of market timing strategies based on machine learning predictions. As in Hastie et al. (2019), our key interest is in the mis-specified model. While Hastie et al. (2019) focus on a specific form of mis-specification and its ridgeless limit, we derive general expressions for asymptotic expected returns and volatility in terms of signal correlations.

Our paper also relates closely to a growing empirical literature that uses machine learning methods to analyze stock returns. The state-of-the-art market return prediction uses high-dimensional models with shrinkage and demonstrates robust out-of-sample predictive power. Rapach et al. (2010) use predictors from Goyal and Welch (2008) and forecast combination methods (which they show exert a strong shrinkage effect). Ludvigson and Ng (2007) and Kelly and Pruitt (2013) use principal components regression and partial least squares, respectively, to leverage large predictor sets for market return prediction and achieve shrinkage through dimension reduction. Dong et al. (2022) use 100 long-short "anomaly" portfolios to forecast the market return using a variety of forecasting strategies to implement shrinkage (more generally, see the recent survey by Rapach and Zhou, 2022). An emerging literature uses machine learning methods to forecast large panels of individual stock returns or portfolios, including Rapach and Zhou (2020), Kozak et al. (2020), Freyberger et al. (2020), Gu et al. (2020), and Chen et al. (Forthcoming) (also see the survey by Kelly and Xiu, 2022). Our paper offers theoretical justification for the successes of machine learning prediction documented in the asset pricing literature. Our theoretical results call for researchers to consider even larger information sets and higher-dimensional approximations

[9] See also Richards et al. (2021) who obtain less general results in an asymptotic setting (as in our paper).

to further improve return forecasts (a rationale justified by our empirical analysis). Finally, our paper is related to Martin and Nagel (2021) and Da et al. (2022) who examine market efficiency implications of the high-dimensional prediction problem faced by investors, to Fan et al. (2022b) who touch upon the "double descent" phenomenon in their analysis of structural machine learning models, and to financial econometrics applications of random matrix theory such as Fan et al. (2008), Ledoit and Wolf (2020), and Fan et al. (2022a).

The paper is organized as follows. In Section 2 we lay out the theoretical environment. Section 3 presents the foundational results from random matrix theory from which we derive our main theoretical results. Section 4 characterizes the behavior of machine learning portfolios in the correctly specified setting and emphasizes the intuition behind the portfolio benefits of high-complexity prediction models. Section 5 extends these results to the more practically relevant setting of mis-specified models. We present our main empirical results in Section 6, and Section 7 concludes. The appendix contains a variety of supplementary theoretical results and empirical robustness analyses. We invite readers that are primarily interested in the qualitative theoretical points and the empirical analysis to skip the technical material of Sections 2 and 3.

## 2 Environment

This section describes our modeling assumptions and outlines the criteria by which we evaluate machine learning portfolios.

### 2.1 Asset Dynamics

**Assumption 1** *There is a single asset whose excess return behaves according to*

$$R_{t+1} = S_t'\beta + \varepsilon_{t+1} \tag{3}$$

*with* $\varepsilon_{t+1}$ *i.i.d.,* $E[\varepsilon_{t+1}] = E[\varepsilon_{t+1}^3] = 0, E[\varepsilon_{t+1}^2] = \sigma^2$, $E[\varepsilon_{t+1}^4] < \infty$,[10] *and* $S_t$ *a* $P$*-vector of predictor variables. Without loss of generality, everywhere in the sequel, we normalize* $\sigma^2 = 1$.

Assumption 1 establishes the basic return generating process. Most notably, conditional expected returns depend on a potentially high-dimensional information set embodied by the predictors, $S$. The interpretation of this assumption is not necessarily that asset returns are subject to many fundamental driving forces. Instead, it espouses the machine learning perspective discussed in the introduction: The DGP's functional form is unknown but may be approximated with richly parameterized models using a high-dimensional nonlinear expansion $S$ of some underlying feature set.

The covariance structure of $S$ plays a central role in the behavior of machine learning predictions and portfolios. Assumption 2 imposes basic regularity conditions on this covariance.

**Assumption 2** *There exist independent random vectors* $X_t \in \mathbb{R}^P$ *with four finite first moments, and a symmetric,* $P$*-dimensional positive semi-definite matrix* $\Psi$ *such that*

$$S_t = \Psi^{1/2} X_t.$$

*Furthermore,* $E[X_{i,t}] = E[X_{i,t}^3] = 0$ *and* $E[X_{i,t}^2] = 1$, $i = 1, \cdots, P$. *Furthermore, the fourth moments* $E[X_{i,t}^4]$ *are uniformly bounded and* $X_{i,t}$ *satisfy the Lindeberg condition*

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

As we show below, the theoretical properties of machine learning portfolios depend heavily on the *distribution of eigenvalues* of $\Psi$. We are interested in limiting behavior in the high model complexity regime, i.e., as $P, T \to \infty$, with $P/T \to c > 0$. Assumption 3 ensures that estimates of $\Psi$ are well-behaved in this limit.

[10] The assumption of zero skewness does not affect our results, but simplifies the analytical expressions.

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

$$F^\Psi(x) = \frac{1}{P}\sum_{k=1}^{P} \mathbf{1}_{\lambda_k(\Psi)\leq x} \tag{4}$$

*converges to a non-random probability distribution* $H$ *supported on* $[0, +\infty)$.[11] *Furthermore,* $\Psi$ *is uniformly bounded as* $P \to \infty$. *We will use*

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

*to denote asymptotic moments of the eigenvalues of* $\Psi$.

Our last assumption governs the behavior of the true predictive coefficient, $\beta$.

**Assumption 4** *We assume* $\beta = \beta_P$ *is random,* $\beta = (\beta_i)_{i=1}^P \in \mathbb{R}^P$, *independent*[12] *of* $S$ *and* $R$, *and satisfies* $E[\beta] = 0$, *and* $E[\beta\beta'] = P^{-1}b_{*,P}I$ *for some constant* $b_{*,P} = E[\|\beta\|^2]$,[13] *and satisfies* $b_{*,P} \to b_*$ *almost surely, for some* $b_* > 0$. *Furthermore,* $E[\beta_i^4] \leq cP^{-2}$ *for some* $c > 0$, *and* $\beta$ *satisfy the same Lindeberg condition as* $X$.

The randomness of $\beta$ in Assumption 4 is a device that allows us to characterize the prediction and portfolio problem for generic predictive coefficients. The assumption that $\beta$ is mean zero is inconsequential; we could allow for a non-zero mean and restate our analysis in terms of variances rather than second moments. $E[\beta\beta'] = P^{-1}b_{*,P}I$ imposes that the predictive content of signals is rotationally symmetric. In other words, predictability is uniformly distributed across signals. This may seem restrictive, as commonly used return predictors would not satisfy Assumption 4. But it is closely aligned with the structure of feed-forward neural networks, in which raw features are mixed and nonlinearly propagated into final generated features whose ordering is essentially randomized by the initialization

[11] If 0 is in the support of $H$, then $\Psi$ is strictly degenerate, meaning that some signals are redundant.
[12] The assumption of a random coefficient vector $\beta$ is related to that in Gagliardini et al. (2016).
[13] This identity follows because $b_* = \operatorname{tr} E[\beta\beta'] = E[\operatorname{tr}(\beta\beta')] = E[b_*]$.

step of network training. Furthermore, the random feature methodology that we use in our empirical analysis satisfies Assumption 4 by construction.[14]

When $\beta$ is random and rotationally symmetric, we can focus on average portfolio behavior across signals, which implies that only the traces of the relevant matrices matter, as opposed to entire matrices (which are the source of technical intractability). The proportionality of $E[\beta\beta']$ to $P^{-1}$, and likewise the finite limiting $\ell_2$ norm of $\beta$, controls the "true" Sharpe ratio. It ensures that Sharpe ratios of timing strategies remain bounded as the number of predictors grows. In other words, our setting is one with many signals, each contributing a little bit of predictability.

A key aspect of our paper, and one rooted in Assumptions 2 and 4, is that realized out-of-sample returns are independent of the specific realization of $\beta$. This is due to a law of large numbers in the $P \to \infty$ limit and is guaranteed by the following lemma.[15]

**Lemma 1** *As $P \to \infty$ we have*

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

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

### 2.2 Timing Strategies and Performance Evaluation

We study timing strategy returns, defined as

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

[14] From a technical standpoint, it is possible to derive explicit expressions for portfolio performance without this assumption, but the expressions become more complex. In this case, the asymptotic behavior depends on the distribution of projections of $\beta$ on the eigenvectors of $\Psi$ (the signal principal components). See, Hastie et al. (2019). In particular, when $\beta$ is concentrated on the top principal components, the phenomenon of benign overfit emerges (Bartlett et al. (2020), Tsigler and Bartlett (2020)), and the optimal ridge regularization is zero. We leave this generalization for future research.

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

where $\pi_t$ is a timing weight that scales the position in the asset up and down to exploit time-varying in the asset's expected returns.

We are interested in timing strategies that optimize the unconditional Sharpe ratio,

$$SR = \frac{E[R^{\pi}_{t+1}]}{\sqrt{E[(R^{\pi}_{t+1})^2]}}. \tag{5}$$

While there are other possible performance criteria, we focus on this for its simplicity and ubiquity. It is implied by the quadratic utility function at the foundation of mean-variance portfolio theory. Academics and real-world investors rely nearly universally on the unconditional Sharpe ratio when evaluating empirical trading strategies. The use of centered versus uncentered second moment in the denominator is without loss of generality.[16]

Our analysis centers on the following timing strategy functional form:

$$\pi_t(\beta) = S_t'\beta. \tag{6}$$

This strategy takes positions equal to the asset's conditional expected return. Note that this timing strategy optimizes the *conditional* Sharpe ratio. It achieves the same Sharpe ratio as the conditional Markowitz solution, $\pi_t^{\text{Cond. MV}} = E_t[R_{t+1}]/\text{Var}_t[R^2_{t+1}] = S_t'\beta$, according to equation (3). While strategy $\pi_t$ is conditionally mean-variance efficient, it is not the optimizer of the unconditional objective in (5), which takes the form $\pi_t^{\text{Uncond. MV}} = S_t'\beta/(1+(S_t'\beta)^2)$.[17] In the proof of Proposition 1 in the Appendix, we show that $\pi_t$ in equation (6) and $\pi_t^{\text{Uncond. MV}}$ are equal up to third-order terms.[18] We study $\pi_t = S_t'\beta$ for the simplicity of its linearity in both $\beta$ and $S_t$, but note that our conclusions are identical for $\pi_t^{\text{Uncond. MV}}$ because, in the limit as $P \to \infty$, the normalization factor $1+(S_t'\beta)^2$ converges to a constant.[19]

[16] Define $\widetilde{SR} = \frac{E[R^{\pi}_{t+1}]}{\sqrt{\text{Var}[(R^{\pi}_{t+1})]}}$. Direct calculation yields $SR = \frac{1}{\sqrt{1+\widetilde{SR}^{-2}}}$.

[17] See Hansen and Richard (1987); Ferson and Siegel (2001); Abhyankar et al. (2012).

[18] In particular, the Sharpe ratio in equation (5) is less than one due to the Cauchy-Schwarz inequality. We show the difference in Sharpe ratios for $\pi_t$ versus $\pi_t^{\text{Uncond. MV}}$ is on the order of the Sharpe ratio cubed.

[19] By a version of Lemma 1, $1+(S_t'\beta)^2 \to 1 + b_*\psi_{*,1}$.

Proposition 1 states the behavior of timing strategy $\pi_t = S_t'\beta$ when $T \to \infty$ and $P/T \to 0$ (i.e., when the predictive parameter $\beta$ is known).

**Proposition 1 (Infinite Sample)** *The unconditional first and second moments of returns to the infeasible market timing strategy $\pi_t = S_t'\beta$ are*

$$E[\pi_t R_{t+1}] \to b_* \psi_{*,1} > 0 \quad \textit{and} \quad E\left[(\pi_t R_{t+1})^2\right] \to \left(3(b_* \psi_{*,1})^2 + b_* \psi_{*,1}\right).$$

*The infeasible market timing Sharpe ratio is*

$$SR \to \frac{1}{\sqrt{3 + (b_* \psi_{*,1})^{-1}}} < \left(\frac{1}{3}\right)^{1/2}. \tag{7}$$

For comparison, under Assumptions 1 to 4, the unconditional first and second moments of the un-timed asset return are (see Lemma 1)

$$E[R_{t+1}] = 0, \quad \text{and} \quad E[R_{t+1}^2] \to 1 + b_* \psi_{*,1}.$$

That is, our assumptions imply the un-timed asset has a zero Sharpe ratio. This is just a normalization so that any positive market timing Sharpe ratio can be interpreted as pure excess performance arising from timing ability.

### 2.3 Relating Predictive Accuracy to Portfolio Performance

We are ultimately interested in understanding the portfolio properties of a feasible timing strategy, $\hat{\pi}_t = \hat{\beta}' S_t$. This is, of course, intimately tied to the prediction accuracy of the estimator $\hat{\beta}$, summarized by its expected mean square forecast error ($MSE$) on an independent test sample. This is the fundamental notion of estimator "risk" from statistical theory, though we use the term "$MSE$" here to avoid confusion with portfolio riskiness. We

can write $MSE$ as

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

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

Note that a strategy $\pi_t = \beta' S_t$ based on the infeasible true $\beta$ satisfies $E[\pi_t R_{t+1}] = E[\beta'\Psi\beta] = E[\pi_t^2]$.[20] In this case, the $MSE$ collapses to $E[R_{t+1}^2] - E[\pi_t R_{t+1}]$ and is minimized, meaning that the leverage taken is exactly justified by the predictive benefits of the strategy. This can also be stated in terms of the infeasible $R^2$ based on equation (3) and Lemma 1:

$$R^2 = \frac{\beta'\Psi\beta}{\beta'\Psi\beta + 1} \to \frac{b_*\psi_{*,1}}{b_*\psi_{*,1} + 1}.$$

Thus, there is a monotonic mapping from the infeasible timing strategy expected return to the true $R^2$, and from the infeasible Sharpe ratio to the true $R^2$ (see equation (7)).

## 3 Machine Learning and Random Matrices

The central premise of machine learning is that large data sets can be used in flexible model specifications to improve prediction. This can be understood in the environment above by considering the regime in which the number of predictors, $P$, is large, perhaps even larger than $T$. Our main objective is thus to understand the behavior of optimal timing portfolios

[20] Indeed, $E[(\beta' S_t)^2] = E[\beta' S_t S_t'\beta] = \beta'\Psi\beta$.

as the prediction model becomes increasingly complex, i.e., when $P \to \infty$. Because this involves estimating infinite-dimensional parameters, traditional large $T$ asymptotics do not apply, and we instead resort to random matrix theory. In this section, we discuss the ridge estimator and present random matrix theory results at the foundation of our theoretical characterization of high-complexity timing strategies.

### 3.1 Least Squares Estimation

Throughout, we analyze (regularized) least squares estimators taking the form

$$\hat{\beta}(z) = \left( zI + T^{-1} \sum_t S_t S_t' \right)^{-1} \frac{1}{T} \sum_t S_t R_{t+1}$$

for a given ridge shrinkage parameter, $z$. The ridge-regularized form is necessary for characterizing $\hat{\beta}(z)$ in the high-complexity regime, $P/T \to c > 1$, though we will see it also has important implications for the behavior of $\hat{\beta}(z)$ when $P/T < 1$.[21]

Consider first the ordinary least squares (OLS) estimator, $\hat{\beta}(0)$. As $P$ approaches $T$ from below, the denominator of the least squares estimator approaches the singularity. This produces explosive variance of $\hat{\beta}(0)$ and, in turn, explosive forecast error variance. As $P \to T$, the model begins to fit the data with zero error, so a common interpretation of the explosive variance of $\hat{\beta}(0)$ is an insidious overfit that does not generalize out-of-sample.

When $P$ moves beyond $T$, there are more parameters than observations and the least squares problem has multiple solutions. A particularly interesting solution invokes the Moore-Penrose pseudo-inverse, $(T^{-1} \sum_t S_t S_t')^+ \frac{1}{T} \sum_t S_t R_{t+1}$.[22] This solution is equivalent

[21] One could alternatively analyze "sparse" least squares models that combine shrinkage with variable selection (e.g., based on LASSO). First, recent evidence of Giannone et al. (2021) suggests sparsity of predictive relationships in economics and finance is likely an illusion. Second, our empirical focus is on non-parametric models that seek to approximate a generic nonlinear function as a linear combination of generated features, and sparsity in the generated feature space is difficult to identify (see, e.g., Ghorbani et al., 2020). Third, analysis with $\ell_1$ shrinkage is significantly more taxing from a theoretical standpoint. We thus leave sparse least squares models to future research.

[22] Recall that the Moore-Penrose pseudo-inverse $A^+$ of a matrix $A$ is defined via $A^+ = (A'A)^{-1}A'$ if $A'A$ is invertible, and $A^+ = A'(AA')^{-1}$ if $AA'$ is invertible.

to the ridge estimator as the shrinkage parameter approaches zero:

$$\hat{\beta}(0^+) = \lim_{z \to 0+} \left( zI + T^{-1} \sum_t S_t S_t' \right)^{-1} \frac{1}{T} \sum_t S_t R_{t+1}.$$

The solution $\hat{\beta}(0^+)$ is often referred to as the "ridgeless" regression estimator. When $P < T$, OLS is the ridgeless estimator. At $P = T$ there is still a unique least squares solution, yet the model can exactly fit the training data (for this reason, $P = T$ is called the "interpolation boundary"). When $P > T$, the ridgeless estimator is one of many solutions that exactly fit the training data, but among these, it is the only solution that achieves the minimum $\ell_2$ norm $\hat{\beta}(z)$ (Hastie et al., 2019). The machine learning literature has recently devoted substantial attention to understanding ridgeless regression in the high-complexity regime. The counter-intuitive insight from this literature is that, beyond the interpolation boundary, allowing the model to become *more* complex in fact *regularizes* the behavior of least squares regression despite using infinitesimal shrinkage. We explore the implications of this idea for market timing in the subsequent sections.

### 3.2 The Role of Random Matrix Theory

We analyze the behavior of $\hat{\beta}(z)$ and associated market timing strategies in the limit as $P \to \infty$. This is possible due to a remarkable connection between ridge regression and random matrix theory.

In regression analysis, the sample covariance matrix of signals, $\hat{\Psi} \coloneqq T^{-1} \sum_t S_t S_t'$, naturally plays a central role. But no general characterization exists for the behavior of $\hat{\Psi}$ in the limit as $P, T \to \infty$. However, the tools of random matrix theory characterize one aspect of $\hat{\Psi}$—the distribution of its eigenvalues. Fortunately, as we show, the prediction and portfolio performance properties of least squares estimators rely only on the eigenvalue distribution of $\hat{\Psi}$. Thus random matrix theory facilitates a rich understanding of machine learning portfolios. Here we elaborate on the core results from the random matrix theory we build upon.

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

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

where $\lambda_i(\Psi)$ are the eigenvalues of $\Psi$. Using formula (4), we can rewrite this identity as

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

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

$$m_{\Psi}(z) \coloneqq \int \frac{1}{x - z} dH(x) = \lim_{P\to\infty} \frac{1}{P}\operatorname{tr}\left((\Psi - zI)^{-1}\right). \tag{9}$$

The function $m_{\Psi}(z)$ is the *limiting* Stieltjes transform of the eigenvalue distribution of $\Psi$. Equation (9) is a powerful step towards understanding the least squares estimator in the machine learning regime (and hence machine learning predictions and portfolios). It states that key properties of the limiting inverse of the ridge-regularized signal covariance matrix can be completely characterized if we just know $\Psi$'s eigenvalue distribution.

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

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

In traditional finite $P$ statistics, we would have convergence between the sample covariance $\hat{\Psi}$ and the true covariance $\Psi$ as $T \to \infty$. One might be tempted to think that $\lim_{P\to\infty} \frac{1}{P} \operatorname{tr}\left((\hat{\Psi} - zI)^{-1}\right)$ and $\lim_{P\to\infty} \frac{1}{P} \operatorname{tr}\left((\Psi - zI)^{-1}\right)$ also converge as $T \to \infty$. But this is not the case. The limiting eigenvalue distributions of $\hat{\Psi}$ and $\Psi$ remain divergent in the limit as $T \to \infty$ if $P/T \to c > 0$. Here we see a first glimpse of the complexity of machine learning and how we can understand it with random matrix theory. In the Appendix (see Theorem 8), we show how $m(-z; c)$ can be computed from $m_\Psi(-z)$ using results of Silverstein and Bai (1995) and Bai and Zhou (2008). In particular, $m(-z; c) > m(-z; 0) = m_\Psi(-z)$ for all $c > 0$.[23] The next result shows that, quite remarkably, if we constrain ourselves to linear ridge regression estimators, all asymptotic expressions depend only on $m(z; c)$ and do not require $m_\Psi$.[24]

**Proposition 2** *We have*

$$\lim_{T\to\infty} \frac{1}{T} \operatorname{tr}((zI + \hat{\Psi})^{-1}\Psi) \;\to\; \xi(z; c) \tag{11}$$

*almost surely, where*

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

The quantity $\operatorname{tr} E[(zI + \hat{\Psi})^{-1}\Psi]$ appears in virtually every expression we analyze to describe portfolio behavior. It depends on an interaction between the sample and true signal covariance matrix and arises in the computation of both the expected return and leverage of

[23] Theorem 8 in the Appendix is a generalized version of the Marčenko and Pastur (1967) theorem that accommodates non-i.i.d. $S_t$. When signals are i.i.d. with $\Psi = I$ and $m_\Psi(z) = (1 - z)^{-1}$, Marčenko and Pastur (1967) show that

$$m(-z; c) = \frac{-((1 - c) + z) + \sqrt{((1 - c) + z)^2 + 4cz}}{2cz}. \tag{10}$$

By direct calculation, (10) is indeed the unique positive solution to (26) when $m_\Psi(z) = (1 - z)^{-1}$. While the eigenvalue distributions of the sample and true covariance matrices do not coincide, Theorem 8 describes the precise nonlinear way they relate to each other. In particular, when $P > T$, the matrix $\hat{\Psi}$ has $P - T$ zero eigenvalues and therefore, $P^{-1} \operatorname{tr}\left((zI + \hat{\Psi})^{-1}\right)$ contains a singular part, $P^{-1}(P - T)z^{-1} = (1 - c^{-1})z^{-1}$.

[24] It is possible to develop *nonlinear* shrinkage estimators analogous to those developed by Ledoit and Wolf (2020) for covariance matrices. Such estimators would require knowledge of the true eigenvalue distribution of $\Psi$ which can be recovered from $m(z; c)$ using equation (26).

the timing strategy (see equation (8)). One would imagine, then, that we need to know the limiting eigenvalue distribution of both matrices (or their Stieltjes transforms, $m$ and $m_\Psi$) to describe $\operatorname{tr} E[(zI + \hat{\Psi})^{-1}\Psi]$. Proposition 2 shows that this is not the case—we only need to know the empirical version, $m(-z; c)$. This is a powerful result. It will allow us to quantify the expected out-of-sample behavior of machine learning portfolios based only on the eigenvalue distribution of the sample signal covariance $\hat{\Psi}$ (which is observable) without requiring us to know the eigenvalues of $\Psi$.[25]

We refer to the constant $c$ as "model complexity," which (as the preceding results show) plays a critical role in understanding model behavior. It describes the limiting ratio of predictors to data points: $P/T \to c$. When $T$ grows at a faster rate than the number of predictors (i.e., $c \to 0$) the limiting eigenvalue distributions of $\hat{\Psi}$ and $\Psi$ in fact converge: $m(-z; 0) = m_\Psi(-z)$. As $c$ becomes positive, these distributions fail to converge, and their divergence is wider for larger $c$. It is, therefore, clear that the behavior of the least squares estimator in the machine learning regime will differ from the true coefficient, even when $T \to \infty$, as long as $c > 0$. As a result, machine learning portfolios will suffer relative to the infeasible performance in Proposition 1 despite abundant data. However, while machine learning portfolios underperform the infeasible strategy, they can continue to generate substantial trading gains. This is true even in the ridgeless case. Additional ridge shrinkage can boost performance even further. In the following sections, we precisely characterize these behaviors.

## 4 Prediction and Performance in the Machine Learning Regime

In this section, we analyze correctly specified models. We present the theoretical characterizations of machine learning models in terms of prediction accuracy and portfolio performance. We then illustrate their behavior in a calibrated theoretical setting.

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

### 4.1 Expected Out-of-sample $R^2$

To understand a model's prediction accuracy in the high-complexity regime, we study its limiting $MSE$, defined as

$$MSE(z;c) = \lim_{T,P\to\infty,\ P/T\to c} E\left[\left(R_{t+1} - S_t'\hat{\beta}(z)\right)^2 |\hat{\beta}(z)\right]. \tag{12}$$

Notably, while $\hat{\beta}(z)$ is random and depends on the sample realization, we show below that the limit in (12) is non-random. The arguments $z$ and $c$ are central to understanding the limiting predictive ability of least squares. Respectively, they describe the extent of ridge shrinkage and the complexity of the DGP (and thus of the correctly specified model).

In finance and economics, it is common to state predictive performance in terms of $R^2$ rather than $MSE$. We denote the limiting out-of-sample $R^2$ as

$$R^2(z;c) = 1 - \frac{MSE(z,c)}{\lim_{T,P\to\infty} E[R_{t+1}^2]},$$

where $E[R_{t+1}^2]$ is the null $MSE$ when $\beta = 0$.

In Section 2.3, we discussed the infeasible maximum $R^2$, or

$$R^2(0;0) = \frac{b_*\psi_{*,1}}{1 + b_*\psi_{*,1}}.$$

This corresponds to a data-rich environment ($c = 0$, so observations vastly outnumber parameters) and OLS regression ($z = 0$). $R^2(0;0)$ is the benchmark for evaluating the loss of predictive accuracy due to high model complexity, even when data is abundant. Specifically, the $R^2$ of the least squares estimator in the machine learning regime behaves as follows.

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

$$\begin{aligned}
&\mathcal{E}(z;c) = \lim\ E[\hat{\pi}_t R_{t+1} | \hat{\beta}(z)] = b_* \nu(z;c) \\
&\mathcal{L}(z;c) = \lim\ E[\hat{\pi}_t^2 | \hat{\beta}(z)] = b_* \hat{\nu}(z;c) - c\nu'(z;c) \\
&R^2(z;c) = \ \frac{2\mathcal{E}(z;c) - \mathcal{L}(z;c)}{1 + b_* \psi_{*,1}}
\end{aligned} \tag{13}$$

*where*

$$\begin{aligned}
\nu(z;c) \ &= \psi_{*,1} - c^{-1} z \xi(z;c) &&= \lim P^{-1} \operatorname{tr}(\hat{\Psi}(zI + \hat{\Psi})^{-1}\Psi) &&> \ 0 \\
\nu'(z;c) \ &= \ -c^{-1}(\xi(z;c) + z\xi'(z;c)) &&= \ -\lim P^{-1} \operatorname{tr}(\hat{\Psi}(zI + \hat{\Psi})^{-2}\Psi) &&< \ 0 \\
\hat{\nu}(z;c) \ &= \nu(z;c) + z\nu'(z;c) &&= \lim P^{-1} \operatorname{tr}(\hat{\Psi}^2(zI + \hat{\Psi})^{-2}\Psi) &&> \ 0.
\end{aligned}$$

*As we show in the Appendix, these limits exist in probability.*

*Furthermore,* $R^2(z;c)$ *is monotone increasing in* $z$ *for* $z < z_* = c/b_*$, *and decreasing in* $z$ *for* $z > z_*$. $R^2(z;c)$ *attains its maximum at* $z_* = c/b_*$, *where it is positive and given by*

$$R^2(z_*;c) = R^2(0;0) - \frac{\xi(z_*;c)}{1 \ + \ b_* \psi_{*,1}} = \frac{b_* \nu(z_*;c)}{1 \ + \ b_* \psi_{*,1}} \ > \ 0\,.$$

*In the ridgeless limit, assuming* $H(0+) = 0$, *we have*

$$R^2(0;c) = R^2(0;0) \ - \ (1 \ + \ b_* \psi_{*,1})^{-1} \begin{cases} (c^{-1} - 1)^{-1}, & c < 1 \\ \mu(c), & c > 1. \end{cases} \tag{14}$$

*with some* $\mu(c) > 0,\ \mu(1+) = +\infty$. *Lastly, we have*

$$\lim_{c \to \infty} R^2(0;c) = 0 \ > \ \lim_{c \to 1} R^2(0;c) = -\infty\,. \tag{15}$$

When the prediction model is complex ($c > 0$), the limiting eigenvalues of $\hat{\Psi}$ and $\Psi$ diverge, and this unambiguously reduces the predictive $R^2$ relative to the infeasible best,

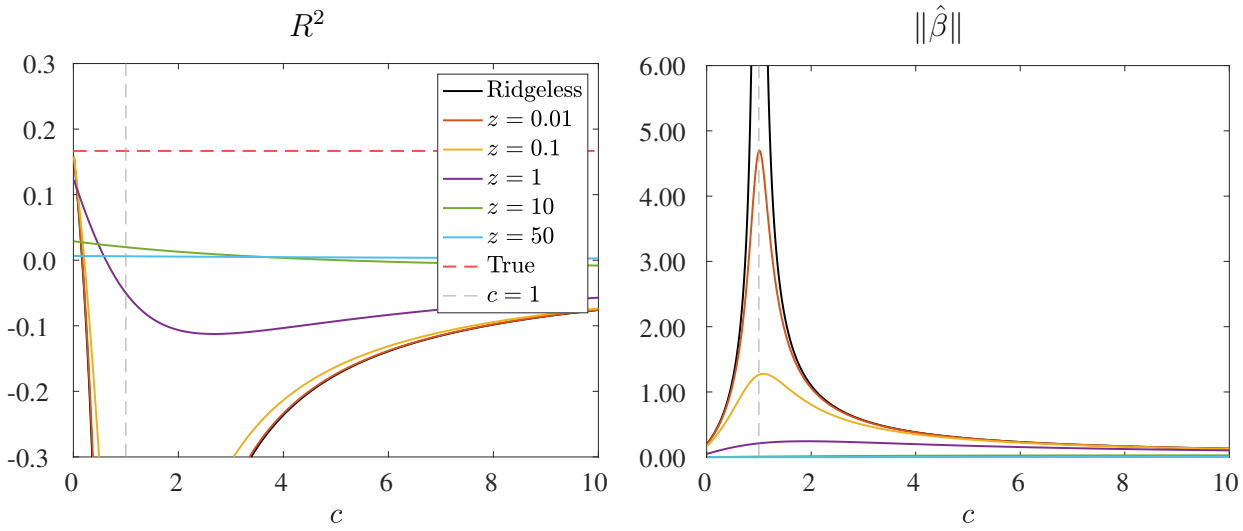



**Figure 1:** Expected Out-of-sample $R^2$ and Norm of Least Squares Coefficient

**Note.** Limiting out-of-sample $R^2$ and $\hat{\beta}$ norm as a function of $c$ and $z$ from Proposition 3 assuming $\Psi$ is the identity matrix and $b_* = 0.2$.

$R^2(0;0)$. Intuitively, because the frictionless $R^2(0;0)$ is fixed, as $c$ increases, the investor must learn the same amount of predictability but spread across many sources, and this dimensionality expansion hinders statistical inference. The degradation in predictive accuracy due to complexity can be so severe that expected out-of-sample $R^2$ becomes extremely negative, particularly in the ridgeless case. Shrinkage can mitigate this and help preserve accuracy amid complexity. Shrinkage controls variance but introduces bias. Proposition 3 points out that the amount of shrinkage that optimizes the bias-variance tradeoff is $z_* = c/b_*$.[26] More complex settings benefit from heavier shrinkage, while settings with higher signal-to-noise ratio (higher $b_*$) benefit from lighter shrinkage (see, e.g. Hastie et al., 2019). $\mathcal{E}$ and $\mathcal{L}$ are the limiting out-of-sample expected returns and leverage of the timing strategy. Proposition 3 shows that these are the main determinants of out-of-sample $R^2$.

Figure 1 illustrates the theoretical behavior of the least squares estimator derived in Proposition 3. The plots set $\Psi$ to the identity matrix and fix $b_* = 0.2$ (recall $\sigma^2$ is

[26]Note that the optimal shrinkage must be inferred from an estimate of $b_*$. Our theoretical and empirical results indicate a general insensitivity of prediction and timing strategy performance to the choice of $z$ in the high-complexity regime. Because of this, simple shrinkage selection methods like cross-validation tend to perform well.

normalized to one). The left panel draws the expected out-of-sample $R^2$ as a function of model complexity $c$ (shown on the $x$-axis) and ridge penalty $z$ (different curves). In this calibration, the infeasible maximum predictive $R^2$ (that uses the true parameter values) is the dotted red line and provides a reference point. Throughout the paper, we refer to plots like these, which describe the model performance as a function of model complexity, as "VoC curves."

The blue line shows the $R^2$ in the ridgeless limit. When $c \leq 1$, the ridgeless limit corresponds to exactly $z = 0$ (i.e., OLS). On this side of $c = 1$, predictive accuracy deteriorates rapidly as model complexity increases. This captures the well-known property that OLS suffers when the number of predictors is large relative to the number of data points. As $c \to 1$, the denominator of the OLS estimator approaches the singularity, and the expected out-of-sample $R^2$ dives.

To the right of $c = 1$, the number of predictors exceeds the sample size, and the "ridgeless" case is defined as the limit as $z \to 0$ (i.e., when the least squares denominator is calculated via the pseudo-inverse of $\hat{\Psi}$). Counter-intuitively, the $R^2$ begins to *rise* as model complexity increases.[27]

The reason is that, while there are many equivalent $\beta$ solutions that exactly fit[28] the training data when $c > 1$, ridgeless regression selects the solution with the smallest norm. As complexity increases, there are more solutions for ridgeless regression to search over, and thus it can find smaller and smaller betas that still exactly fit the training data. This acts as shrinkage, biasing the beta estimate toward zero. Due to this bias, the forecast variance drops, improving the $R^2$. In other words, despite $z \to 0$, the ridgeless solution still regularizes the least squares estimator, and more so, the larger is $c$. This property of ridgeless least squares is a newly documented phenomenon in the statistics literature and is still an emerging topic of research.[29] It shows that even in very simple data generating

[27] This is an illustration of what the statistics literature refers to as benign overfitting.

[28] That is, $\beta' S_t = R_{t+1}$ for all $t \in [1, \cdots, T]$.

[29] See Spigler et al. (2019), Belkin et al. (2019b), Belkin et al. (2019a), Belkin et al. (2020), and Hastie et al. (2019).

processes, one may be able to improve the accuracy of return forecasts by pushing model dimensionality well beyond sample size.

The remaining curves in Figure 1 show how the out-of-sample $R^2$ is affected by non-trivial ridge shrinkage. Allowing $z > 0$ improves $R^2$ except at very low levels of complexity. This is again a manifestation of the bias-variance tradeoff. When $z > 0$, the norm of $\hat{\beta}$ is controlled, and the associated variance reduction outweighs the effects of bias when the model is complex.

It is useful to place our analysis thus far in the context of the literature. Some formulas of Propositions 2 and 3 have been established in papers on random matrix theory (e.g. Ledoit and Péché, 2011). Hastie et al. (2019) prove an analog of Proposition 3 allowing for arbitrary $\beta$ and expressing all quantities in terms of the distribution of projections of $\beta$ onto the eigenvectors of $\Psi$ (see also Wu and Xu, 2020). Furthermore, they establish non-asymptotic bounds on the rate of convergence. However, both Hastie et al. (2019) and Wu and Xu (2020) require that $\Psi$ is strictly positive definite. By contrast, in our data analysis, we find that $\Psi$ is nearly degenerate. Richards et al. (2021) also allow for more general $\beta$ structures and $\Psi$ matrices, but require that $X_t$ be i.i.d. Gaussian and Dobriban and Wager (2018) require $X_t$ be i.i.d. This is clearly not applicable to the RFFs used in our empirical analysis (or any other nonlinear signal transformations). In contrast to these papers, we establish our results under much weaker conditions on the distribution of $X_{i,t}$ across $i$. This is important for practical applications, where neither the independence of $X_t$ nor equality (or boundedness) of their higher moments can be guaranteed. Lastly, the novel techniques we develop allow us to characterize the out-of-sample performance of mis-specified models. To the best of our knowledge, this characterization is new in the literature (see Section 5).

Our main theoretical contribution is in the subsequent sections, where we derive portfolio performance properties.

### 4.2 Expected Out-of-sample Market Timing Performance

Next, we analyze the behavior of market timing based on the least squares estimate:

$$\hat{\pi}_t(z) = \hat{\beta}(z)' S_t.$$

Formula (13) derives the expected return of this strategy. The following proposition characterizes the expected out-of-sample risk-return tradeoff of market timing in the high-complexity regime.

**Proposition 4** *In the limit when $P,\ T \to \infty,\ P/T \to c$, the limiting second moment of the market timing strategy is*

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

*in probability, with $\mathcal{E}$ and $\mathcal{L}$ given in (13). As a result, the Sharpe ratio satisfies*

$$SR(z;c) = \frac{\mathcal{E}(z;c)}{\sqrt{\mathcal{V}(z;c)}} = \frac{1}{\sqrt{2 + (1 + b_* \psi_{*,1}) \frac{\mathcal{L}(z;c)}{(\mathcal{E}(z;c))^2}}}. \tag{16}$$

*Furthermore, we have:*

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

*ii) $SR(z;c)$ is monotone increasing in $z$ for $z < z_* = c/b_*$ and monotone decreasing in $z$ for $z > z_* = c/b_*$. Thus, the maximal Sharpe ratio is given by*

$$SR(z_*;c) = \frac{1}{\sqrt{2 + (1 + b_* \psi_{*,1}) \frac{1}{b_* \nu(z_*;c)}}} < SR(0;0)\,, \tag{17}$$

*where $\mathcal{E}(0,0)$ and $SR(0,0)$ are the infeasible market timing expected return and Sharpe ratio from Proposition 1.*

The left panel of Figure 2 plots the expected out-of-sample return and the right panel

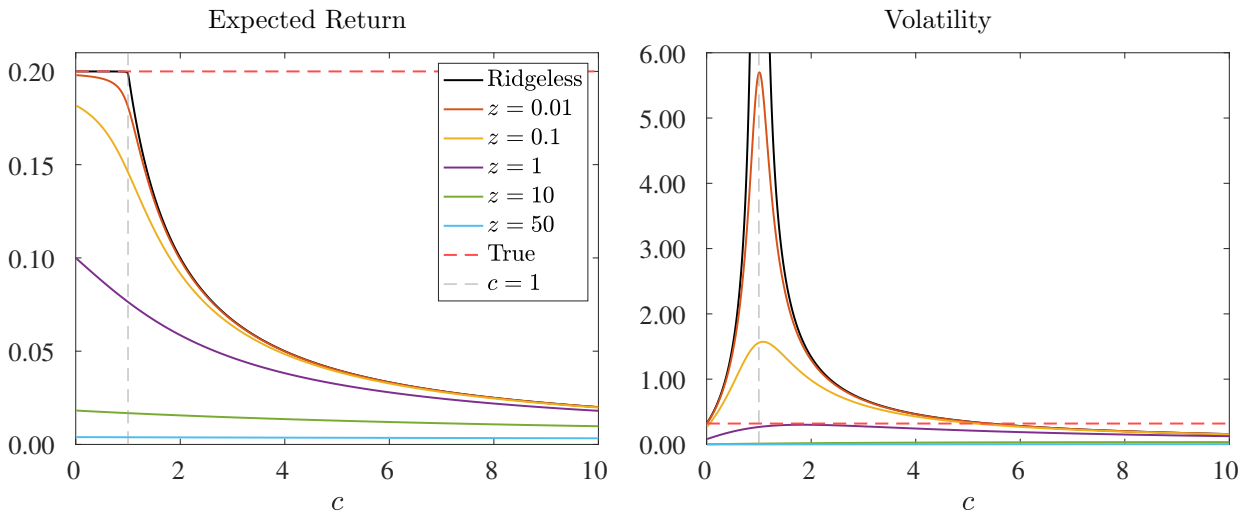



**Figure 2:** Expected Out-of-sample Risk and Return of Market Timing

**Note.** Limiting out-of-sample expected return and volatility of the market timing strategy as a function of $c$ and $z$ from Proposition 3 assuming $\Psi$ is the identity matrix and $b_* = 0.2$.

plots the expected out-of-sample volatility based on Propositions 3 and 4 using the same calibration as Figure 1. Again, the ridgeless case is in blue. The expected returns of least squares timing strategies are always positive because they are quadratic in beta. When $c < 1$ (i.e., in the OLS case), the ridgeless timing strategy achieves the true expected return even though the corresponding $R^2$ is significantly negative in much of this range. The fact that the out-of-sample expected return is unimpaired reflects the unbiasedness of OLS, while the declining $R^2$ reflects the increasing forecast variance as $c$ rises toward one. The return volatility of the timing strategy is likewise increasing in $c$ for $c \in [0, 1]$ due to the rising forecast variance and maxes out at $c = 1$.

When $c > 1$, the ridgeless expected return begins to deteriorate. This is more subtle and is related to the rising $R^2$ discussed above. When model complexity is high, the multiplicity of least squares solutions allows ridgeless regression to find a low norm beta that exactly fits the training data. So, even though $z \to 0$, the ridgeless beta is biased, and the expected return of the strategy falls. At the same time, the volatility of the strategy falls.

The other expected return and volatility curves show that the bias induced by a non-trivial ridge penalty eats into the timing strategy even for $c < 1$. But the bright side of

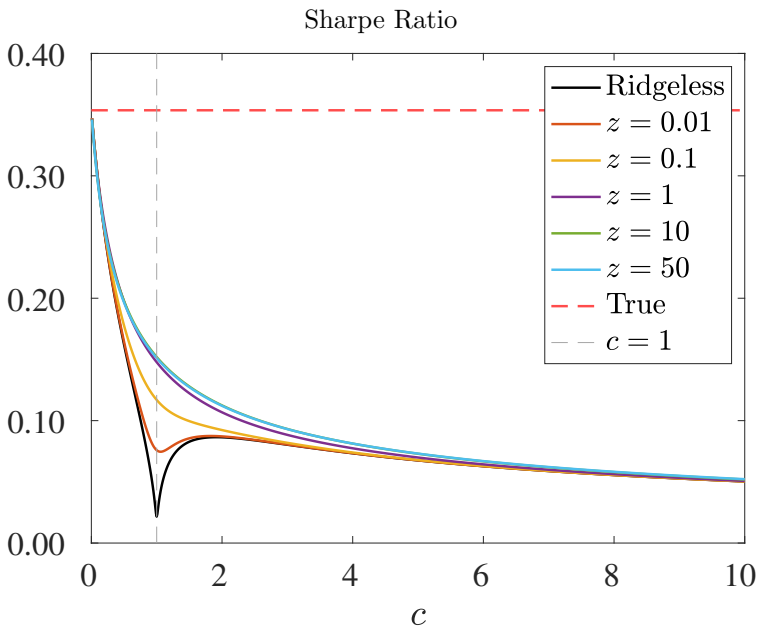



**Figure 3:** Expected Out-of-sample Sharpe Ratio of Market Timing

**Note.** Limiting out-of-sample Sharpe ratio of the market timing strategy as a function of $c$ and $z$ from Proposition 3 assuming $\Psi$ is the identity matrix and $b_* = 0.2$.

this attenuation is a reduction in the strategy's riskiness. For relatively high shrinkage levels like $z = 1$, the volatility of the timing strategy drops even below that of the infeasible best strategy while maintaining a meaningfully positive expected return.

The net effect of these expected return and volatility behaviors is summarized by the market timing strategy's expected out-of-sample Sharpe ratio, given in Proposition 4. The calibrated Sharpe ratio is shown in Figure 3. Recall that the buy-and-hold Sharpe ratio is normalized to zero. The key implication of Proposition 4 is that despite the sometimes massively negative predictive $R^2$, the ridgeless Sharpe ratio is everywhere positive, even for extreme levels of model complexity. At $c = 1$, the Sharpe ratio drops to near zero, not because the strategy is unprofitable (it remains maximally profitable in an expected return sense) but because its volatility explodes.

Another interesting aspect of Figure 3 is that the Sharpe ratio benefits from non-trivial ridge shrinkage regardless of model complexity. Shrinkage is most valuable near $c = 1$, where it reins in volatility substantially more than it reduces expected return. At both low levels

of complexity ($c \approx 0$) and high levels of complexity ($c >> 1$), the Sharpe ratio is relatively insensitive to $z$.

Proposition 4 also implies that when the model is correctly specified, the shrinkage that optimizes the expected out-of-sample $R^2$ also optimizes the Sharpe ratio. This is convenient because it means that one can focus on tuning the prediction model and be confident that the tuned $z$ will optimize timing performance. But two caveats are in order. The first is that this statement applies to the Sharpe ratio, so if investors judge their performance with other criteria, then other levels of shrinkage may be optimal. For example, a risk-neutral investor prefers ridgeless regression despite its comparatively poor performance in $R^2$. Second, this statement requires correct specification. If the empirical model is mis-specified, the optimal amount of shrinkage can differ depending on whether the objective is to maximize out-of-sample $R^2$ or Sharpe ratio.

### 4.3 A Note on $R^2$

At this point, we already see that a timing strategy with negative $R^2$ can have high average out-of-sample returns and thus positive out-of-sample Sharpe ratios.[30] More plainly, the positivity of out-of-sample $R^2$ is *not* a necessary condition for an economically valuable timing strategy. The least squares timing strategies in our framework all have strictly positive out-of-sample expected return and Sharpe ratio regardless of shrinkage or model complexity (despite having enormously negative $R^2$ in many cases).

This is an important contrast versus the mapping from $R^2$ to the timing Sharpe ratio proposed by Campbell and Thompson (2008), which is an often-used heuristic for interpreting the economic benefits of a predictive $R^2$. Their mapping is population mapping, meaning it corresponds to the special case of an analyst using a correctly specified model with $c = 0$ (i.e., infinitely more data than parameters). In contrast, our analysis characterizes expected

[30] To see this in a simple example, consider a model with one predictor and imagine estimating a predictive coefficient that happens to be a large scalar multiple of the truth. In this case, the $R^2$ will be pushed negative, but the predictions will be perfectly correlated with the true expected return. Thus, the expected return of the timing strategy will be positive. Furthermore, because the Sharpe ratio is independent of scale effects, this timing strategy will equal the actual Sharpe ratio of the DGP.

out-of-sample $R^2$ and Sharpe ratios for generic $c$ and even with mis-specified models (Section 5).

Out-of-sample $R^2$ and Sharpe ratio measurements serve different purposes. $R^2$ helps evaluate forecast accuracy. The Sharpe ratio is appropriate for evaluating the economic value of forecasts in asset allocation contexts. Much of the empirical literature in return prediction and market timing focuses its evaluations on out-of-sample predictive $R^2$ (see, e.g. Goyal and Welch, 2008). Proposition 4 ensures that we can worry less about the positivity of out-of-sample $R^2$ from a prediction model and focus more on the out-of-sample performance of timing strategies based on those predictions.

## 5 Machine Learning and Model Mis-specification

So far, we have studied the behavior of machine learning portfolios as a function of the complexity of the true DGP while assuming we have the correctly specified model. Under correct specification, the complexity comparative statics in Figures 1 to 3 change both the empirical and the true model as we vary $c$. So, these theoretical comparative statics cannot be taken to the data. Nevertheless, theory grounded on correct model specification is powerful for developing a conceptual understanding of machine learning portfolios.

A more empirically relevant theoretical setting would consider a single true DGP. Then, it would consider empirical models that are always a misspecified approximation to this DGP. Finally, it would make comparisons by increasing the complexity of the empirical model to achieve an increasingly accurate approximation of the true DGP. We will develop this theory now.

We consider a true DGP with $P$ predictors. We consider an expanding set of empirical models to approximate the DGP. Each model is indexed by $P_1 = 1, \cdots, P$ and corresponds to an economic agent observing only a subset of the signals, $S_t^{(1)} = (S_{i,t})_{i=1}^{P_1}$. We use $S_t^{(2)} = (S_{i,t})_{i=P_1+1}^{P}$ to denote the remaining unobserved signals. The signal covariance matrix

corresponding to this partition is

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

Naturally, mis-specified estimator behavior depends on the correlation structure of observed and unobserved signals captured by the off-diagonal blocks of $\Psi$.

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

**Assumption 5** *For any sequence $P_1 \to \infty$ such that $P_1/P = q > 0$, the eigenvalue distribution of the matrix $\Psi_{1,1}$ converges to a non-random probability distribution $H(x;q)$. We say that signals are sufficiently mixed if $H(x;q)$ is independent of $q$. We will also use*

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

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

In a mis-specified model, the (regularized) least squares estimator is

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

where

$$\hat{\Psi}_{1,1} = T^{-1} \sum_t S_t^{(1)} (S_t^{(1)})' \ \in \ \mathbb{R}^{P_1 \times P_1}.$$

We also introduce the following auxiliary objects:

$$\begin{aligned} \xi_{2,1}(z;cq;q) &= \lim_{T\to\infty} T^{-1} \operatorname{tr} E[(zI + \hat{\Psi}_{1,1})^{-1} \Psi_{1,2} \Psi_{1,2}'] \ \geq \ 0 \\ \widehat{\xi}_{2,1}(z;cq;q) &= \lim_{T\to\infty} T^{-1} \operatorname{tr} E[(zI + \hat{\Psi}_{1,1})^{-1} \Psi_{1,1} (zI + \hat{\Psi}_{1,1})^{-1} \Psi_{1,2} \Psi_{1,2}'] \ \geq \ 0. \end{aligned} \tag{18}$$

The quantities in (18) account for covariances between observed and unobserved signals. While the existence of the limits in (18) cannot be guaranteed in general, the expectations are uniformly bounded for $z > 0$ (since so are the $\Psi$ matrices). Hence, by passing to a subsequence of $T, P$, we can always assume the limits in (18) exist. In the appendix, we show that these limits actually exist for a class of correlation structures.

With the additional assumptions for the mis-specified setting in place, we have the following analog of Propositions 2, 3, and 4.

**Proposition 5** *In the limit* $T, P, P_1 \to \infty,\ P/T \to c,\ P_1/P \to q \in (0, 1]$,

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

*in probability, where*

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

*and*

$$m(-z; cq; q) = \lim P_1^{-1} \operatorname{tr}((zI + \hat{\Psi}_{1,1})^{-1})\,.$$

*Furthermore,*

$$\begin{aligned}
\nu(z; cq; q) \ &= \psi_{*,1}(q) - (qc)^{-1} z\xi(z; cq; q) \ > \ 0 \\
\nu'(z; cq; q) \ &= \ -(qc)^{-1}(\xi(z; cq; q) + z\xi'(z; cq; q)) \ < \ 0 \\
\hat{\nu}(z; cq; q) \ &= \nu(z; cq; q) + z\nu'(z; cq; q) \ > \ 0\,.
\end{aligned}$$

*In addition, we have*

*i) The expected return on the market timing strategy converges in probability to*

$$\mathcal{E}(z;cq;q) \coloneqq \lim E[\hat{\pi}_t(z)R_{t+1}|\hat{\beta}] = b_* \, q \left( \nu(z;cq;q) \; + \frac{(cq)^{-1}\xi_{2,1}(z;cq;q)}{1+\xi(z;cq;q)} \right)$$

*ii) Expected leverage converges in probability to*

$$\mathcal{L}(z;cq;q) \coloneqq \lim E[\hat{\pi}_t(z)^2|\hat{\beta}] = q\Big(b_*\hat{\nu}(z;cq;q) - c(1+b_*[\psi_{*,1}(1) - q\psi_{*,1}(q)])\nu'(z;cq;q)\Big) + \Delta(z;cq;q)\,,$$

*where*

$$\Delta(z;cq;q) = b_* \frac{(qc)^{-1}\widehat{\xi}_{2,1}(z;cq;q) \; + \; 2(1+\xi(z;cq;q))\nu'(z;cq;q)\xi_{2,1}(z;cq;q)}{(1+\xi(z;cq;q))^2}\,.$$

*iii) $R^2$ converges in probability to*

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

*iv) The second moment of the market timing strategy converges in probability to*

$$\mathcal{V}(z;cq;q) \coloneqq \lim E\left[(\hat{\pi}_t(z)R_{t+1})^2\right] = 2(\mathcal{E}(z;cq;q))^2 + (1+b_*\psi_{*,1})\,\mathcal{L}(z;cq;q)\,.$$

*v) And, as a result, the Sharpe ratio satisfies*

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

In general, the behavior of quantities in Proposition 5 depends in a complex fashion on the correlations between observable and unobservable signals, as captured by the quantities (18). When both quantities (18) are zero, expressions significantly simplify. It is straightforward to show that both quantities in (18) are zero if the matrices $\Psi_{1,2}, \Psi_{2,1}$ have uniformly bounded traces. For example, this is when $\Psi_{1,2}$ has a finite, uniformly bounded rank when $P,\ P_1 \to \infty$

(due to, say, a finite-dimensional factor structure in the signals). We thus obtain the following result.

**Proposition 6** *Suppose that* $\operatorname{tr}(\Psi_{1,2}\Psi_{2,1}) = o(P)$.[31] *Then,* $\xi_{2,1} = \widehat{\xi}_{2,1} = 0$. *Furthermore,*

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

*(ii) both* $R^2(z;cq;q)$ *and* $SR(z;cq;q)$ *are monotone increasing in* $z$ *for* $z < z_* = c(1 + b_*(\psi_{*,1}(1) - q\psi_{*,1}(q)))/b_*$ *and monotone decreasing in* $z$ *for* $z > z_*$.

*(iii) in the ridgeless limit as* $z \to 0$, *we have*

$$\mathcal{E}(0;cq;q) = b_* q \left(\psi_{*,1}(q) - (cq)^{-2} m_*(cq;q)^{-1} \mathbf{1}_{q>1/c}\right)$$

$$\mathcal{L}(0;cq;q) = \mathcal{E}(0;cq;q) + \left(1 + b_*(\psi_{*,1}(1) - q\psi_{*,1}(q))\right) \begin{cases} ((cq)^{-1} - 1)^{-1}, & q < 1/c \\ \tilde{\mu}(cq;q), & q > 1/c \end{cases}$$

$$\mathcal{V}(0;cq;q) = 2(\mathcal{E}(0;cq;q))^2 + (1 + b_*\psi_{*,1})\,\mathcal{L}(0;cq;q)$$

$$SR(0;cq;q) = \frac{\mathcal{E}(0;cq;q)}{\sqrt{\mathcal{V}(0;cq;q)}}$$

*for some* $m_*(cq;q) > 0$ *and some* $\tilde{\mu}(cq;q) < 0$ *with* $\tilde{\mu}(1+;c) = -\infty$. *In particular, if* $\Psi$ *is proportional to the identity matrix,* $\Psi = \psi_{*,1} I$, *then*

$$\mathcal{E}(0;cq;q) = b_* \psi_{*,1} \min\{q, c^{-1}\} \tag{20}$$

*is constant for* $q > 1/c$.

The comparative statics of Section 4.2 highlight how, even when the empirical model is correctly specified, complexity hinders the model's ability to hone in on the true DGP

[31] This is the case, for example, when $\Psi_P = D_P + Q_P$ where $\limsup_{P\to\infty} \operatorname{rank} Q_P < \infty$, while $D_P$ are diagonal matrices, and $D_P$, $Q_P$ are uniformly bounded. In this case, we can replace $\Psi_P$ with $D_P$ in all expressions. Perhaps more tangibly, this condition obtains when the signals satisfy a finite-dimensional factor structure. Furthermore, if the signals have similar idiosyncratic variance, they satisfy the necessary mixing condition.

because there is not enough data to support the model's heavy parameterization. That analysis shows that when models are correctly specified, the best performance (in terms of $R^2$ and Sharpe ratio) comes from simple models. Naturally, a small, correctly specified model will converge on the truth faster than a large, correctly specified model. But this is not a very helpful comparison.

The fundamental difference in this section is that while raising $cq$ brings the usual statistical challenges of heavy parameterization without much data, the added complexity also brings the benefit of improving the empirical model's approximation of the true DGP. A simple model will tend to suffer from poor approximation and thus fare poorly in terms of both statistical metrics like $R^2$ and portfolio metrics like expected return and Sharpe ratio. Thus, our mis-specification analysis tackles the most important question about high-complexity: Does the improvement in approximation justify the statistical cost of heavy parameterization when it comes to out-of-sample forecast and portfolio performance? The answer is yes, as established by the following theorem.

**Theorem 7 (Virtue of Complexity)** *Suppose that signals are sufficiently mixed (so that $H(x;q)$ does not depend on $q$) and $\text{tr}(\Psi_{1,2}\Psi_{2,1}) = o(P)$. Then, with the optimal amount of shrinkage $z_*$, the Sharpe ratio $SR(z_*(q;c);cq;q)$ and $R^2(z_*(q;c);cq;q)$ are strictly monotone increasing and concave in $q \in [0,1]$.*

Figures 4, 5, and 6 illustrate the behavior of mis-specified machine learning predictions and portfolios derived in Proposition 5. In this calibration, the true unknown DGP is assumed to have a complexity of $c = 10$. We continue to calibrate $\Psi$ as identity and $b_* = 0.2$. We analyze the behavior of approximating empirical models that range in complexity from very simple ($cq \approx 0$ and thus severely mis-specified) to highly complex ($q = 1,\ cq = 10$ and thus correctly specified). The left panel of Figure 4 shows the expected out-of-sample $R^2$. The cost of mis-specification for low $c$ is seen as a shift downward in the $R^2$ relative to Figure 1. The challenges of model complexity highlighted in previous sections play an important role here as well. Intermediate levels of complexity ($c \approx 1$) dilate the size of beta estimates

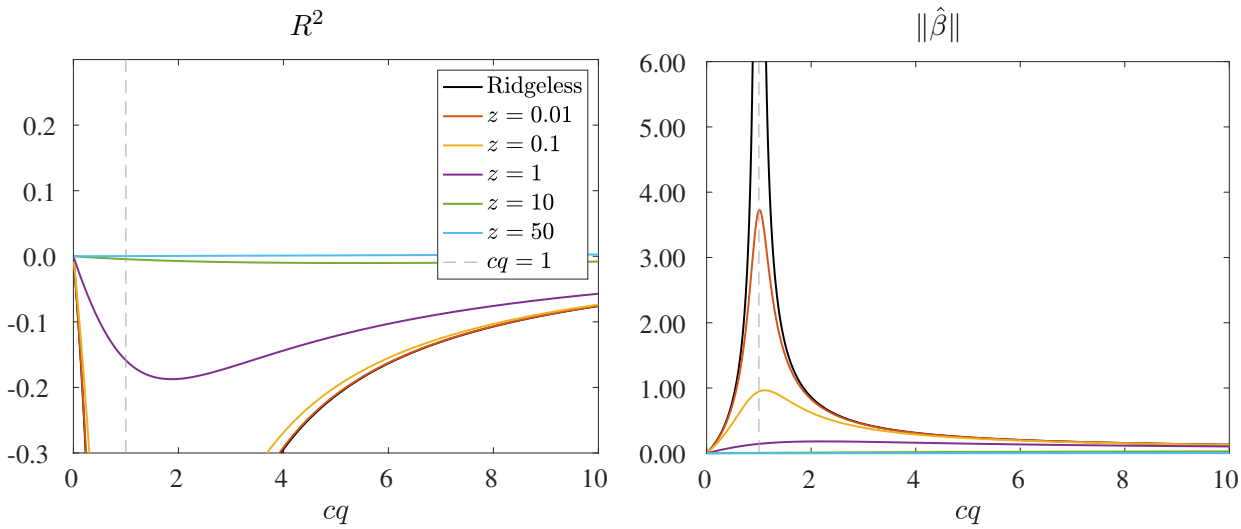



**Figure 4:** Expected Out-of-sample Prediction Accuracy From Mis-specified Models

**Note.** Limiting out-of-sample $R^2$ and $\hat{\beta}$ norm as a function of $c$ and $z$ from Proposition 6 assuming $\Psi$ is the identity matrix, $b_* = 0.2$, and the complexity of the true model is $c = 10$.

(Figure 4, right panel), driving down the $R^2$ and inflating portfolio volatility (Figure 5, right panel). These effects abate once again for $c > 1$ due to the implicit regularization of high-complexity ridgeless regression, just as in the earlier analysis. More generally, the patterns for $R^2$, $\hat{\beta}$ norm, and portfolio volatility share similar qualitative patterns to those in Figure 1.

The most important difference versus Figure 1 is the pattern for the out-of-sample expected return of the market timing strategy (Figure 5, right panel). Expected returns are now low for simple strategies due to their poor approximation of the DGP. Increasing model complexity monotonically increases expected timing returns. In the ridgeless case, the benefit of added complexity reaches its maximum of $\mathcal{E}(0; 1; c^{-1}) = b_* \psi_{*,1} c^{-1}$ when $cq = 1$. A surprising fact is that the ridgeless expected return is exactly flat as complexity rises beyond $cq = 1$, in which case the benefits of incremental improvements in DGP approximation are exactly offset by the gradually rising bias of ridgeless shrinkage; see formula (20).

This new fact that the expected return rises monotonically with model complexity in the mis-specified setting induces a similar pattern in the out-of-sample Sharpe ratio, shown in Figure 6. Rather than decreasing in complexity as we saw in the correctly specified setting,

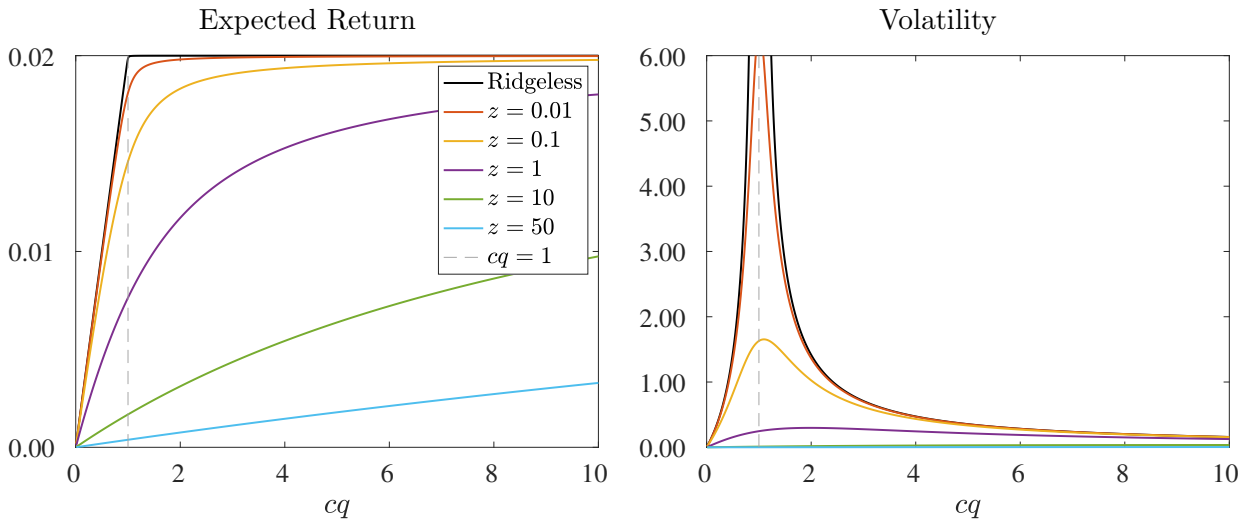



**Figure 5:** Expected Out-of-sample Risk and Return From Mis-specified Models

**Note.** Limiting out-of-sample expected return and volatility of the market timing strategy as a function of $c$ and $z$ from Proposition 6 assuming $\Psi$ is the identity matrix, $b_* = 0.2$, and the complexity of the true model is $c = 10$.

the expected return improvement from additional complexity leads the Sharpe ratio to also increase with complexity. Consistent with Theorem 7, this is particularly true with non-trivial ridge shrinkage but is even true in the ridgeless case as long as $cq$ is sufficiently far from unity. In summary, in the realistic case of mis-specified empirical models, complexity is a virtue. It improves the expected out-of-sample market timing performance in terms of both expected return and Sharpe ratio.

It is instructive to compare our findings with the phenomenon of double descent, which is that, absent regularization, out-of-sample $MSE$ has a non-monotonic pattern in model complexity (Belkin et al., 2019b; Hastie et al., 2019). The mirror image of double descent in $MSE$ is the "double ascent" behavior of the ridgeless Sharpe ratio (Figure 6). As Theorem 7 shows, Sharpe ratio double ascent is an artifact of insufficient shrinkage. With the right amount of shrinkage, complexity becomes a virtue even in the low complexity regime (when $cq < 1$): the hump disappears, and "double ascent" turns into "permanent ascent."

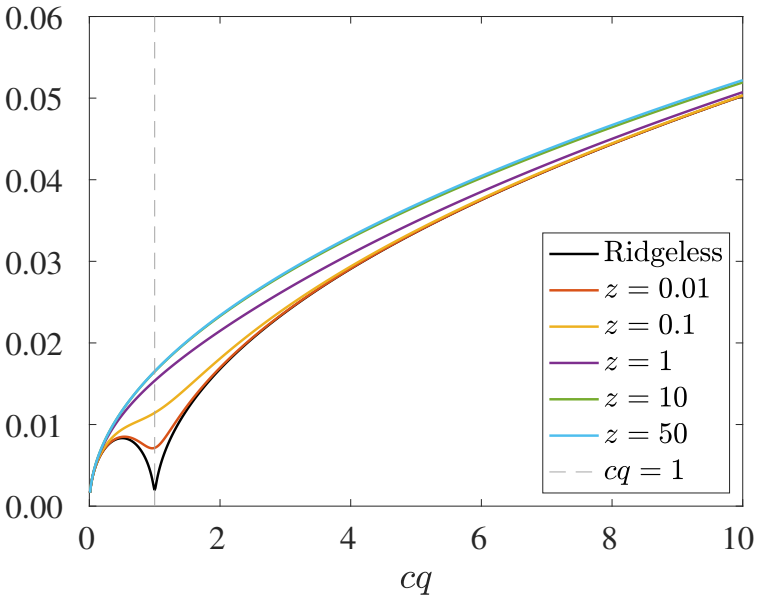



**Figure 6:** Expected Out-of-sample Sharpe Ratio From Mis-specified Models

**Note.** Limiting out-of-sample Sharpe ratio of the market timing strategy as a function of $c$ and $z$ from Proposition 6 assuming $\Psi$ is the identity matrix, $b_* = 0.2$, and the complexity of the true model is $c = 10$.

# 6 Virtue of Complexity: Empirical Evidence From Market Timing

In this section, we present empirical analyses that are direct empirical analogs to the theoretical comparative statics for mis-specified models in Section 5.

## 6.1 Data

Our empirical investigation centers on a cornerstone of empirical asset pricing research—forecasting the aggregate stock market return. To make the conclusions from this analysis as easy to digest as possible, we perform our analysis in a conventional setting with conventional data. Our forecast target is the monthly return of the CRSP value-weighted index. The information set we use for prediction consists of the 15 predictor variables from Goyal and Welch (2008) available monthly over the sample 1926–2020.[32]

We volatility standardize returns and predictors using backward-looking standard deviations that preserve the out-of-sample nature of our forecasts. Returns are standardized

[32]This list includes (using mnemonics from their paper): dfy, infl, svar, de, lty, tms, tbl, dfr, dp, dy, ltr, ep, b/m, and ntis, as well as one lag of the market return.

by their trailing 12-month return standard deviation (to capture their comparatively fast-moving conditional volatility). In contrast, predictors are standardized using an expanding window historical standard deviation (given the much higher persistence of most predictors). We require 36 months of data to ensure enough stability in our initial predictor standardization, so the final sample we bring to our analysis began in 1930. We perform this standardization to align the empirical analysis with our homoskedastic theoretical setting. Still, our results are insensitive to this step (none of our findings are sensitive to variations in how standardizations are implemented).

### 6.2 Random Fourier Features

We seek models taking the form of equation (3). In order to evaluate our theory, we also seek a framework that will allow us to smoothly transition from low complexity models to high-complexity. To do so, we adopt an influential methodology from the machine learning literature known as random Fourier features, or RFF (Rahimi and Recht, 2007, 2008).[33] Let $G_t$ denote our $15 \times 1$ vector of predictors. The RFF methodology converts $G_t$ into a pair of new signals

$$S_{i,t} = P^{-\frac{1}{2}} \left[\sin(\gamma \omega_i' G_t), \cos(\gamma \omega_i' G_t)\right]', \quad \omega_i \sim i.i.d.N(0, I). \tag{21}$$

$S_{i,t}$ uses the vector $\omega_i$ to form a random linear combination of $G_t$, which is then fed through the trigonometric functions.[34] The advantage of RFF is that for a fixed set of input data, $G_t$, we can create an arbitrarily large (or small) set of features based on the information in $G_t$

[33] Rahimi and Recht (2007) describe how RFF approximation accuracy improves as you increase the level of model complexity. In the limit of zero complexity ($P, T \to \infty$, $P/T \to 0$), RFF regression approximates any sufficiently smooth non-linear function arbitrarily well. Subsequent papers (see for example Rudi and Rosasco, 2017) further characterize rates of convergence. The case of non-zero complexity is less well understood. Recent results (Mei and Montanari, 2019; Mei et al., 2022; Ghorbani et al., 2020) show that, for non-zero complexity, random features methods cannot learn the true function and only learn its projection on a specific functional sub-space.

[34] The parameter $\gamma$ controls the Gaussian kernel bandwidth in the generation of random Fourier features. Random features can be generated in several ways (for a survey see Liu et al., 2021). Our choice of functional form in (21) is guided by Sutherland and Schneider (2015) who document tighter error bounds for this functional approximation relative to some alternative random feature formulations. However, we have found that our results are insensitive to using other random feature schemes.

through the nonlinear transformation in (21). If one desires a very low-dimensional model in (3), say $P = 2$, one can generate a single pair of RFFs. For a very high-dimensional model, say $P$ = 10,000, one can instead draw many random weight vectors $\omega_i$, $i = 1, ..., 5{,}000$. The larger the number of random features, the richer the approximation (3) provides to the general functional form $E[R_{t+1}|G_t] = f(G_t)$ where $f$ is some smooth nonlinear function. Indeed, the RFF approach is a wide two-layer neural network with fixed weights in the first layer (in the form of $\omega_i$) and optimized weights in the second layer (in the form of the regression estimates for $\beta$).

### 6.3 Out-of-sample Performance

To conduct the empirical analogue of the theoretical analysis in Figures 4, 5, and 6, we consider a one-year, five-year and ten-year year rolling training windows ($T$ = 12, 60, or 120) and a large set of RFFs (as high as $P$ = 12,000). These choices are guided by our desire to investigate the role of model complexity, defined in the empirical analysis as $c = P/T$. The advantages of short training samples like $T = 12$ are i) that we can reach extreme levels of model complexity with smaller $P$ and thus less computing burden, and ii) it shows that the virtue of complexity can be enjoyed in small samples. But none of our conclusions are sensitive to this choice as we document all of the same patterns for training windows of $T = 60$ and 120.

To draw "VoC curves" along the lines of Figures 4, 5, and 6, we estimate a sequence of out-of-sample predictions and trading strategies for various degrees of model complexity ranging from $P = 2$ to $P$ = 12,000 and varying degrees of ridge shrinkage ranging from $\log_{10}(z) = -3, ..., 3$. One repetition of our analysis proceeds as follows:

(i) Generate 12,000 RFFs according to (21) with bandwidth parameter $\gamma$.[35]

(ii) Fix a model defined by the number of features $P \in \{2, ..., 12{,}000\}$ and a ridge shrinkage

[35] We set $\gamma = 2$. Our results are generally insensitive to $\gamma$, as discussed in the robustness section below.

parameter $\log_{10}(z) \in \{-3, ..., 3\}$. The set of predictors $S_t$ for regression (3) correspond to the first $P$ RFFs from (i).

(iii) Given the model in (ii), and fixing a training window $T \in \{12, 60, 120\}$, conduct a recursive out-of-sample prediction and market timing strategy. For each $t \in \{T, ..., 1{,}091\}$, estimate (3) using training observations $\{(R_t, S_{t-1}), ..., (R_{t-T+1}, S_{t-T})\}$.[36] Then, from the estimated regression coefficient, construct out-of-sample return forecast $\hat{\beta}' S_t$ and timing strategy return $\hat{\beta}' S_t R_{t+1}$.

(iv) From the sequence of out-of-sample predictions and strategy returns in (iii), calculate the average $\|\hat{\beta}\|^2$ across training samples, the out-of-sample $R^2$, and the out-of-sample average return, volatility, and Sharpe ratio of the timing strategy.

The inherent randomness of RFFs means that estimates of out-of-sample performance tend to be noisy for models with low $P$. Therefore we repeat the analysis (i)–(iv) 1,000 times with independent draws of the RFFs, and then average the performance statistics across repetitions.

The VoC curves in Figures 7 and 8 plot out-of-sample prediction and market timing performance as a function of model complexity and ridge shrinkage for the case $T = 12$. The wide range of complexity that we consider (e.g., $c \in [0, 1000]$ when $T = 12$) can make it difficult to read plots. To better visualize the results while emphasizing both behaviors near the interpolation boundary and behavior for extreme complexity, we break the $x$-axis at an intermediate value of $c$.

The first conclusion from these figures is that the out-of-sample empirical behavior of machine learning predictions is a strikingly close match to the VoC curves predicted by our theory. In particular, compare the empirical results of Figure 7 to the theoretical results under model mis-specification from Figure 4. The beta estimates and out-of-sample $R^2$ demonstrate explosiveness at the interpolation boundary and recovery in the high-complexity

[36] Prior to estimation, we volatility standardize the training sample RFFs $\{S_{t-1}, ..., S_{t-T}\}$ and out-of-sample RFFs $S_t$ by their standard deviations in the training sample.

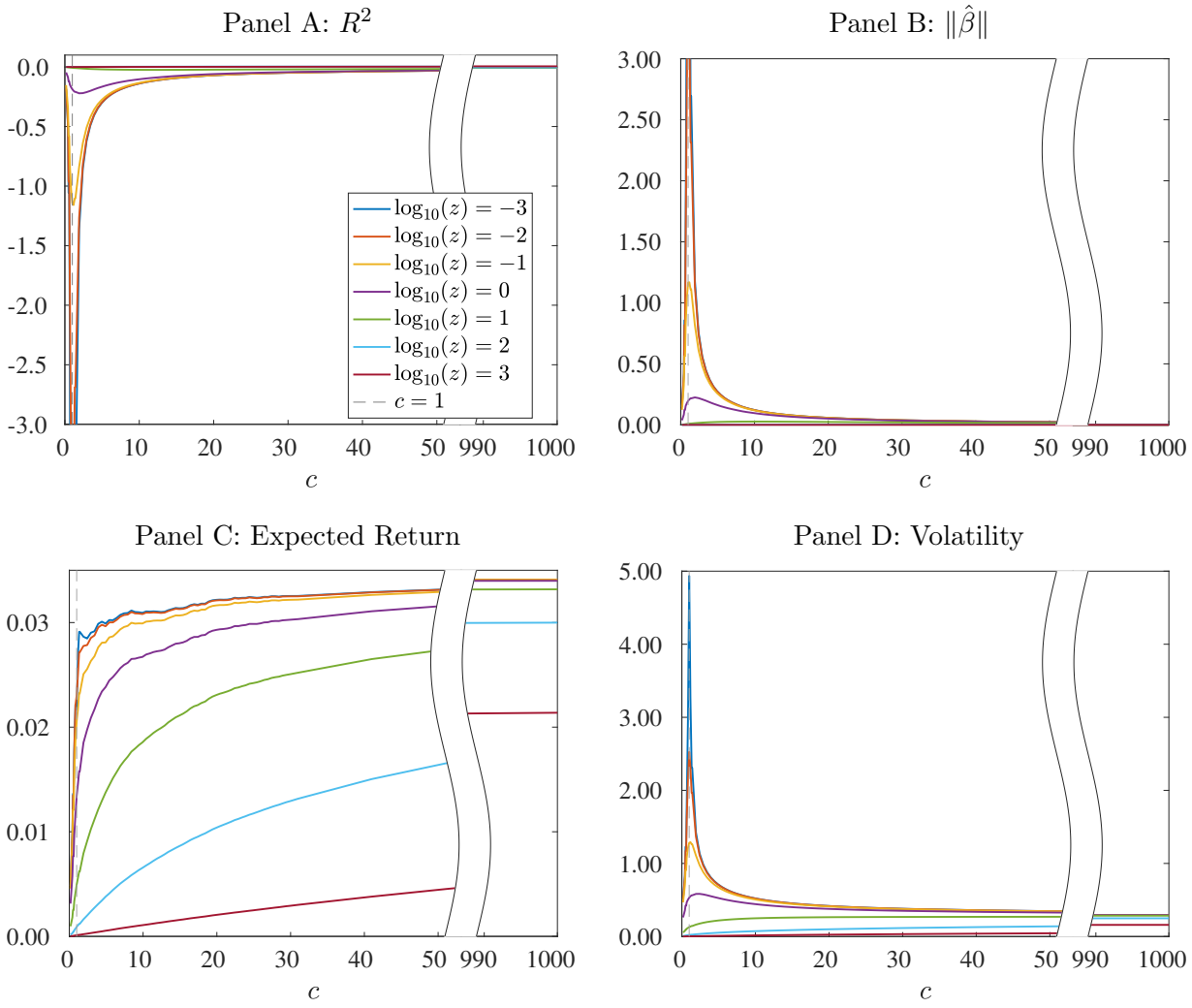



**Figure 7:** Out-of-sample Market Timing Performance ($T = 12$)

**Note.** Out-of-sample prediction accuracy and portfolio performance estimates for empirical analysis described in Section 6.3. Training window is $T = 12$ months and RFF count $P$ (or $cT$) ranges from 2 to 12,000 with $\gamma = 2$.

regime. Figures 12 and 13 (reported in the appendix in the interest of space) document identical patterns for training windows of 60 and 120 months.

Extreme behavior at the interpolation boundary makes it difficult to fully appreciate the patterns in $R^2$. Figure 14 in the appendix provides more detail by plotting the out-of-sample $R^2$ zooming-in on the range $[-10\%, 1\%]$. Here we see more clearly that high complexity and regularization combine to produce a positive out-of-sample $R^2$. In this plot, regularization comes in two forms, directly through higher $z$ and more subtly through higher $c$ (which

allows ridgeless regression to find solutions with small $\hat{\beta}$ norm). For large $z$, the $R^2$ is almost everywhere positive for all training windows.

The most intriguing aspect of Figure 7 is the clear increasing pattern in out-of-sample expected returns as model complexity rises. For $z = 10^{-3}$, which roughly approximates the ridgeless case, we see a nearly linear upward trend in average returns as $c$ rises from 0 to 1. Beyond $c = 1$, the ridgeless expected return is nearly flat, just as predicted by equation (20) in Proposition 6. For higher levels of ridge shrinkage, the rise in expected return is more gradual and continues into the range of extreme model complexity. Appendix Figures 12 and 13 again document an identical expected return pattern for longer training windows.

The increasing pattern in out-of-sample expected return and the decreasing pattern in volatility above $c = 1$ translate into a generally increasing pattern in the out-of-sample market timing Sharpe ratio, shown in Figure 8. The exception is a brief dip near $c = 1$ at low levels of regularization as the spike in variance compresses the Sharpe ratio. For high complexity, the Sharpe ratio generally exceeds 0.4.

In our theoretical setting, we normalize the expected return of the un-timed asset to zero. This is, of course, not the case for the US market return. Therefore, to adjust for buy-and-hold market exposure, we calculate the out-of-sample alpha, alpha $t$-statistic, and information ratio (IR) of the timing strategy return via time series regression on the un-timed market. Figure 8 shows that the market timing alpha and IR inherit the same patterns as the average return and Sharpe ratio. In the high-complexity regime, we find information ratios around 0.3 and significant alpha $t$-statistics ranging from 2.6 to 2.9 depending on the amount of ridge shrinkage. Figure 9 repeats this analysis for training windows of 60 and 120 months, where we find similar information ratios of roughly 0.25 with $t$-statistics above 2.0 for high-complexity models.

What do market timing strategies look like in the high-complexity regime? Figure 10 plots $\hat{\pi}(z, c)$ for the highest complexity and shrinkage configurations of our empirical model ($P = 12,000$ and $z = 10^3$, averaged across 1,000 sets of random feature weights). The three lines correspond to training windows of 12, 60, and 120 months. Positions show the

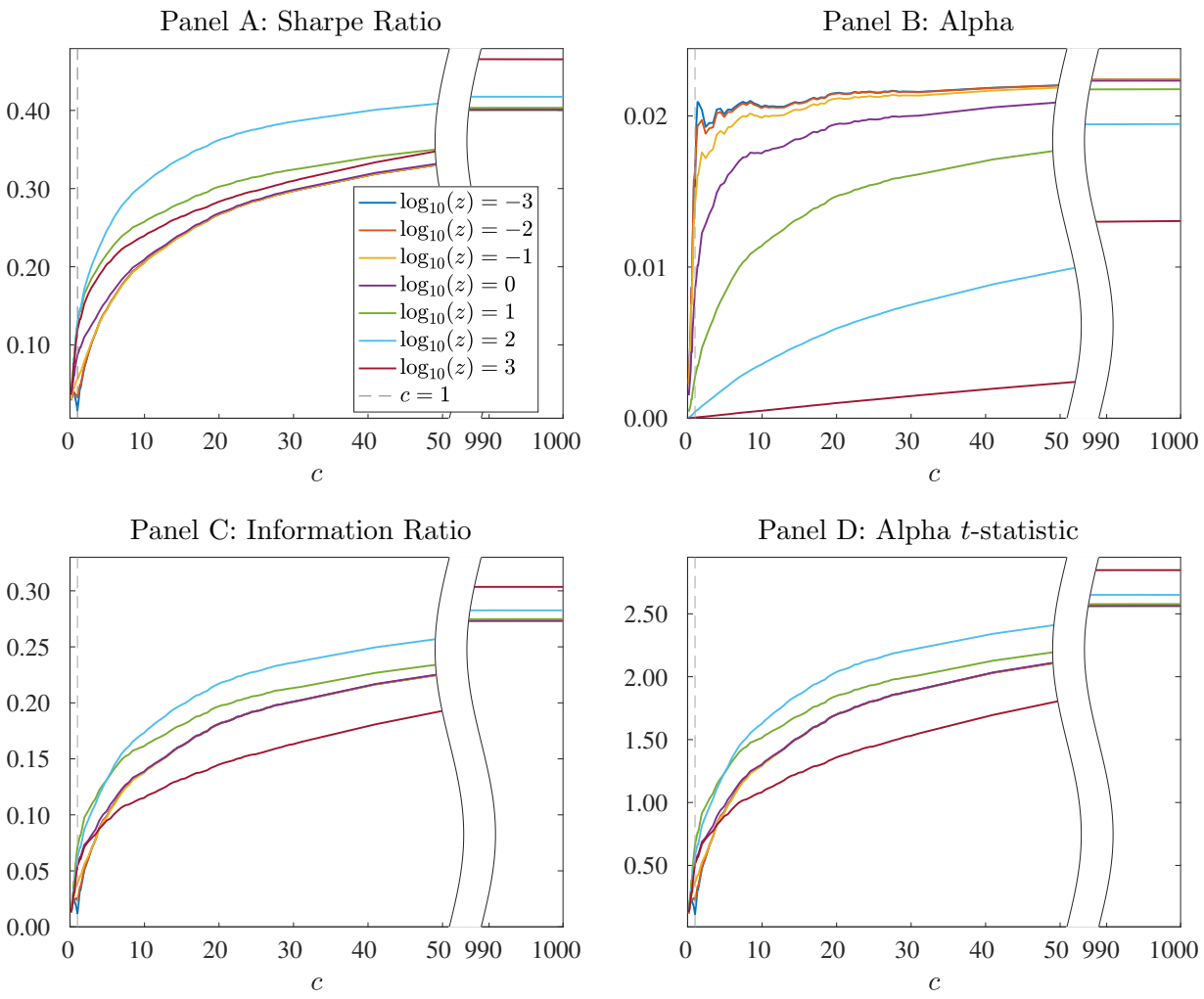



**Figure 8:** Out-of-sample Market Timing Performance ($T = 12$)

**Note.** Out-of-sample prediction accuracy and portfolio performance estimates for empirical analysis described in Section 6.3. Training window is $T = 12$ months and RFF count $P$ (or $cT$) ranges from 2 to 12,000 with $\gamma = 2$. Alphas are versus a static position in the volatility-standardized market portfolio.

same patterns for all training windows; their time series correlations are 90% ($T = 12$ with $T = 60$), 87% ($T = 12$ with $T = 120$), and 97% ($T = 60$ with $T = 120$).[37] The plot shows 6-month moving averages of raw positions for better readability (our trading results are based on the raw positions and not the moving averages).

The timing positions in Figure 10 are remarkable. First, they show that the high-complexity strategy is long-only at heart. Negative bets are infrequent and small relative

[37] While the time series patterns in positions are the same for all training windows, the scale of positions is smaller for longer training windows. This is because the "leverage" of a strategy is driven by the norm of beta, and this is typically smaller for larger $T$.

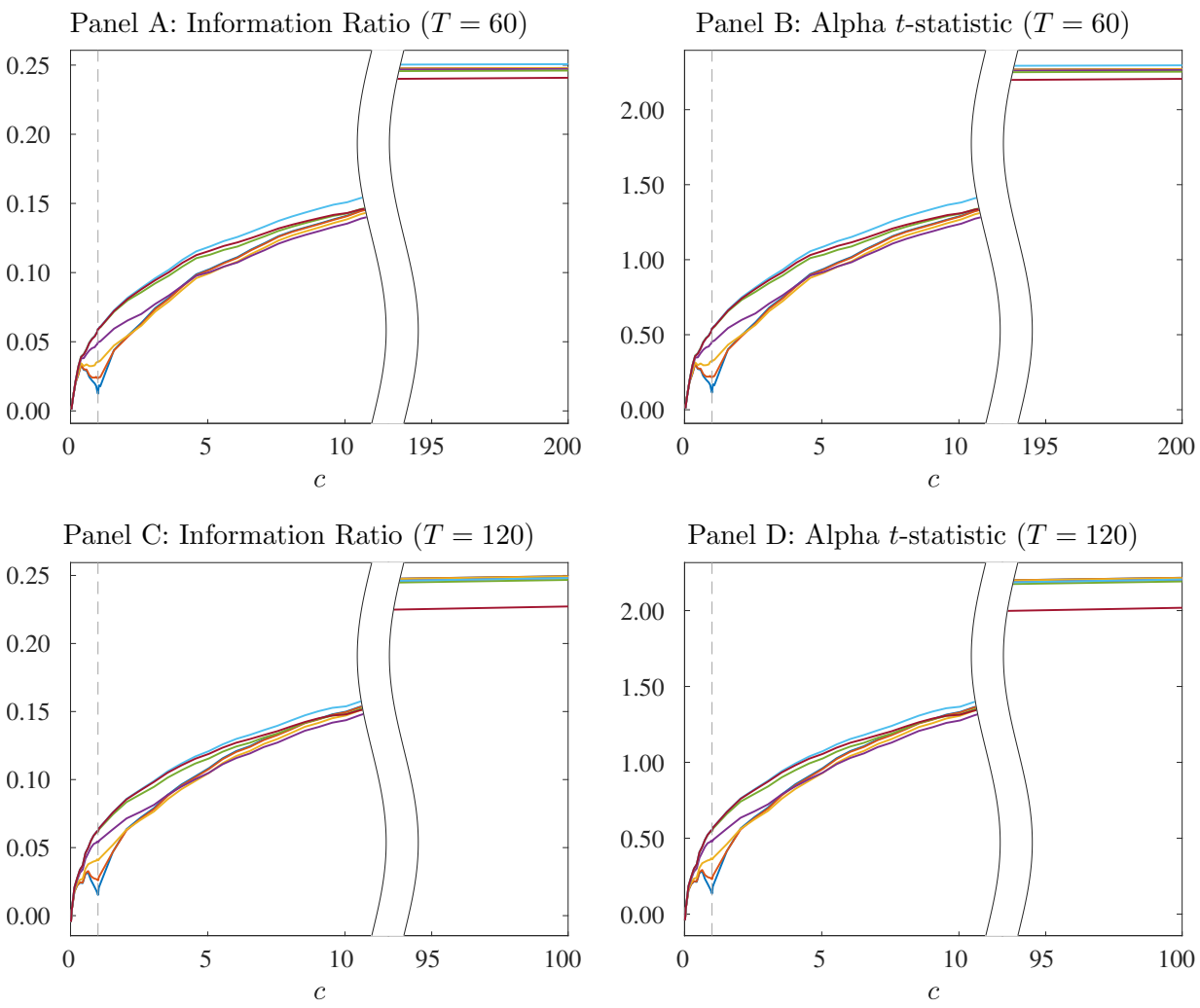



**Figure 9:** Out-of-sample Market Timing Performance ($T = 60, 120$)

**Note.** Out-of-sample prediction accuracy and portfolio performance estimates for the empirical analysis described in Section 6.3. Training window is $T = 60$ or 120 months and RFF count $P$ (or $cT$) ranges from 2 to 12,000 with $\gamma = 2$. Alphas are versus a static position in the volatility-standardized market portfolio.

to positive bets. The machine learning model thus heeds the guidance of Campbell and Thompson (2008) "that many predictive regressions beat the historical average return, once weak restrictions are imposed on the signs of coefficients and return forecasts." However, unlike Campbell and Thompson (2008), the machine seems to learn this rule without being given an explicit constraint.[38]

Second, the machine learning strategy learns to divest leading up to recessions. NBER

[38]Strictly imposing the Campbell and Thompson (2008) constraint gives a boost in Sharpe ratio from 0.47 to 0.54 in the $T = 12$ case; from 0.42 to 0.50 for $T = 60$; and from 0.41 to 0.49 for $T = 120$.

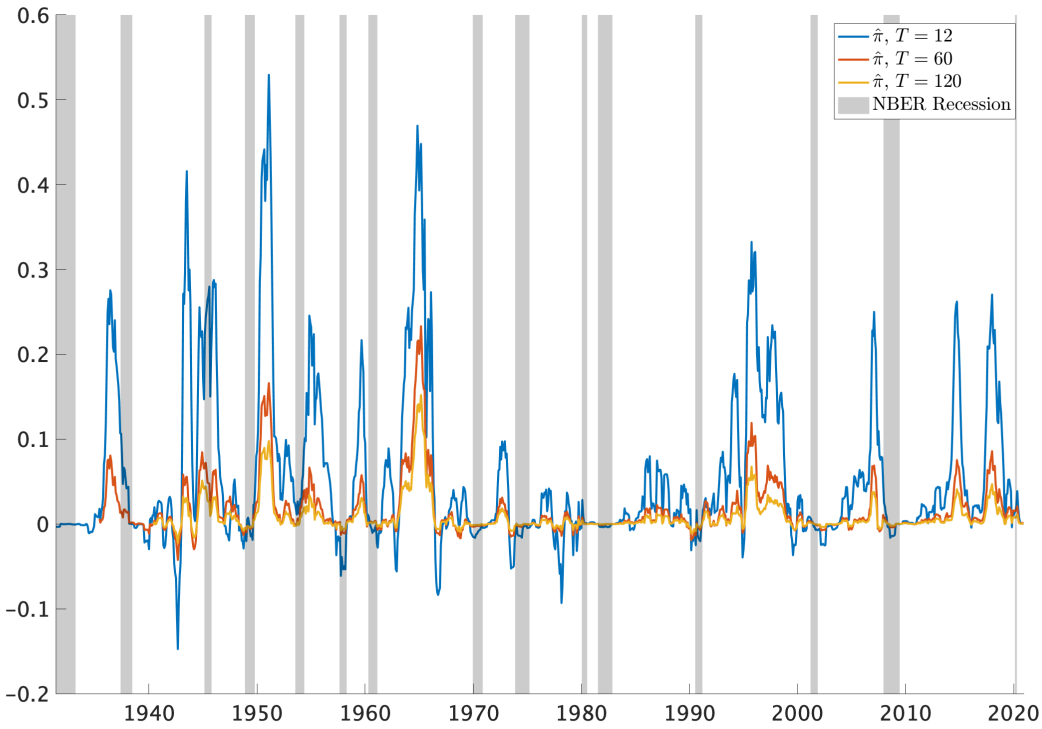



**Figure 10:** Market Timing Positions

**Note.** Out-of-sample prediction accuracy and portfolio performance estimates for empirical analysis described in Section 6.3. Training window is $T = 12$, 60, or 120 months with $P = 12,000$, $z = 10^3$, and $\gamma = 2$. Positions are averaged across 1,000 sets of random feature weights. Plots show the 6-month moving average of positions to improve readability.

recession dates are shown in the gray-shaded regions. For 14 out of 15 recessions in our test sample, the timing strategy substantially reduces its position in the market before the recession (the exception is the eight-month recession of 1945). And it does this on a purely out-of-sample basis.

### 6.4 Comparison With Goyal and Welch (2008)

Our results seem at odds with the primary conclusion of Goyal and Welch (2008). They argue that the enterprise of market return prediction, which has occupied large attention in the asset pricing literature for decades, is by and large a failed endeavor: "these models seem unstable, as diagnosed by their out-of-sample predictions and other statistics; and

these models would not have helped an investor with access only to available information to profitably time the market." But we use the same predictive information studied in that paper. What is the source of the discrepancy?

The conclusions of Goyal and Welch (2008) are based on their findings of consistently negative out-of-sample prediction $R^2$. They do not analyze the performance of timing strategies based on expected returns or Sharpe ratios.[39] We revisit their analysis with a focus on timing strategy performance using the same recursive out-of-sample prediction scheme as in the analysis of Figures 7 and 8. We use rolling 12, 60, and 120-month training windows (Panels A, B, and C, respectively). We focus on a version of what Goyal and Welch (2008) call the "kitchen sink" regression. Our implementation uses 15 monthly predictors in a linear ridgeless regression.[40]

The first finding of Table 1 is that we confirm the conclusions of Goyal and Welch (2008). Note that, with monthly data, a model with 15 regressors already has nontrivial complexity even for long training windows, and for the 12-month training window, its complexity even exceeds one. Monthly return forecasts using linear ridgeless regression behave egregiously. The monthly out-of-sample $R^2$ from ridgeless regression ($z = 0^+$) is large and negative at less than $-100\%$ ($-9764\%$ to be precise!). The timing strategy based on these predictions is also poor. The Sharpe ratio is $-0.11$ and is insignificantly different from zero. This seems perhaps not so terrible given the wildness of the forecasts, but it is due to the fact that the strategy's volatility is so high. Its maximum loss is 98 standard deviations. In light of our theoretical analysis, this agreement with the conclusions of Goyal and Welch (2008) is perhaps unsurprising. With $P = 15$ and $T = 12$, this analysis takes place near the interpolation boundary. Thus, forecasts and timing strategy returns are expected to be highly volatile, as our estimates confirm. In Panels B and C, we repeat the same analysis with

[39] Updating the original Goyal and Welch (2008) analysis, Goyal et al. (2021) provide some evidence of timing strategy performance for market return predictors.

[40] To remain consistent with our other analyses, the forecast target is the monthly market return standardized by its rolling 12-month volatility standardization. We continue to refer to this as "the market" throughout. As discussed in the robustness section, our results across the board are generally insensitive to, and our conclusions entirely unaffected by, whether we work with the raw or volatility standardized market return.

**Table 1:** Comparison With Goyal and Welch (2008)

**Note.** Out-of-sample prediction accuracy and portfolio performance estimates for high-complexity timing strategy returns with $c = 1000$ and $z = 10^3$ in Section 6.3 ("Nonlinear") averaged across 1,000 sets of random feature weights, compared with the linear kitchen sink model of Goyal and Welch (2008) ("Linear") with shrinkage of $z = 0^+$ (ridgeless) and $z = 10^3$. The forecast target is the monthly market return standardized by its rolling 12-month volatility standardization. We report strategy Sharpe ratios (with average return $t$-statistics), information ratios versus the market and versus the linear model with $z = 10^3$ (with alpha $t$-statistics). The panels correspond to training windows of 12, 60, or 120 months. "Max Loss" is in standard deviation units.

| Model | Shrinkage | $R^2$ | SR | $t$ | IR v. Mkt | $t$ | IR v. Linear | $t$ | Max Loss | Skew |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Panel A: 12-month Training Window | | | | | | |
| Linear | $z = 0^+$ | <-100% | -0.11 | -1.0 | -0.16 | -1.6 | - | - | 98.5 | -0.9 |
| | $z = 10^3$ | -3.8% | 0.46 | 4.4 | 0.33 | 3.1 | - | - | 2.4 | -0.1 |
| Nonlinear | $z = 10^3$ | 0.6% | 0.47 | 4.5 | 0.31 | 2.9 | 0.26 | 2.5 | 1.2 | 2.5 |
| | | | | Panel B: 60-month Training Window | | | | | | |
| Linear | $z = 0^+$ | -96.6% | 0.00 | 0.0 | -0.07 | -0.6 | - | - | 35.8 | -11.1 |
| | $z = 10^3$ | -0.5% | 0.44 | 4.1 | 0.10 | 0.9 | - | - | 1.4 | -0.3 |
| Nonlinear | $z = 10^3$ | 0.5% | 0.42 | 3.9 | 0.25 | 2.3 | 0.27 | 2.5 | 0.5 | 1.7 |
| | | | | Panel C: 120-month Training Window | | | | | | |
| Linear | $z = 0^+$ | -26.6% | 0.20 | 1.8 | 0.14 | 1.2 | - | - | 15.4 | -6.5 |
| | $z = 10^3$ | 0.1% | 0.49 | 4.4 | 0.13 | 1.2 | - | - | 0.8 | -0.9 |
| Nonlinear | $z = 10^3$ | 0.3% | 0.41 | 3.7 | 0.24 | 2.2 | 0.24 | 2.2 | 0.3 | 0.9 |

longer training windows ($T = 60$ and 120). Longer training windows lead to less variable ridgeless regression estimates, producing higher (though still negative) $R^2$ and improving the Sharpe ratio.

Our theoretical analysis suggests that, in circumstances like the linear kitchen sink where the regression takes place near the interpolation boundary, the benefits from additional ridge shrinkage are potentially large. Therefore, we re-estimate the Goyal and Welch (2008) kitchen sink regression with the same range of ridge parameters used in our machine learning models. The $R^2$ from even heavily regularized regressions can remain negative, as seen in the out-

of-sample $R^2$ of $-3.8\%$ when $z = 10^3$. However, with this much shrinkage, the benefits of market timing become large. The annualized out-of-sample Sharpe ratio of the strategy is 0.46 with a $t$-statistic of 4.4. This performance is not due to static market exposure. In the column "IR v. Mkt," we report performance after regressing on the volatility-standardized market return. The linear model with $z = 10^3$ has an information ratio of 0.33 ($t = 3.1$) versus the market. Shrinkage also produces more attractive maximum loss and skewness. These patterns align with the behavior predicted by our theoretical analysis. Near the interpolation boundary, models can seem defective in terms of $R^2$, yet they can nonetheless confer large economic benefits for investors. In Panels B and C, we see that shrinkage also benefits performance amid longer training windows. For $T = 120$, the linear strategy Sharpe ratio is 0.49 for $z = 10^3$ (the alpha versus the market is insignificant, however).

The "Nonlinear" model in Table 1 refers to the machine learning timing strategy with $c = 1000$ and $z = 10^3$ (averaged across 1,000 sets of random weight draws). In Panel A, the out-of-sample $R^2$ is 1% per month, with a Sharpe ratio of 0.46 with an information ratio of 0.31 versus the market. It also has a significant information ratio of 0.26 ($t = 2.5$) versus the best linear strategy ($z = 10^3$). One of the most attractive aspects of the machine learning strategy is its low downside risk. Its worst month was a loss of 1.23 standard deviations, and its skewness is positive, 2.48. These attractive tail risk properties of the machine learning model are not reflected in the Sharpe ratio. Still, they would be an important utility boost for investors who care about non-Gaussian risks. Note that the machine learning strategy accomplishes this using the identical information set as the linear strategy; it exploits this information in a high-dimensional, nonlinear way. Using longer training windows (Panels B and C) lead to the same conclusions.

### 6.5 Variable Importance

These results beg the question: How can such large models learn predictive patterns in training windows as short as 12 months, particularly when several raw predictors are highly persistent (e.g., dividend yield and T-bill rate)? The short answer is that a number of the

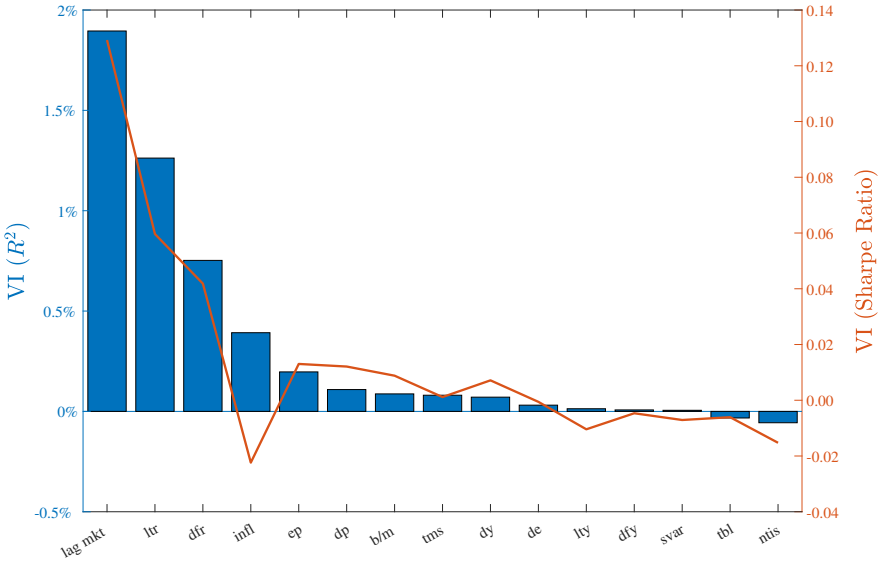



**Figure 11:** Variable Importance

**Note.** Variable importance (VI) for the $i^{th}$ predictor is the change in performance, defined as out-of-sample $R^2$ or Sharpe ratio, moving from the full model with 15 variables to the re-estimated model using 14 variables (excluding variable $i$).

15 raw predictors are, in fact, highly variable over short horizons, and these variables are the most important contributors to the performance of the high-complexity model. To shed more detailed light on this, we analyze the contribution of each variable to overall model performance. We re-estimate the machine learning model omitting each of the 15 predictor variables one by one. We calculate "variable importance" (VI) for the $i^{th}$ predictor as the change in performance (defined as out-of-sample $R^2$ or Sharpe ratio) moving from the full model with 15 variables to the re-estimated model using 14 variables (excluding variable $i$).

Figure 11 plots the results for the 12-month training window (with $P = 12,000$, $z = 10^3$, and averaged across 1,000 sets of random feature weights). The three most important variables are also the three predictors with the highest average variation in 12-month windows (i.e., the least persistent predictors).[41] Excluding the lagged market return ("lag mkt"), long-term bond return ("ltr"), or default return ("dfr") from the random features model reduces the out-of-sample monthly prediction $R^2$ by 1.9%, 1.3%, and 0.8%, respectively. In other words, the complex model is particularly adept at leveraging information in short-horizon

[41] Figure 15 in the appendix reports the average variation of each predictor in 12-month training windows

fluctuations among predictors. The variable importance calculations tell the same story when we measure it in terms of $R^2$ (bars) or Sharpe ratio (line).

Variable importance helps us identify which of the 15 predictors are the most dominant information sources. But our results further show that the key differentiator of the high-complexity model is its ability to extract nonlinear prediction effects. The first evidence of this is its alpha versus the linear model shown in Table 1 above. The linear model has access to the same predictors, but incorporating nonlinearities generates significant alpha over the linear model.

The variable importance results show some linear predictors have very impressive individual performance. To show that machine learning performance is not driven by these simple linear effects, Appendix Table 2 reports information ratios of the machine learning strategy on the linear univariate timing strategy of each predictor (the univariate timing strategy is defined as the product of a predictor at time $t$ with the market return at $t+1$).

The machine learning model has a large and highly significant information ratio over every linear strategy. We also calculate its information ratio versus all 15 univariate strategies simultaneously ("All").[42] In this case, we find an information ratio of 0.32 ($t = 2.9$), providing further direct evidence for the nonlinear benefits of complexity.

Naturally, interpretation is a challenge for complex nonlinear models. Appendix Figure 16 makes progress in this direction by illustrating the nonlinear prediction patterns associated with each of the 15 predictors. To trace the impact of predictor $i$ on expected returns, we fix the prediction model estimated from a given training sample and fix the values of all variables other than $i$ at their values at the time of the forecast. Next, we vary the value of the $i^{th}$ predictor from its full sample min (corresponding to $-1$ on the plots) to its full

[42] We cannot run an in-sample versus all 15 univariate strategies simultaneously because this is equivalent to using the in-sample tangency portfolio of the 15 timing strategies as a benchmark. This is not an apples-to-apples comparison because the machine learning strategy is out-of-sample, so it should be benchmarked to a similarly out-of-sample strategy. To this end, we build the out-of-sample tangency portfolio of the 15 timing strategies (scaled to have an expected volatility of 20%) using an expanding window. We use this combined strategy as the regressor when calculating alpha for the "all" case.

sample max (corresponding to +1) and record how the return prediction varies. Then we average this prediction response function across all training windows and plot the result.

The figure illustrates a few interesting patterns. First, we see that when certain indicators of macroeconomic risk are at their lowest (in particular, stock market variance "svar" and credit spreads on risky corporate debt "dfy"), the machine learning model forecasts positive returns. However, once these variables reach even moderate levels, the return prediction drops to zero. This is consistent with the time series pattern in Figure 10, which shows that timing positions (i.e., expected returns) drop to zero heading into recessions. In fact, all predictors demonstrate a similar "risk on/risk off" predictive pattern in which certain values trigger positive market bets, and otherwise, they advocate positions near zero.

### 6.6 The Extent of Nonlinearity and Other Robustness

It is interesting to note that the linear strategy and the nonlinear machine learning strategy each have beneficial performance relative to buy-and-hold. Yet, they are distinct from each other (for example, the nonlinear strategy has significant alpha versus the linear strategy). The parameter $\gamma$ controls the degree of nonlinearity in the RFF approximation. It turns out that the linear kitchen sink regression is equivalent to an RFF model in the limit when $\gamma \approx 0$. In particular, note that

$$\sin(\gamma\omega_i' G_t) \;=\; \gamma\omega_i' G_t \;+\; O(\gamma^2),\; \cos(\gamma\omega_i' G_t) \;=\; 1 - \gamma\omega_i' G_t \;+\; O(\gamma^2). \tag{22}$$

Suppose for simplicity that we only have the sin features. Then, defining $\Omega \;=\; \frac{1}{P^{1/2}}(\omega_i)_{i=1}^P \in \mathbb{R}^{15\times P}$, we have that the model is equivalent to a model with random *linear* features, $S_t = \Omega' G_t$.[43]

This begs the question: Is there an optimal degree of nonlinearity? In general, the answer is no. In the high-complexity regime, different choices of $\gamma$ all deliver *different* approximations of the true DGP, with none strictly dominating the others. Mei et al. (2022) show that high

[43] See Proposition 9 in Appendix E.

model complexity poses an insurmountable obstacle for any random feature regression—it is impossible to learn the "true" dependency $R_{t+1} = f(G_t) + \varepsilon_{t+1}$ when the model is complex. In this case, different random feature generators recover different aspects (projections) of the truth on different subspaces. As a result, we would expect linear and nonlinear random features to contain complementary information, and this is clearly reflected in the results of Table 1.[44]

We assess robustness of our results to various degrees of nonlinearity ($\gamma = 0.5$ or 1, versus $\gamma = 2$ in our main analysis) in Appendix F. We also investigate the effect of excluding volatility standardization of the market return. The brief summary of these analyses is that our conclusions are robust to each variation in empirical design.

Next, we analyze the robustness of our main findings in subsamples. We report model performance splitting the test sample into halves, shown in appendix Figures 20, 21, and 22 for training windows $T = 12$, 60, and 120, respectively. The left side of the each figure reports machine learning timing strategy out-of-sample performance from 1930–1974, and the right side from 1975–2020. The figures show that the patterns of out-of-sample timing strategy performance with respect to complexity and shrinkage do not depend on the subsample. Average out-of-sample returns rise monotonically with complexity and decrease with ridge shrinkage; volatility abates once we move past the interpolation boundary and is further dampened by shrinkage. Information ratios rise with complexity and are fairly insensitive to shrinkage. In the interest of space, we do not plot the out-of-sample $R^2$ or $\hat{\beta}$ norm, but these also follow identical patterns to those for the full sample.

While the patterns are the same across subsamples, the magnitudes differ. Average returns in the second sample are about half as large as the first. But volatilities are roughly the same, so information ratios are about half as large in the second sample. This is consistent with the machine's trading patterns plotted in Figure 10. Starting around 1968, it finds

[44] Relatedly, the machine learning model and the linear kitchen sink (with $z = 10^3$) have alpha versus each other, suggesting that there are benefits to model averaging. For example, an equal-weighted average of the two strategies (after they are re-scaled to have the same volatility) produces a Sharpe ratio of 0.53 and a significant information ratio versus the market of 0.37.

notably fewer buying opportunities and, when it does, takes smaller positions than in the earlier sample.

Finally, we compare the performance of the machine learning model with a 12-month training window to a 12-month time series momentum strategy (Moskowitz et al., 2012). If regressors are highly persistent, they will appear roughly static in a typical 12-month window. In this case, forecasts from a high-complexity regression will behave very similarly to time series momentum.[45] In Appendix G we explain this issue in more detail. We also show that our results are not driven by this "short window and persistent regressor" mechanism. Instead, as emphasized in Section 6.5, our machine learning model performance is driven by relatively high-frequency fluctuations among the predictors. We also show that the machine learning timing strategy has economically large and statistically significant alpha over time series momentum.

## 7 Conclusion

The field of asset pricing is in the midst of a boom in research applications using machine learning. The asset management industry is experiencing a parallel boom in adopting machine learning to improve portfolio construction. However, the properties of portfolios based on such richly parameterized models are not well understood.

In this article, we offer some new theoretical insight into the expected out-of-sample behavior of machine learning portfolios. Building on recent advances in the theory of high-complexity models from the machine learning literature, we demonstrate a theoretical "virtue of complexity" for investment strategies derived from machine learning models. Contrary to conventional wisdom, we prove that market timing strategies based on ridgeless least squares generate positive Sharpe ratio improvements for arbitrarily high levels of model complexity. In other words, the performance of machine learning portfolios can be theoretically improved by pushing model parameterization far beyond the number of training

[45] We are grateful to the Editor for pointing this out.

observations, even when minimal regularization is applied. We provide a rigorous foundation for this behavior rooted in techniques from random matrix theory. We complement these technical developments with intuitive descriptions of the key statistical mechanisms.

In addition to establishing the virtue of complexity, we demonstrate that out-of-sample $R^2$ from a prediction model is generally a poor measure of its economic value. We prove that a market timing model can earn large economic profits when $R^2$ is large and negative. This naturally recommends that the finance profession focus less on evaluating models in terms of forecast accuracy and more on evaluating in economic terms, for example, based on the Sharpe ratio of the associated strategy. We compare and contrast the implications of model complexity for machine learning portfolio performance in correctly specified versus mis-specified models.

Finally, we compare theoretically predicted behavior to the empirical behavior of machine learning-based trading strategies. The theoretical virtue of complexity aligns remarkably closely with patterns in real-world data. In a canonical empirical finance application—market return prediction and concomitant market timing strategies—we find out-of-sample information ratios on the order of 0.3 relative to a market buy-and-hold strategy, and these improvements are highly statistically significant. The emerging strategies have some remarkable attributes, behaving as long-only strategies that divest the market leading up to recessions. Our high-complexity models learn this behavior without guidance from researcher priors or modeling constraints.

Our results are *not* a license to add arbitrary predictors to a model. Instead, we encourage i) including all plausibly relevant predictors and ii) using rich nonlinear models rather than simple linear specifications. Doing so confers prediction and portfolio benefits, even when training data is scarce, particularly when accompanied by prudent shrinkage. Even when the number of raw predictors is small, gains are achieved using those predictors in highly parameterized nonlinear prediction models.

This recommendation clashes with the philosophy of parsimony frequently espoused by economists and famously articulated by the statistician George Box:

> *Since all models are wrong, the scientist cannot obtain a 'correct' one by excessive elaboration. On the contrary, following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.* (Box, 1976)

Our theoretical analysis (along with that of Belkin et al., 2019b; Hastie et al., 2019; Bartlett et al., 2020, among others) shows the flaw in this view—Occam's razor may instead be Occam's blunder. Theoretically, we show that a small model is preferable only if it is correctly specified. But, as Box (1976) emphasizes, models are never correctly specified. The logical conclusion is that large models are preferable under fairly general conditions. The machine learning literature demonstrates the preferability of large models in a wide range of real-world prediction tasks. Our results indicate that the same is likely true in finance and economics.

Our findings point to a number of interesting directions for future work, such as studying the theoretical behavior of high-complexity models in cross-sectional trading strategies and more extensive empirical investigation into the virtue of complexity across different asset markets.

## References


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

**Ali, Alnur, J Zico Kolter, and Ryan J Tibshirani**, "A continuous-time view of early stopping for least squares regression," in "The 22nd International Conference on Artificial Intelligence and Statistics" PMLR 2019, pp. 1370–1378.

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

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

**Bartlett, Maurice S**, "An inverse matrix adjustment arising in discriminant analysis," *The Annals of Mathematical Statistics*, 1951, *22* (1), 107–111.

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

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

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

___ **,** ___ **, Siyuan Ma, and Soumik Mandal**, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proceedings of the National Academy of Sciences*, 2019, *116* (32), 15849–15854.

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

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

**Campbell, John Y and Samuel B Thompson**, "Predicting excess stock returns out of sample: Can anything beat the historical average?," *The Review of Financial Studies*, 2008, *21* (4), 1509–1531.

**Cenesizoglu, Tolga and Allan Timmermann**, "Do return prediction models add economic value?," *Journal of Banking & Finance*, 2012, *36* (11), 2974–2987.

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

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

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

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

**Dong, Xi, Yan Li, David E Rapach, and Guofu Zhou**, "Anomalies and the expected market return," *The Journal of Finance*, 2022, *77* (1), 639–681.

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

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

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

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

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

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

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

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

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

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

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

**___ , ___ , and Athanasse Zafirov**, "A Comprehensive Look at the Empirical Performance ofEquity Premium Prediction II," *Available at SSRN 3929119*, 2021.

**Gu, Shihao, Bryan Kelly, and Dacheng Xiu**, "Empirical asset pricing via machine learning," *The Review of Financial Studies*, 2020, *33* (5), 2223–2273.

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

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

**Hornik, Kurt, Maxwell Stinchcombe, and Halbert White**, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks," *Neural networks*, 1990, *3* (5), 551–560.

**Jacot, Arthur, Franck Gabriel, and Clément Hongler**, "Neural tangent kernel: Convergence and generalization in neural networks," *Advances in neural information processing systems*, 2018, *31.*

**Kelly, Bryan and Dacheng Xiu**, "Financial Machine Learning," Yale Working Paper 2022.

___ **and Seth Pruitt**, "Market expectations in the cross-section of present values," *The Journal of Finance*, 2013, *68* (5), 1721–1756.

**Koijen, Ralph and Stijn Van Nieuwerburgh**, "Predictability of returns and cash flows," *Annual Review of Financial Economics*, 2011, *3*, 467–491.

**Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh**, "Shrinking the cross-section," *Journal of Financial Economics*, 2020, *135* (2), 271–292.

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

___ **and Sandrine Péché**, "Eigenvectors of some large sample covariance matrix ensembles," *Probability Theory and Related Fields*, 2011, *151* (1), 233–264.

**Leitch, Gordon and J Ernest Tanner**, "Economic forecast evaluation: profits versus the conventional error measures," *The American Economic Review*, 1991, pp. 580–590.

**Liu, Fanghui, Xiaolin Huang, Yudong Chen, and Johan AK Suykens**, "Random Features for Kernel Approximation: A Survey on Algorithms, Theory, and Beyond," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2021.

**Ludvigson, Sydney C and Serena Ng**, "The empirical risk–return relation: A factor analysis approach," *Journal of Financial Economics*, 2007, *83* (1), 171–222.

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

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

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

___ **, Theodor Misiakiewicz, and Andrea Montanari**, "Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration," *Applied and Computational Harmonic Analysis*, 2022, *59*, 3–84.

**Moskowitz, Tobias J, Yao Hua Ooi, and Lasse Heje Pedersen**, "Time series momentum," *Journal of financial economics*, 2012, *104* (2), 228–250.

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

___ **and** ___ , "Weighted sums of random kitchen sinks: replacing minimization with randomization in learning.," in "Nips" Citeseer 2008, pp. 1313–1320.

**Rapach, David and Guofu Zhou**, "Forecasting stock returns," in "Handbook of economic forecasting," Vol. 2, Elsevier, 2013, pp. 328–383.

___ **and** ___ , "Asset pricing: Time-series predictability," *Oxford Research Encyclopedia of Economics and Finance*, 2022.

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

___ , **Jack K Strauss, and Guofu Zhou**, "Out-of-sample equity premium prediction: Combination forecasts and links to the real economy," *The Review of Financial Studies*, 2010, *23* (2), 821–862.

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

**Rudi, Alessandro and Lorenzo Rosasco**, "Generalization properties of learning with random features," *Advances in neural information processing systems*, 2017, *30*.

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

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

**Sutherland, Danica J and Jeff Schneider**, "On the error of random fourier features," *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, 2015, pp. 862–871.

**Tsigler, A. and P. L. Bartlett**, "Benign overfitting in ridge regression," 2020.

**Wu, Denny and Ji Xu**, "On the Optimal Weighted $\ell_2$ Regularization in Overparameterized Linear Regression," *Advances in Neural Information Processing Systems*, 2020, *33*, 10112–10123.

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

# INTERNET APPENDIX

## A Proofs

**Proof of Lemma 1**. The proof of Lemma 1 follows directly from Proposition 2.1 in Yaskov (2016). □

**Proof of Proposition 1**. We define $\pi_t^Q = \pi_t(\beta)/(1+(S_t'\beta)^2)$ to be the optimal strategy maximizing the unconditional Sharpe ratio. First we consider $\pi_t^Q$. Then,

$$E[\pi_t^Q R_{t+1}] = E[\pi_t^Q (S_t'\beta)] = E[\frac{(S_t'\beta)^2}{\sigma^2 + (S_t'\beta)^2}]$$

whereas

$$E_t[R_{t+1}^2] = \sigma^2 + (S_t'\beta)^2$$

and hence

$$E[(\pi_t^Q)^2 R_{t+1}^2] = E[\frac{(S_t'\beta)^2 E_t[R_{t+1}^2]}{(\sigma^2 + (S_t'\beta)^2)^2}] = E[\frac{(S_t'\beta)^2}{\sigma^2 + (S_t'\beta)^2}]\,.$$

Thus,

$$SR(R^{\pi^Q}) = \left( E[\frac{(S_t'\beta)^2}{\sigma^2 + (S_t'\beta)^2}] \right)^{1/2}\,.$$

At the same time, for the $\pi_t$ portfolio, we have

$$E[\pi_t R_{t+1}] = E[(\beta' S_t)^2] = E[\beta'\Psi\beta] = \beta'\Psi\beta \tag{23}$$

whereas, defining $\tilde{\beta} = \Psi^{1/2}\beta$ and using that $S_t = \Psi^{1/2} X_t$, we get

$$\begin{aligned} &\sigma^4 E[(\pi_t)^2 R_{t+1}^2] = \sigma^4 E[(\pi_t)^2 E_t[R_{t+1}^2]] = E[((S_t'\beta)^2)^2(\sigma^2 + (S_t'\beta)^2)] \\ &= \sigma^2\beta'\Psi\beta \;+\; E[(S_t'\beta)^4] = \sigma^2\beta'\Psi\beta \;+\; E[(X_t'\tilde{\beta})^4] \\ &= \sigma^2\beta'\Psi\beta \;+\; E[\sum_{i_1,i_2,i_3,i_4} X_{i_1}X_{i_2}X_{i_3}X_{i_4}\tilde{\beta}_{i_1}\tilde{\beta}_{i_2}\tilde{\beta}_{i_3}\tilde{\beta}_{i_4}] \end{aligned} \tag{24}$$

Since all first- and third-order moments of $X$ are zero, the only terms that survive are those there two pairs of $i$ indices are identical, or all of them are identical. For the first one, there are three possibilities, and all second moments of $X_i$ equal one. This gives

$$E[\sum_{i_1,i_2,i_3,i_4} X_{i_1}X_{i_2}X_{i_3}X_{i_4}\tilde{\beta}_{i_1}\tilde{\beta}_{i_2}\tilde{\beta}_{i_3}\tilde{\beta}_{i_4}] = 3\|\tilde{\beta}\|^2 \;+\; \sum_i (E[X_{i,t}^4] - 3)\tilde{\beta}_i^4$$

and hence

$$\sigma^4 E[(\pi_t)^2 R_{t+1}^2] = \sigma^2 \beta' \Psi \beta \;+\; 3(\beta' \Psi \beta)^2 \;+\; \sum_i (E[X_{i,t}^4] - 3)\tilde{\beta}_i^4 \tag{25}$$

The claim of the proposition follows by using Taylor approximation and

$$\frac{(S_t'\beta)^2}{\sigma^2 + (S_t'\beta)^2} = \frac{(S_t'\beta)^2}{\sigma^2}(1 - \frac{(S_t'\beta)^2}{\sigma^2}) \;+\; O(\|\beta\|^6)\,.$$

□

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

**Theorem 8** *For any $c > 0$ and $z < 0$, the distribution of eigenvalues of $\hat{\Psi}$ in the limit as $P, T \to \infty$, $P/T \to c$ converges to a distribution whose Stieltjes transform, $m(z; c)$, is the unique positive solution to the equation*

$$m(z;c) = \frac{1}{1 \;-\; c \;-\; c\, z\, m(z;c)}\, m_\Psi \left( \frac{z}{1 \;-\; c \;-\; c\, z\, m(z;c)} \right). \tag{26}$$

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

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

We will need an auxiliary

**Lemma 2** *For any sequence of bounded matrices $A_P$, we have*

$$P^{-1} S_t' A_P S_t - P^{-1} \operatorname{tr}(A_P \Psi) \;\to\; 0 \tag{27}$$

*is probability.*

**Proof of Lemma 2**. The proof follows directly from Proposition 2.1 in Yaskov (2016).

□

**Lemma 3** *We have*

$$P^{-1} \operatorname{tr}(Q_P (zI + \hat{\Psi}_T)^{-1}) \;-\; E[P^{-1} \operatorname{tr}(Q_P (zI + \hat{\Psi}_T)^{-1})] \;\to\; 0 \tag{28}$$

*almost surely for any sequence of uniformly bounded matrices $Q_P$.*

**Proof of Lemma 3**. The proof follows by the same arguments as in Bai and Zhou (2008). Let $\Psi_{T,t} = \frac{1}{T} \sum_{\tau \neq t} S_\tau S_\tau'$. By the Sherman-Morrison formula (see Bartlett (1951)),

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

(29)

Let $E_\tau$ denote the conditional expectation given $S_{\tau+1}, \cdots, S_T$. Let also

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

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

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

and therefore

$$\begin{aligned}
E[q_T(z)] - q_T(z) \;&=\; E_0[q_T(z)] - E_T[q_T(z)] = \sum_{t=1}^{T} (E_{t-1}[q_T(z)] - E_t[q_T(z)]) \\
&= \sum_{t=1}^{T} (E_{t-1} - E_t)[q_T(z)] \\
&= \sum_{t=1}^{T} (E_{t-1} - E_t)[q_T(z) - \frac{1}{P} \operatorname{tr}(zI + \Psi_{T,t})^{-1} Q_P] \qquad (30) \\
&= \frac{1}{P} \sum_{t=1}^{T} (E_{t-1} - E_t)[\operatorname{tr}(zI + \hat{\Psi}_T)^{-1} Q_P - \operatorname{tr}(zI + \hat{\Psi}_{T,t})^{-1} Q_P] \\
&= -\frac{1}{P} \sum_{\tau=1}^{T} (E_{t-1} - E_t)[\gamma_t] \,,
\end{aligned}$$

where we have used (29) and defined

$$\gamma_t = \operatorname{tr} \left( \frac{1}{T} (zI + \hat{\Psi}_{T,t})^{-1} S_t (I + \frac{1}{T} S_t' (zI + \hat{\Psi}_{T,t})^{-1} S_t)^{-1} S_t' (zI + \hat{\Psi}_{T,t})^{-1} Q_P \right) \qquad (31)$$

We will need the following known properties of the trace:

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

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

*and*

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

Thus,

$$\begin{aligned}\|\gamma_t\| &\leq \|Q_P\| \operatorname{tr}\left(\frac{1}{T}(zI+\hat{\Psi}_{T,t})^{-1}S_t(I+\frac{1}{T}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}\right) \\ &\leq z^{-1}\operatorname{tr}\left(\frac{1}{T}(zI+\hat{\Psi}_{T,t})^{-1/2}S_t(I+\frac{1}{T}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1/2}\right) \\ &= z^{-1}\operatorname{tr}(B(zI+B)^{-1}) \leq z^{-1},\end{aligned} \tag{32}$$

where

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

Thus,

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

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

$$\begin{aligned}E[|q_T(z)-E[q_T(z)]|^q] &\leq K_q P^{-q} E\left(\sum_{t=1}^{T}|(E_{t-1}-E_t)[\gamma_t]|^2\right)^{q/2} \\ &\leq K_q(2N/z)^q P^{-q/2}\left(\frac{P}{T}\right)^{-q/2}.\end{aligned} \tag{33}$$

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

**Lemma 5** *Suppose that*

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

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

**Proof**. It is known that if

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

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

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

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

The proof of Lemma 3 is complete. □

**Proof of Proposition 2**. The proof is based on several steps.

- Let

$$\hat{\Psi}_{T,t} = \frac{1}{T}\sum_{\tau \neq t} S_\tau S_\tau' \,. \tag{34}$$

Then, by the Sherman-Morrison formula (29),

$$\begin{aligned}(zI + \hat{\Psi}_T)^{-1} S_t &= (zI + \hat{\Psi}_{T,t})^{-1} S_t \\ &- \frac{1}{T}(zI + \hat{\Psi}_{T,t})^{-1} S_t S_t' (zI + \hat{\Psi}_{T,t})^{-1} S_t \frac{1}{1 + (T)^{-1} S_t' (zI + \hat{\Psi}_{T,t})^{-1} S_t} \\ &= (zI + \hat{\Psi}_{T,t})^{-1} S_t \frac{1}{1 + (T)^{-1} S_t' (zI + \hat{\Psi}_{T,t})^{-1} S_t} \,.\end{aligned} \tag{35}$$

- By Lemma 2,

$$P^{-1}\, S_t'(zI + \hat{\Psi}_{T,t})^{-1} S_t \; - P^{-1} \operatorname{tr}(\Psi (zI + \hat{\Psi}_{T,t})^{-1}) \;\to\; 0 \tag{36}$$

in probability. At the same time, by Lemma 3,

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

alsmost surely. Thus,

$$P^{-1}\, S_t'(zI + \hat{\Psi}_{T,t})^{-1} S_t \; - E[P^{-1} \operatorname{tr}(\Psi (zI + \hat{\Psi}_{T,t})^{-1})] \;\to\; 0 \tag{37}$$

is probability.

- Theorem 8 implies that

$$P^{-1} \operatorname{tr} E[(zI + \hat{\Psi}_T)^{-1}] \;\to\; m(-z; c) \tag{38}$$

- Now, we have

$$\begin{aligned}1 &= P^{-1} \operatorname{tr} E[(zI + \hat{\Psi}_T)^{-1}(zI + \hat{\Psi}_T)] = P^{-1} \operatorname{tr} E[(zI + \hat{\Psi}_T)^{-1}] z \;+\; P^{-1} \operatorname{tr} E[(zI + \hat{\Psi}_T)^{-1} \hat{\Psi}_T] \\ &= zm(-z, c) \;+\; P^{-1} \operatorname{tr} E[(zI + \hat{\Psi}_T)^{-1} \frac{1}{T} \sum_t S_t S_t'] \\ &= \{symmetry\ across\ t\} \;=\; zm(-z, c) \;+\; P^{-1} \operatorname{tr} E[(zI + \hat{\Psi}_T)^{-1} \frac{1}{N} S_t S_t'] \\ &= \{using\ Sherman - Morrison\ (35)\} \\ &= zm(-z, c) \;+\; P^{-1} \operatorname{tr} E[(zI + \hat{\Psi}_{T,t})^{-1} S_t \frac{1}{1 + (T)^{-1} S_t' (zI + \hat{\Psi}_{T,t})^{-1} S_t} S_t'] \\ &= zm(-z, c) \;+\; E[\frac{P^{-1} S_t' (zI + \hat{\Psi}_{T,t})^{-1} S_t}{1 + (T)^{-1} S_t' (zI + \hat{\Psi}_{T,t})^{-1} S_t}]\end{aligned}$$

(39)

Now, $E[T^{-1}\operatorname{tr}(\Psi(zI+\hat{\Psi}_{T,t})^{-1})] \leq \|\Psi\| z^{-1}$ and hence is uniformly bounded. Let us pick a sub-sequence of $T$ converging to infinity and such that $E[T^{-1}\operatorname{tr}(\Psi(zI+\hat{\Psi}_{T,t})^{-1})] \to q$ for some $q > 0$. By (36),

$$\frac{P^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}{1+(T)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t} \;\to\; \frac{c^{-1}q}{1+q}$$

in probability and this sequence is uniformly bounded. Hence,

$$E\Big[\frac{P^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}{1+(T)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}\Big] \;\to\; \frac{c^{-1}q}{1+q}$$

and we get

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

Thus, the limit of $\xi(z;c) = E[T^{-1}\operatorname{tr}(\Psi(zI+\hat{\Psi}_{T,t})^{-1})]$ is independent of the sub-sequence of $T$ and satisfies the required equation.

The proof of Proposition 2 is complete.

□

**Proof of Proposition 3**. First we show

$$\beta'\Psi\hat{\beta} \;\to\; b_*(\psi_{*,1} - c^{-1}z\xi(z)) \tag{40}$$

in probability, and then we establish the identity

$$\operatorname{tr}(\Psi\hat{\beta}\hat{\beta}') \to\; b_*(\psi_{*,1} - 2zc^{-1}\xi(z) - z^2c^{-1}\xi'(z)) + \xi(z) + z\xi'(z) \tag{41}$$

in probability. We start with the observation that

$$\frac{1}{T}\sum_{t=1}^{T} S_tR_{t+1} = \frac{1}{T}\sum_{t=1}^{T} S_t(S_t'\beta + \varepsilon_{t+1}) = \hat{\Psi}_T\beta \;+\; q_T\,, \tag{42}$$

where we have defined

$$q_T = \frac{1}{T}\sum_{t=1}^{T} S_t'\varepsilon_{t+1}\,. \tag{43}$$

Therefore,

$$\hat{\beta} = (zI+\hat{\Psi}_T)^{-1}(\hat{\Psi}_T\beta \;+\; q_T) \tag{44}$$

By a standard application of the law of large numbers, $q_T \to 0$ in $L_2$ and hence also in probability. We will be using $a \approx b$ to denote that $a - b \to 0$ in probability.

Using (44) and Assumption 4, we have (using that $\varepsilon_t$ are independent of $S_t$ and have zero means) that

$$\begin{aligned}
&\beta' E[S_t S_t'] \hat{\beta} \\
&= \beta' E[S_t S_t'](zI + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T \beta \ + \ q_T) \\
&\approx \beta' \Psi \, (zI + \hat{\Psi}_T)^{-1} \hat{\Psi}_T \beta \\
&= \{by \ Lemma \ 1\}\} \\
&\overset{prob}{\to} \ b_* \, P^{-1} \operatorname{tr} E[\Psi \, (zI + \hat{\Psi}_T)^{-1} \hat{\Psi}_T] \\
&= b_* P^{-1} \operatorname{tr} E[\Psi \, (zI + \hat{\Psi}_T)^{-1}(zI + \hat{\Psi}_T - zI)] \\
&= b_* P^{-1} \operatorname{tr} E[\Psi - z\Psi \, (zI + \hat{\Psi}_T)^{-1}] \\
&= b_* P^{-1} \left( \operatorname{tr} \Psi \ - \ z \operatorname{tr} E[(zI + \hat{\Psi}_T)^{-1} \Psi] \right) \\
&= \{by \ Proposition \ 2\} \\
&\to_{T \to \infty} \ b_* \nu(z) \, .
\end{aligned} \tag{45}$$

At the same time,

$$\begin{aligned}
&\operatorname{tr}(\Psi \hat{\beta} \hat{\beta}') \\
&= \operatorname{tr}(\Psi (zI + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T \beta \ + \ q_T)(\hat{\Psi}_T \beta \ + \ q_T)'(zI + \hat{\Psi}_T)^{-1}) \\
&= \operatorname{tr}(\Psi (zI + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T \beta \ + \ q_T)(\beta' \hat{\Psi}_T \ + \ q_T')(zI + \hat{\Psi}_T)^{-1}) \\
&\approx \operatorname{tr}(\Psi (zI + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T \beta \beta' \hat{\Psi}_T \ + \ q_T q_T')(zI + \hat{\Psi}_T)^{-1})
\end{aligned} \tag{46}$$

where we have used the fact that the terms that are linear in $q_T$ converge to zero in probability. Now,

$$\begin{aligned}
E[q_T q_T' | S] &= \frac{1}{T^2} E[\sum_t S_t \varepsilon_{t+1} \sum_{t_1} \varepsilon_{t_1+1} S_{t_1}' | S] \\
&= \frac{1}{T^2} E[\sum_{t,t_1} S_t \varepsilon_{t+1} \varepsilon_{t_1+1}' S_{t_1}' | S] \\
&= \frac{1}{T^2} E[\sum_t S_t \varepsilon_{t+1} \varepsilon_{t+1}' S_t' | S] \\
&= \frac{1}{T^2} \sum_t S_t E[\varepsilon_{t+1} \varepsilon_{t+1}' | S] S_t' \\
&= \frac{1}{T^2} \sum_t S_t \sigma^2 S_t' \\
&= \frac{1}{T^2} \sum_t S_t S_t' = \frac{1}{T} \hat{\Psi}_T \, ,
\end{aligned} \tag{47}$$

and it is straightforward to show that the contributions coming from

$$T^{-2}\sum_{t,t_1} S_t(\varepsilon_{t+1}\varepsilon'_{t_1+1}-1)S'_{t_1}$$

are converge to zero in probability. Therefore, (46) takes the form

$$\begin{aligned}
&\mathrm{tr}(\Psi\hat{\beta}\hat{\beta}') \\
&= \mathrm{tr}(\Psi(zI+\hat{\Psi}_T)^{-1}(\hat{\Psi}_T\beta\beta'\hat{\Psi}_T \;+\; q_Tq_T')(zI+\hat{\Psi}_T)^{-1}) \\
&= \mathrm{tr}(\Psi(zI+\hat{\Psi}_T)^{-1}(\hat{\Psi}_T\beta\beta'\hat{\Psi}_T \;+\; \frac{1}{T}\hat{\Psi}_T)(zI+\hat{\Psi}_T)^{-1}) \\
&= \mathrm{tr}(\Psi(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T\beta\beta'\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}) \\
&+ \mathrm{tr}(\Psi(zI+\hat{\Psi}_T)^{-1}\frac{1}{T}\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}) \\
&= \mathrm{tr}(\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}\Psi(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T\beta\beta') \\
&+ \mathrm{tr}(\Psi(zI+\hat{\Psi}_T)^{-1}\frac{1}{T}(zI+\hat{\Psi}_T-zI)(zI+\hat{\Psi}_T)^{-1}) \\
&= \{by\ Lemmas\ 1,\ 3\ and\ Vitali's\ thorem\} \\
&\overset{prob}{\to} b_*P^{-1}\,\mathrm{tr}\,E[\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}\Psi(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T] \\
&+ \frac{1}{T}\,\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}(zI+\hat{\Psi}_T)(zI+\hat{\Psi}_T)^{-1}]) \\
&- z\frac{1}{T}\,\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}(zI+\hat{\Psi}_T)^{-1}]) \\
&= b_*P^{-1}\,\mathrm{tr}\,E[\Psi(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}] \\
&+ \frac{1}{T}\,\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}]) \\
&- z\frac{1}{T}\,\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-2}]) \\
&= Term1 \;+\; Term2 \;+\; Term3\,.
\end{aligned} \tag{48}$$

We now proceed with each term:

$$\begin{aligned}
&(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1} \;= \{\text{all matrices commute}\} = (zI+\hat{\Psi}_T)^{-2}\hat{\Psi}_T^2 \\
&= (zI+\hat{\Psi}_T)^{-2}(\hat{\Psi}_T^2+2z\hat{\Psi}_T+z^2I-2z\hat{\Psi}_T-z^2I) \\
&= (zI+\hat{\Psi}_T)^{-2}(\hat{\Psi}_T^2+2z\hat{\Psi}_T+z^2I-2z(\hat{\Psi}_T+zI)+z^2I) \\
&= (zI+\hat{\Psi}_T)^{-2}\Big((zI+\hat{\Psi}_T)^2-2z(\hat{\Psi}_T+zI)+z^2I\Big) \\
&= I \;-\; 2z(zI+\hat{\Psi}_T)^{-1}+z^2(zI+\hat{\Psi}_T)^{-2}\,.
\end{aligned} \tag{49}$$

Therefore,

$$\begin{aligned}
&Term1 = b_*P^{-1}\,\mathrm{tr}\,E[\Psi(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}] \\
&= b_*P^{-1}\,\mathrm{tr}\,E[\Psi(I \;-\; 2z(zI+\hat{\Psi}_T)^{-1}+z^2(zI+\hat{\Psi}_T)^{-2})]\,,
\end{aligned} \tag{50}$$

and

$$Term2 = \frac{1}{T}\operatorname{tr}(\Psi E[(zI + \hat{\Psi}_T)^{-1}]) \;\to\; \xi(z)$$

by Proposition 2, and hence

$$\frac{d}{dz}\frac{1}{T}\operatorname{tr}(\Psi E[(zI + \hat{\Psi}_T)^{-1}]) \;\to\; \frac{d}{dz}\xi(z) \tag{51}$$

by the Vitali theorem. However,

$$\frac{d}{dz}\operatorname{tr}(\Psi E[(zI + \hat{\Psi}_T)^{-1}]) = -\operatorname{tr}(\Psi E[(zI + \hat{\Psi}_T)^{-2}]) \tag{52}$$

and hence

$$\frac{1}{T}\operatorname{tr}(\Psi E[(zI + \hat{\Psi}_T)^{-2}]) \;\to\; -\frac{d}{dz}\xi(z)\,. \tag{53}$$

Summarizing, we get

$$Term3 \;\to\; z\frac{d}{dz}\xi(z)\,,$$

whereas

$$Term1 \;\to\; b_* P^{-1}\operatorname{tr} E[\Psi(I - 2z(zI+\hat{\Psi}_T)^{-1}+z^2(zI+\hat{\Psi}_T)^{-2})] \;\to\; b_*(\psi_{*,1}-2zc^{-1}\xi(z)-z^2c^{-1}\xi'(z)) \tag{54}$$

and hence

$$\begin{aligned}
&\operatorname{tr}(\Psi E[\hat{\beta}\hat{\beta}']) = \; Term1 \;+\; Term2 \;+\; Term3 \\
&\overset{prob}{\to} b_*(\psi_{*,1} - 2zc^{-1}\xi(z) - z^2c^{-1}\xi'(z)) + \xi(z) + z\frac{d}{dz}\xi(z) \\
&= \; b_*\,\hat{\nu}(z;c) \;-\; c\,\nu'(z;c)
\end{aligned} \tag{55}$$

Now, by (8), we have

$$MSE \;\to\; E[R_{t+1}^2] \;-\; 2\,E[\hat{\beta}' S_t S_t' \beta] \;+\; \operatorname{tr} E[\hat{\beta}\beta'\Psi] \tag{56}$$

and therefore equations (45) and (55) imply that

$$MSE \;\to\; E[R_{t+1}^2] \;-\; 2\,\mathcal{E}(z;c) + \;\mathcal{L}(z;c) \tag{57}$$

and hence

$$R^2(z;c) \;=\; 1 - \frac{MSE}{E[R_{t+1}^2]} \;\to\; \frac{2\,\mathcal{E}(z;c) + \;\mathcal{L}(z;c)}{E[R_{t+1}^2]} \tag{58}$$

whereas

$$\begin{aligned} 2\mathcal{E}(z;c) - \mathcal{L}(z;c) &= \lim P^{-1} \operatorname{tr}(2b_*\hat{\Psi}(zI+\hat{\Psi}) - b_*\hat{\Psi}^2 - c\hat{\Psi})(zI+\hat{\Psi})^{-2}\Psi) \\ &= \lim P^{-1} \operatorname{tr}(\hat{\Psi}(2b_*z - c) + b_*\hat{\Psi}^2)(zI+\hat{\Psi})^{-2}\Psi) \end{aligned} \tag{59}$$

and the optimality of $z_* = c/b_*$ follows because the function $f(z) = ((2b_*z - c)\lambda + b_*\lambda^2)/(z+\lambda)^2$ attains its maximum at $z = z_*$ for any value of $\lambda > 0$. The proof of the first part of Proposition 3 is complete.

To study the ridgeless limit, we will need the following auxiliary result.

**Lemma 6** *Suppose that $c > 1$. Then,*

$$m(z;c) = (1-c^{-1})z^{-1} + m_*(c) + n_*(c)z \ + \ O(z^2), \ z \to \ 0\,. \tag{60}$$

*Furthermore,*

$$\begin{aligned} m_*(c) &= c^{-1}((\sigma_*\psi_{*,1})^{-1}c^{-1} + \sigma_*^{-1}\psi_{*,2}\psi_{*,1}^{-3}c^{-2}) \ + \ O(c^{-4}) \\ n_*(c) &= c^{-1}\left(-(\sigma_*\psi_{*,1})^2c^2 + 3\sigma_*^2\psi_{*,2}c\right)^{-1} \ + \ O(c^{-5})\,. \end{aligned} \tag{61}$$

**Proof of Lemma 6**. Let $\sigma_* = 1$.

**Case 1:** $c > 1$ Substituting

$$\tilde{m}(-z;c) = (1-c)z^{-1} \ + \ cm(-z;c)\,, \tag{62}$$

into the equation of Theorem 8, we get that $\tilde{m}$ satisfies

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

Our goal is to understand what happens when $z \to 0$. We have

$$\int \frac{(1-(c-1)\tilde{m}x)\,dH(x)}{\tilde{m}(1+\tilde{m}\,x)} = 0$$

always has a finite solution $\tilde{m}_*(0,c) > 0$ because

$$\frac{\int \frac{dH(x)}{\tilde{m}(1+\tilde{m}\,x)}}{\int \frac{x\,dH(x)}{(1+\tilde{m}\,x)}}$$

is monotone decreasing in $\tilde{m}$, from $+\infty$ to 0 and hence it crosses the level $c-1$ somewhere. Thus, $\tilde{m}_*(c)$ is the unique solution to

$$\frac{\int \frac{dH(x)}{\tilde{m}(1+\tilde{m}\,x)}}{\int \frac{x\,dH(x)}{(1+\tilde{m}\,x)}} = c - 1\,. \tag{63}$$

and $\tilde{m}(z)$ stays bounded and smooth when $z \to 0+$ by the implicit function theorem.

Furthermore, substituting $\tilde{m}(0,c) = ac^{-1} + bc^{-2}$, we get

$$\int \frac{dH(x)}{(ac^{-1}+bc^{-2})(1+(ac^{-1}+bc^{-2})\,x)} = (c-1)\int \frac{x\,dH(x)}{(1+(ac^{-1}+bc^{-2})\,x)} \tag{64}$$

that is (up to negligible terms)

$$\begin{aligned} & a^{-1}c\int (1-bc^{-1}/a+(bc^{-1}/a)^2)(1-(ac^{-1}+bc^{-2})\,x+(ac^{-1}+bc^{-1})^2\,x^2)dH(x) \\ &= \ (c-1)\int x(1-(ac^{-1}+bc^{-2})\,x+(ac^{-1}+bc^{-2})^2\,x^2)dH(x) \end{aligned} \tag{65}$$

Equating the coefficient on $c$ gives

$$a^{-1}c = c\sigma_*\psi_*$$

while the constant coefficient gives

$$-ba^{-2} - \sigma_*\psi_{*,1} = -a\sigma_*^2\psi_{*,2} - \sigma_*\psi_{*,1}$$

and hence

$$a = (\sigma_*\psi_{*,1})^{-1},\ b = a^3\sigma_*^2\psi_{*,2} = \sigma_*^{-1}\psi_{*,2}/\psi_{*,1}^3$$

and

$$m_*(c) = c^{-1}\tilde{m}_*(c) \underbrace{\sim}_{c\to\infty} c^{-1}((\sigma_*\psi_{*,1})^{-1}c^{-1} + \sigma_*^{-1}\psi_{*,2}\psi_{*,1}^{-3}c^{-2}) \tag{66}$$

Thus,

$$cm'(-z;c) = (1-c)z^{-2} \ + \ \tilde{m}'(-z;c) = (1-c)z^{-2} \ + \ O(1)$$

Differentiating the identity

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

with respect to $z$, we get

$$\tilde{m}'(0)\left(-\int \frac{(1+2\tilde{m}_*x)dH(x)}{(\tilde{m}_*(1+\tilde{m}_*\,x))^2} + (c-1)\int \frac{x^2\,dH(x)}{(1+\tilde{m}_*\,x)^2}\right) = 1\,.$$

Furthermore,

$$\int \frac{dH(x)}{\tilde{m}_*(1+\tilde{m}_*\,x)} \ - \ (c-1)\int \frac{x\,dH(x)}{(1+\tilde{m}_*\,x)} \ = 0\,,$$

and therefore

$$\begin{aligned}(c-1)\int \frac{x^2\,dH(x)}{(1+\tilde{m}_*\,x)^2} \;&<\; (c-1)\int \frac{x\,dH(x)}{\tilde{m}_*(1+\tilde{m}_*\,x)} = \int \frac{dH(x)}{\tilde{m}_*^2(1+\tilde{m}_*\,x)} \\ =\; \int \frac{(1+\tilde{m}_*\,x)dH(x)}{\tilde{m}_*^2(1+\tilde{m}_*\,x)^2} \;&<\; \int \frac{(1+2\tilde{m}_*\,x)dH(x)}{\tilde{m}_*^2(1+\tilde{m}_*\,x)^2}\end{aligned} \tag{67}$$

and the claim follows with

$$n_*(c) = c^{-1}\left(-\int \frac{(1+2\tilde{m}_*x)dH(x)}{(\tilde{m}_*(1+\tilde{m}_*\,x))^2} + (c-1)\int \frac{x^2\,dH(x)}{(1+\tilde{m}_*\,x)^2}\right)^{-1} \;<\; 0\,.$$

We have

$$(c-1)\int \frac{x^2\,dH(x)}{(1+\tilde{m}_*\,x)^2} = (c-1)\frac{1}{\tilde{m}_*^2}\int \frac{((1+\tilde{m}_*\,x)^2-1-2\tilde{m}_*x)\,dH(x)}{(1+\tilde{m}_*\,x)^2}$$

Furthermore,

$$\begin{aligned}cn_*(c) \;&\sim\; \left(-a^{-2}c^2\int \frac{(1+2(ac^{-1}+bc^{-2})x)dH(x)}{((1+bc^{-1}/a)(1+(ac^{-1}+bc^{-2})\,x))^2} + (c-1)\int \frac{x^2\,dH(x)}{(1+(ac^{-1}+bc^{-2})\,x)^2}\right)^{-1} \\ &\sim\; \left(-a^{-2}c^2\int (1+2ac^{-1}x-2bc^{-1}/a-2ac^{-1}x)dH(x) + c\int x^2(1-2ac^{-1}x)\,dH(x)\right)^{-1} \\ &\sim\; \left(-a^{-2}c^2+(2a^{-3}b+\sigma_*^2\psi_{*,2})c\right)^{-1} \\ &=\; \left(-(\sigma_*\psi_{*,1})^2c^2+(2(\sigma_*\psi_{*,1})^3\sigma_*^{-1}\psi_{*,2}/\psi_{*,1}^3+\sigma_*^2\psi_{*,2})c\right)^{-1} \\ &=\; \left(-(\sigma_*\psi_{*,1})^2c^2+3\sigma_*^2\psi_{*,2}c\right)^{-1}\end{aligned} \tag{68}$$

when $c\to\infty$. □

We will now use this lemma to prove the behavior of the ridgeless limit. We have

$$\begin{aligned}\xi(z) &= -1+c^{-1}/(c^{-1}-1+zm(-z;c)) \\ &=\; -1+c^{-1}/(zm_*(c)+z^2n_*(c)+O(z^3)) \\ &=\; -1+c^{-1}(zm_*(c))^{-1}/(1+zn_*(c)/m_*(c)+O(z^2)) \\ &=\; -1+c^{-1}(zm_*(c))^{-1}(1-zn_*(c)/m_*(c)+O(z^2)) \\ &= -1+c^{-1}(zm_*(c))^{-1} \;-\; c^{-1}n_*(c)m_*(c)^{-2} \;+\; O(z)\end{aligned} \tag{69}$$

and hence

$$\nu'(z) = -c^{-1}(\xi+z\xi') = -c^{-1}(-1-c^{-1}n_*(c)m_*(c)^{-2} \;+\; O(z))$$

converges to a finite limit when $z\to 0$. Thus,

$$\mathcal{L}(z;c) = b_*(\nu+z\nu') - c\nu' = b_*(\psi_{*,1}-c^{-2}m_*(c)^{-1}) + (-1-c^{-1}n_*(c)m_*(c)^{-2}) \;+\; O(z)$$

Hence,

$$2\mathcal{E}(0;c) \;-\; \mathcal{L}(0;c) = b_*(\psi_{*,1} - c^{-2}m_*(c)^{-1}) + (1 + c^{-1}n_*(c)m_*(c)^{-2})$$

The proof of Proposition 3 is complete. □

**Lemma 7** *Let $a = \sigma_*$. We have*

$$1 - zm(z) = \psi_{*,1}az^{-1} - z^{-2}(\psi_{*,2} + c\psi_{*,1}^2)a^2 + z^{-3}a^3(\psi_{*,3} + 3c\psi_{*,2}\psi_{*,1} + c^2\psi_{*,1}^3) \;+\; O(z^{-4}) \quad (70)$$

*for $z \to \infty$.*

**Proof of Lemma 7**. Then, Theorem 8 implies

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

implying that $zm(z) \to 1$ when $z \to \infty$, whereas

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

and therefore

$$1 - zm(z) \;\sim\; z^{-1}a\psi_{*,1}\,,$$

and

$$\begin{aligned}
&1 - zm(-z) - \psi_{*,1}az^{-1} \\
&= \; (1 - c + czm(z)) \int \frac{xdH(x)}{x(1 - c + czm(-z)) + z} - \psi_{*,1}az^{-1} \\
&= \; (1 - cz^{-1}a\psi_{*,1} + O(z^{-2}))z^{-1} \int \frac{xdH(x)}{xz^{-1}(1 - cz^{-1}a\psi_{*,1} + O(z^{-2})) + 1} - \psi_{*,1}az^{-1} \\
&\sim \; (1 - cz^{-1}a\psi_{*,1} + O(z^{-2}))z^{-1} \int \frac{xdH(x)}{xz^{-1} + 1} - \psi_{*,1}az^{-1} \qquad (71) \\
&\sim \; (1 - cz^{-1}a\psi_{*,1} + O(z^{-2}))z^{-1} \int (x - x^2z^{-1})dH(x) - \psi_{*,1}az^{-1} \\
&\sim \; z^{-1}\psi_{*,1}a - \psi_{*,2}a^2z^{-2} - cz^{-2}a^2\psi_{*,1}^2 - \psi_{*,1}az^{-1} \;+\; O(z^{-3}) \\
&= \; -z^{-2}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \;+\; O(z^{-3})
\end{aligned}$$

Now, we can expand to the higher order. We have

$$1 - c + czm(-z) = 1 - c(1 - zm(-z)) = 1 - cz^{-1}(\psi_{*,1}a - z^{-1}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \;+\; O(z^{-2}))$$

and hence

$$
\begin{aligned}
&1 - zm(-z) - \psi_{*,1} a z^{-1} + z^{-2}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \\
&= \ (1 - cz^{-1}(\psi_{*,1}a - z^{-1}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \ + \ O(z^{-2}))) \\
&\times \int \frac{xdH(x)}{x(1 - cz^{-1}(\psi_{*,1}a - z^{-1}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \ + \ O(z^{-2}))) + z} - \psi_{*,1} a z^{-1} + z^{-2}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \\
&= \ (1 - cz^{-1}(\psi_{*,1}a - z^{-1}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \ + \ O(z^{-2}))) \\
&\times z^{-1} \int \frac{xdH(x)}{xz^{-1}(1 - cz^{-1}\psi_{*,1}a) + 1 + O(z^{-3})} - \psi_{*,1} a z^{-1} + z^{-2}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \\
&\sim \ (1 - cz^{-1}(\psi_{*,1}a - z^{-1}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \ + \ O(z^{-2})))z^{-1} \int x(1 - xz^{-1}(1 - cz^{-1}\psi_{*,1}a) + x^2 z^{-2}) \\
&- \psi_{*,1} a z^{-1} + z^{-2}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \\
&\sim \ (1 - cz^{-1}(\psi_{*,1}a - z^{-1}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \ + \ O(z^{-2})))z^{-1} \left( \psi_{*,1}a - z^{-1}\psi_{*,2}a^2 + z^{-2}a^3(\psi_{*,3} + c\psi_{*,2}\psi_{*,1}) \right) \\
&- \psi_{*,1} a z^{-1} + z^{-2}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \\
&= \ \psi_{*,1} a z^{-1} - z^{-2}\psi_{*,2}a^2 + z^{-3}a^3(\psi_{*,3} + c\psi_{*,2}\psi_{*,1}) \\
&- cz^{-2}\psi_{*,1}a(\psi_{*,1}a - z^{-1}\psi_{*,2}a^2) + cz^{-3}(\psi_{*,2} + c\psi_{*,1}^2)a^2\psi_{*,1}a \ + \ O(z^{-4}) - \psi_{*,1} a z^{-1} + z^{-2}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \\
&= \ z^{-3}a^3(\psi_{*,3} + c\psi_{*,2}\psi_{*,1}) \\
&- cz^{-2}\psi_{*,1}a(-z^{-1}\psi_{*,2}a^2) + cz^{-3}(\psi_{*,2} + c\psi_{*,1}^2)a^2\psi_{*,1}a \ + \ O(z^{-4}) \\
&= \ z^{-3}a^3(\psi_{*,3} + 3c\psi_{*,2}\psi_{*,1} + c^2\psi_{*,1}^3) \ + \ O(z^{-4}) \, .
\end{aligned}
\tag{72}
$$

The proof of Lemma 7 is complete. □

# B Proofs for the Mis-specified Model

We will be using a slightly simpler notation $S_{t,1} = S_t^{(1)}$ and $S_{t,2} = S_t^{(2)}$. Then,

$$
\begin{aligned}
&MSE = E[\|R_{t+1} - S_{t,1}\hat{\beta}\|^2] \\
&= \ \operatorname{tr} E[R_{t+1}R_{t+1}'] \ - \ 2E[\beta' S_t' S_{t,1}\hat{\beta}_1] \ + \ \operatorname{tr} E[S_{t,1}\hat{\beta}_1\hat{\beta}_1' S_{t,1}'] \\
&= \ \operatorname{tr} E[R_{t+1}R_{t+1}'] \ - \ 2E[\beta' S_t' S_{t,1}\hat{\beta}_1] \ + \ \operatorname{tr} E[\Psi_{1,1}\hat{\beta}_1\hat{\beta}_1']
\end{aligned}
\tag{73}
$$

where $\hat{\beta}_1$ is the estimate of the first component of the whole $\beta$ vector. We will also denote $c_1 = cq = P_1/T$ and omit the dependence on $q$ in all the functions. Finally, we will use the notation $\xi_{1,1}(z) = \lim T^{-1} \operatorname{tr} E[(zI + \hat{\Psi})^{-1}\Psi]$ to denote $\xi(z; cq; q)$.

The following is true.

**Lemma 8** *We have*

$$\begin{aligned}\mathcal{E}(z;c_1) &= \lim E[\beta' S_t' S_{t,1} \hat{\beta}_1] \\ &= b_* \frac{c_1}{c}(\psi_{*,1} - c_1^{-1} z\xi(z)) \ + \ b_* \frac{c^{-1}\xi_{2,1}(z)}{1+\xi_{1,1}(z)}\end{aligned} \tag{74}$$

*where*

$$\xi_{2,1}(z) = \lim_{T\to\infty} \frac{1}{T} \operatorname{tr} E[\Psi_{1,2}\Psi_{2,1}(zI + \hat{\Psi}_{T,1,t})^{-1}] \tag{75}$$

**Proof of Lemma 8**. We have

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

and

$$\frac{1}{T}\sum_t S_{t,1}' R_{t+1} = \frac{1}{T}\sum_{t=1}^T S_{t,1}'(S_t\beta + \varepsilon_{t+1}) = \ \hat{\Psi}_T\beta \ + \ q_T\,, \tag{76}$$

where

$$q_T = \frac{1}{T}\sum_{t=1}^T S_{t,1}'\varepsilon_{t+1} \tag{77}$$

and

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

where

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

Therefore,

$$\hat{\beta} = (zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_{T,1}\beta_1 + \hat{\Psi}_{T,2}\beta_2 \ + \ q_T)\,. \tag{78}$$

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

zero means) that

$$
\begin{aligned}
& E[\beta' S_t' S_{t,1} \hat{\beta}] \\
&= E[(\beta_1' \Psi_{1,1} + \beta_2' \Psi_{2,1})(zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_{T,1}\beta_1 + \hat{\Psi}_{T,2}\beta_2)] \\
&= \operatorname{tr} E[\Psi_{1,1}\,(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_T \beta\beta'] \\
&+ E[\beta_1' \Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\beta_2] \\
&+ E[\beta_2' \Psi_{2,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}\beta_1] \\
&+ E[\beta_2' \Psi_{2,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\beta_2] \\
&= \{by\ Lemma\ 1\} \\
&\xrightarrow{prob} b_*\, P^{-1} \operatorname{tr} E[\Psi_{1,1}\,(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}] \;+\; P^{-1} b_* \operatorname{tr} E[\Psi_{2,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}]\,.
\end{aligned}
\tag{79}
$$

The first term is

$$
\begin{aligned}
& P^{-1} \operatorname{tr} E[\Psi_{1,1}\,(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}] \\
&= P^{-1} \operatorname{tr} E[\Psi_{1,1}\,(zI + \hat{\Psi}_{T,1})^{-1}(zI + \hat{\Psi}_{T,1} - zI)] \;\to\; \frac{c_1}{c}(\psi_{*,1} - c_1^{-1} z \xi_{1,1}(z))\,.
\end{aligned}
\tag{80}
$$

To compute the second term in (79), we will need the following lemma.

**Lemma 9**

$$
\frac{1}{P} \operatorname{tr} E[\Psi_{2,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}] \;\to\; c^{-1}\xi_{2,1}(z)/(1 + \xi_{1,1}(z)) \tag{81}
$$

**Proof of Lemma 9**. We have that, by symmetry over time, and using the Sherman-Morrison formula (29), we get

$$
\begin{aligned}
& \frac{1}{P} \operatorname{tr} E[\Psi_{2,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}] \\
&= \frac{1}{P} \operatorname{tr} E[\Psi_{2,1}(zI + \hat{\Psi}_{T,1})^{-1} \frac{1}{T}\sum_{t=1}^{T} S_{t,1} S_{t,2}'] \\
&= \frac{1}{P} \operatorname{tr} E[\Psi_{2,1}(zI + \hat{\Psi}_{T,1})^{-1} S_{t,1} S_{t,2}'] \\
&= \frac{1}{P} \operatorname{tr} E[\Psi_{2,1}\Big((zI + \hat{\Psi}_{T,1,t})^{-1} \\
&- \frac{1}{T}(zI + \hat{\Psi}_{T,1,t})^{-1} S_{t,1}(I + \frac{1}{T} S_{t,1}'(zI + \hat{\Psi}_{T,1,t})^{-1} S_{t,1})^{-1} S_{t,1}'(zI + \hat{\Psi}_{T,1,t})^{-1}\Big) S_{t,1} S_{t,2}'] \\
&= \frac{1}{P} \operatorname{tr} E[\Psi_{2,1}(zI + \hat{\Psi}_{T,1,t})^{-1} S_{t,1} S_{t,2}'] \\
&- \frac{1}{P} \operatorname{tr} E[\Psi_{2,1}(zI + \hat{\Psi}_{T,1,t})^{-1} S_{t,1}(I + C_T)^{-1} C_T S_{t,2}'] \\
&= \frac{1}{P} \operatorname{tr} E[\Psi_{2,1}(zI + \hat{\Psi}_{T,1,t})^{-1} \Psi_{2,1}'] \\
&- \frac{1}{P} E[S_{t,2}' \Psi_{2,1}(zI + \hat{\Psi}_{T,1,t})^{-1} S_{t,1}(1 + C_T)^{-1} C_T]
\end{aligned}
\tag{82}
$$

where we have defined

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

By Lemma 2 and (36),

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

in probability. Furthermore, $(1 + C_T)^{-1} C_T$ is uniformly bounded.

By a similar argument,

$$\frac{1}{T} S'_{t,2} \Psi_{2,1}(zI + \hat{\Psi}_{T,1,t})^{-1} S_{t,1} \to \xi_{2,1}(z) \tag{83}$$

in probability, and these variables have uniformly bounded $L_2$ norms. We will need another auxiliary lemma.

**Lemma 10** *Suppose that $X_T - X \to 0$ and $Y_T - Y \to 0$ in $L_2$, and all variables have uniformly bounded $L_2$ norms. Then, $E[X_T Y_T] - E[XY] \to 0$.*

**Proof**. We have

$$E[X_T Y_T] - E[XY] = E[(X_T - X)Y_T] \; + \; E[X(Y_T - Y)]$$

and the claim follows from the Cauchy-Schwarz inequality. □

Thus,

$$\begin{aligned}
&\frac{1}{P} \operatorname{tr} E[\Psi_{2,1}(zI + \hat{\Psi}_{T,1})^{-1} \hat{\Psi}_{T,2}] \\
&= \; \frac{1}{P} \operatorname{tr} E[\Psi_{2,1}(zI + \hat{\Psi}_{T,1,t})^{-1} \Psi'_{2,1}] \\
&- \frac{1}{P} \operatorname{tr} E[\Psi_{2,1}(zI + \hat{\Psi}_{T,1,t})^{-1} S_{t,1}(I + C_T)^{-1} C_T S'_{t,2}] \\
&\to \; c^{-1}\xi_{2,1}(z) \; - \; c^{-1}\xi_{2,1}(z)\xi_{1,1}(z)/(1 + \xi_{1,1}(z)) \\
&= \; c^{-1}\xi_{2,1}(z)/(1 + \xi_{1,1}(z)) \,,
\end{aligned} \tag{84}$$

The proof of Lemma 9 is complete. □

Lemma 8 follows now from (79) □

**Lemma 11** *We have*

$$\begin{aligned}
&\mathcal{L}(z) = \lim \operatorname{tr}(\Psi_{1,1} E[\hat{\beta}\hat{\beta}']) \\
&= \; \frac{c_1}{c} b_*(\psi_{*,1}(q) - 2zc_1^{-1}\xi_{1,1}(z) - z^2 c_1^{-1}\xi'_{1,1}(z)) \; + \; (1 + b_* P^{-1} \operatorname{tr} \Psi_{2,2})(\xi_{1,1}(z) + z\xi'_{1,1}(z)) \\
&+ b_*(1 + \xi(z))^{-2} c_1^{-1} \widehat{\xi}_{2,1} \\
&- 2b_*(\xi_{1,1}(z) + z\xi'_{1,1}(z))(1 + \xi_{1,1}(z))^{-1} c_1^{-1} \xi_{2,1}(z)
\end{aligned} \tag{85}$$

**Proof of Lemma 11**. Let $\hat{\Psi}_T(1,:)$ be the first row in the $2 \times 2$ block representation of $\hat{\Psi}$. Then,

$$\begin{aligned}
&\operatorname{tr}(\Psi_{1,1} E[\hat{\beta}\hat{\beta}']) \\
&= \operatorname{tr}(\Psi_{1,1} E[(zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T(1,:)\beta \ + \ q_T)(\hat{\Psi}_T(1,:)\beta \ + \ q_T)'(zI + \hat{\Psi}_{T,1})^{-1}]) \\
&= \operatorname{tr}(\Psi_{1,1} E[(zI + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T(1,:)\beta \ + \ q_T)(\beta'\hat{\Psi}_T(1,:)' \ + \ q_T')(zI + \hat{\Psi}_{T,1})^{-1}]) \qquad (86) \\
&= \operatorname{tr}(\Psi_{1,1} E[(zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T(1,:)\beta\beta'\hat{\Psi}_T(1,:)' \ + \ q_T q_T')(zI + \hat{\Psi}_{T,1})^{-1}]) \\
&= \operatorname{tr}(\Psi_{1,1} E[(zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T(1,:)\beta\beta'\hat{\Psi}_T(1,:)' \ + \ q_T q_T')(zI + \hat{\Psi}_{T,1})^{-1}])
\end{aligned}$$

Formula (47) still holds with $\hat{\Psi}$ replaced by $\hat{\Psi}_{1,1}$ and calculations in (48) imply

$$\operatorname{tr}(\Psi_{1,1} E[(zI + \hat{\Psi}_{T,1})^{-1} q_T q_T' (zI + \hat{\Psi}_{T,1})^{-1}]) \ \to \ \xi_{1,1}(z) + z\xi_{1,1}'(z)\,. \qquad (87)$$

It remains to deal with

$$\begin{aligned}
&\operatorname{tr}(\Psi_{1,1} E[(zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T(1,:)\beta\beta\hat{\Psi}_T(1,:))(zI + \hat{\Psi}_{T,1})^{-1}]) \\
&= \operatorname{tr}(\Psi_{1,1} E[(zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_{T,1}\beta_1\beta_1\hat{\Psi}_{T,1})(zI + \hat{\Psi}_{T,1})^{-1}]) \\
&+ \operatorname{tr}(\Psi_{1,1} E[(zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_{T,1,2}\beta_2\beta_2\hat{\Psi}_{T,1,2})(zI + \hat{\Psi}_{T,1})^{-1}]) \qquad (88) \\
&\overset{prob}{\to} P^{-1} b_* \operatorname{tr}(\Psi_{1,1} E[(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}^2(zI + \hat{\Psi}_{T,1})^{-1}]) \\
&+ P^{-1} b_* \operatorname{tr} E[\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1,2}\hat{\Psi}_{T,1,2}(zI + \hat{\Psi}_{T,1})^{-1}]
\end{aligned}$$

by Lemmas 1 and 3. The same calculations as above imply that

$$P^{-1} b_* \operatorname{tr}(\Psi_{1,1} E[(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}^2(zI+\hat{\Psi}_{T,1})^{-1}]) \ \to \ \frac{c_1}{c} b_*(\psi_{*,1}(q) - 2zc_1^{-1}\xi_{1,1}(z) - z^2 c_1^{-1}\xi_{1,1}(z))\,. \qquad (89)$$

Thus, it remains to deal with the second term in (88). We have

$$\begin{aligned}
&P^{-1} b_* \operatorname{tr} E[\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1,2}\hat{\Psi}_{T,1,2}(zI + \hat{\Psi}_{T,1})^{-1}] \\
&= P^{-1} b_* \operatorname{tr} E[\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1,2}\hat{\Psi}_{T,1,2}(zI + \hat{\Psi}_{T,1})^{-1}] \\
&= P^{-1} b_* \frac{1}{T^2} \operatorname{tr} E[\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1} \sum_{t_1,t_2} S_{t_1,1} S_{t_1,2}' S_{t_2,2} S_{t_2,1}' (zI + \hat{\Psi}_{T,1})^{-1}] \\
&= P^{-1} b_* \frac{1}{T^2} \operatorname{tr} E[\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1} \Big( T S_{t_1,1} S_{t_1,2}' S_{t_1,2} S_{t_1,1}' \qquad (90) \\
&+ T(T-1)\, S_{t_1,1} S_{t_1,2}' S_{t_2,2} S_{t_2,1}' \Big) (zI + \hat{\Psi}_{T,1})^{-1}] \\
&= Term1 \ + \ Term2\,.
\end{aligned}$$

Here,

$$Term1 = P^{-1}b_*\frac{1}{T}\operatorname{tr} E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}S_{t_1,1}S'_{t_1,2}S_{t_1,2}S'_{t_1,1}(zI+\hat{\Psi}_{T,1})^{-1}] \tag{91}$$

and

$$Term2 = (1-T^{-1})P^{-1}b_*\operatorname{tr} E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}S_{t_1,1}S'_{t_1,2}S_{t_2,2}S'_{t_2,1}(zI+\hat{\Psi}_{T,1})^{-1}]\,. \tag{92}$$

Using the Sherman-Morrison formula (29) and defining $C_T = S'_{t,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S_{t,1}$, we get

$$(zI+\hat{\Psi}_{T,1})^{-1}S_{t_1,1} = (zI+\hat{\Psi}_{T,1,t})^{-1}S_{t_1,1}(1+C_T)^{-1}\,, \tag{93}$$

and therefore

$$\begin{aligned} &Term1 \\ &= P^{-1}b_*\frac{1}{T}\operatorname{tr} E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S_{t_1,1}(1+C_T)^{-1} \\ &\times S'_{t_1,2}S_{t_1,2}(1+C_T)^{-1}S'_{t_1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}] \\ &= P^{-1}b_*\frac{1}{T}\operatorname{tr} E[S'_{t_1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S_{t_1,1}S'_{t_1,2}S_{t_1,2}(1+C_T)^{-2}] \end{aligned} \tag{94}$$

Now, Lemmas 2, and 3, and the Vitali Theorem together with the fact that $S'_t$ is independent of $\hat{\Psi}_{T,1,t}$ imply that

$$\frac{1}{T}S'_{t,1}(zI+\hat{\Psi}_{T,1,t})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S_{t,1} \;\to\; \frac{1}{T}E[\operatorname{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1})] \tag{95}$$

in $L_2$, whereas

$$P^{-1}S'_{t_1,2}S_{t_1,2}(1+C_T)^{-2} \;\to\; P^{-1}\operatorname{tr}\Psi_{2,2}/(1+\xi_{1,1}(z))^2$$

in $L_2$. Therefore, Lemma 10 implies that

$$Term1 \to b_*\widehat{\xi}_{1,1}(z)P^{-1}\operatorname{tr}\Psi_{2,2}/(1+\xi_{1,1}(z))^2 \tag{96}$$

where we have defined

$$\widehat{\xi}_{1,1,T}(z) = \frac{1}{T}E[\operatorname{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1})]$$

We will now need the following lemma.

**Lemma 12** *We have*

$$\frac{1}{T}E[\operatorname{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1})] \;\to\; \widehat{\xi}_{1,1}(z) = (\xi_{1,1}(z)+z\xi'_{1,1}(z))(1+\xi_{1,1}(z))^2 \tag{97}$$

**Proof of Lemma 12.** We have

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

by (11) and therefore

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

Lemmas 2, and 3, and the Vitali Theorem imply that

$$\begin{aligned}&\frac{1}{T}S'_{t_1,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S'_{t_1,1}\\&-\frac{1}{T}\,\mathrm{tr}\,E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}]\;\to\;0\end{aligned}\tag{98}$$

is probability. In the next equation, to simplify the expressions, we will use $X_T\approx Y_T$ to denote the fact that $X_T-Y_T\to 0$ as $T\to\infty$. By (93) and (98),

$$\begin{aligned}\xi_{1,1}(z)\;&\approx\;\frac{1}{T}\,\mathrm{tr}\,E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}]\\
&=\;\frac{1}{T}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_{T,1})(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}]\\
&\approx\;-z\xi_{1,1}(z)\;+\;\frac{1}{T}\,\mathrm{tr}\,E[\hat{\Psi}_{T,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}]\\
&=\{\hat{\Psi}_{T,1}=T^{-1}\sum_t S_{t,1}S'_{t,1}\}\\
&=\;-z\xi'_{1,1}(z)\;+\;\frac{1}{T^2}\sum_t\mathrm{tr}\,E[S_{t,1}S'_{t,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}]\\
&=\;-z\xi'_{1,1}(z)\;+\;\frac{1}{T}\,\mathrm{tr}\,E[S_{t,1}S'_{t,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}]\\
&=\;-z\xi'_{1,1}(z)\;+\;\frac{1}{T}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_{T,1})^{-1}S_{t,1}S'_{t,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}]\\
&=\;-z\xi'_{1,1}(z)\\
&+\;\frac{1}{T}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_{T,1,t})^{-1}S_{t_1,1}(1+C_T)^{-1}S'_{t_1,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}]\\
&=\;-z\xi'_{1,1}(z)\\
&+\;\frac{1}{T}\,\mathrm{tr}\,E[(1+C_T)^{-2}S'_{t_1,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S_{t_1,1}]\\
&\approx\;-z\xi'_{1,1}(z)\;+\;(1+\xi_{1,1}(z))^{-2}\widehat{\xi}_{1,1}(z)\end{aligned}\tag{99}$$

and the claim follows. The proof of Lemma 12 is complete. □

Thus, it remains to deal with $Term2$ in (90). By (93),

$$
\begin{aligned}
Term2 \;&\approx\; P^{-1}b_*\,\mathrm{tr}\,E[\Psi_{1,1}(zI+\hat\Psi_{T,1})^{-1}S_{t_1,1}S'_{t_1,2}S_{t_2,2}S'_{t_2,1}(zI+\hat\Psi_{T,1})^{-1}]\\
&=\; P^{-1}b_*\,\mathrm{tr}\,E[S'_{t_2,1}(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,1}(zI+\hat\Psi_{T,1})^{-1}S_{t_1,1}S'_{t_1,2}S_{t_2,2}]\\
&\approx\; P^{-1}b_*\,\mathrm{tr}\,E[(1+C_T)^{-1}S'_{t_2,1}(zI+\hat\Psi_{T,1,t_2})^{-1}\\
&\times\Psi_{1,1}(zI+\hat\Psi_{T,1,t_1})^{-1}S_{t_1,1}(1+C_T)^{-1}S'_{t_1,2}S_{t_2,2}]\\
&\approx\; P^{-1}b_*\,\mathrm{tr}\,E[(1+C_T)^{-1}S'_{t_2,1}\\
&\times\left((zI+\hat\Psi_{T,1,t_1,t_2})^{-1}-\frac{1}{T}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S_{t_1,1}(1+C_T)^{-1}S'_{t_1,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\right)\\
&\times\Psi_{1,1}\left((zI+\hat\Psi_{T,1,t_1,t_2})^{-1}-\frac{1}{T}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S_{t_2,1}(1+C_T)^{-1}S'_{t_2,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\right)\\
&S_{t_1,1}(1+C_T)^{-1}S'_{t_1,2}S_{t_2,2}]\\
&=\; P^{-1}b_*\,\mathrm{tr}\,E[(1+C_T)^{-1}S'_{t_2,1}\\
&\times\left((zI+\hat\Psi_{T,1,t_1,t_2})^{-1}-\frac{1}{T}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S_{t_1,1}(1+C_T)^{-1}S'_{t_1,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\right)\\
&\times\Psi_{1,1}\left((zI+\hat\Psi_{T,1,t_1,t_2})^{-1}-\frac{1}{T}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S_{t_2,1}(1+C_T)^{-1}S'_{t_2,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\right)\\
&S_{t_1,1}(1+C_T)^{-1}S'_{t_1,2}S_{t_2,2}]\\
&=\; P^{-1}b_*\,\mathrm{tr}\,E[(1+C_T)^{-1}S'_{t_2,1}\\
&\times\Bigg((zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\\
&-\frac{1}{T}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S_{t_1,1}(1+C_T)^{-1}S'_{t_1,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\\
&-(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\frac{1}{T}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S_{t_2,1}(1+C_T)^{-1}S'_{t_2,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\\
&+\frac{1}{T}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S_{t_1,1}(1+C_T)^{-1}S'_{t_1,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\\
&\times\frac{1}{T}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}S_{t_2,1}(1+C_T)^{-1}S'_{t_2,1}(zI+\hat\Psi_{T,1,t_1,t_2})^{-1}\Bigg)S_{t_1,1}(1+C_T)^{-1}S'_{t_1,2}S_{t_2,2}]\\
&=\; Term21+Term22+Term23+Term24\,.
\end{aligned}
\tag{100}
$$

Note that the different $1+C_T$ factors differ from each other slightly, but we will abuse the notation and treat them as identical. Dealing with them separately requires minor modifications in the proofs. By direct calculation,

$$
E[S'_{t_2,1}QS_{t_2,2}|S_{t_2}]=\mathrm{tr}(Q\Psi_{2,1}) \tag{101}
$$

for any $Q$ independent of $S_{t_2}$. Thus,

$$\begin{aligned} Term21 &= P^{-1} b_* \operatorname{tr} E[(1+C_T)^{-1} S'_{t_2,1} \\ &\times \left( (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \Psi_{1,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \right) S_{t_1,1} (1+C_T)^{-1} S'_{t_1,2} S_{t_2,2}] \\ &= P^{-1} b_* \operatorname{tr} E[(1+C_T)^{-2} S'_{t_2,1} Q S_{t_2,2}] \\ &= b_* E[(1+C_T)^{-2} P^{-1} \operatorname{tr}(Q \Psi_{2,1})] \,, \end{aligned} \tag{102}$$

where we have defined

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

By a modification of Lemmas 2 and 3, we get

$$\begin{aligned} P^{-1} \operatorname{tr}(Q \Psi_{2,1}) &= P^{-1} \operatorname{tr} \left( (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \Psi_{1,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_1,1} S'_{t_1,2} \Psi_{2,1} \right) \\ &= P^{-1} \operatorname{tr} \left( S'_{t_1,2} \Psi_{2,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \Psi_{1,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_1,1} \right) \\ &\overset{prob}{\to} P^{-1} \operatorname{tr} E[\Psi_{1,2} \Psi_{2,1} \left( (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \Psi_{1,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \right)] \,, \end{aligned} \tag{103}$$

where, as we explain in the main text, we pass to a subsequence if necessary to ensure the limit exists. Thus, by (11),

$$Term21 \to b_* (1 + \xi(z))^{-2} c_1^{-1} \widehat{\xi}_{2,1} \,. \tag{104}$$

Proceeding to the next term in (100), we get

$$\begin{aligned} &Term22 \\ &= P^{-1} b_* \operatorname{tr} E[(1+C_T)^{-1} S'_{t_2,1} \\ &\times \left( -\frac{1}{T} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_1,1} (1+C_T)^{-1} S'_{t_1,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \Psi_{1,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \right) \\ &S_{t_1,1} (1+C_T)^{-1} S'_{t_1,2} S_{t_2,2}] \end{aligned} \tag{105}$$

We have

$$\begin{aligned}
&\frac{1}{T}S'_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1} \\
&\to \ \frac{1}{T}\operatorname{tr} E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}] \\
&= \ \widehat{\xi}_{1,1}(z)
\end{aligned} \tag{106}$$

is probability by Lemmas 2 and 3 and the Vitali Theorem. Hence,

$$\begin{aligned}
&Term22 \\
&\to \ P^{-1}b_* \operatorname{tr} E[(1+C_T)^{-1}S'_{t_2,1} \\
&\times \left( -(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}(1+C_T)^{-1}\widehat{\xi}_{1,1}(z)\right)(1+C_T)^{-1}S_{t_1,2}S'_{t_2,2}] \\
&\to \ -b_*\widehat{\xi}_{1,1}(z)(1+\xi_{1,1}(z))^{-3}\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}S'_{t_1,2}] \\
&\to \ -b_*\widehat{\xi}_{1,1}(z)(1+\xi_{1,1}(z))^{-3}c_1^{-1}\xi_{2,1}(z)\,,
\end{aligned} \tag{107}$$

where we have used Lemma 10 to pass to the limit.[46] Proceeding to the next term in (100), we get

$$\begin{aligned}
&Term23 \ \approx \ P^{-1}b_*E[(1+C_T)^{-1}S'_{t_2,1} \\
&\left( -(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_2,1}(1+C_T)^{-1}S'_{t_2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right) \\
&S_{t_1,1}(1+C_T)^{-1}S'_{t_1,2}S_{t_2,2}] \\
&= -b_*E[X_TY_T]
\end{aligned} \tag{108}$$

where we have defined

$$X_T = -(1+C_T)^{-1}S'_{t_2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_2,1}$$

and

$$Y_T = P^{-1}(1+C_T)^{-2}S'_{t_1,2}S_{t_2,2}S'_{t_2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}\,.$$

By Lemma 12 and (11), $X_T \to (1+\xi_{1,1}(z))^{-1}\widehat{\xi}(z)$ in $L_2$, whereas $Y_T$ has a bounded $L_2$-norm. Then, a small modification of Lemma 10 implies that

$$E[X_TY_T] \ - \ (1+\xi_{1,1}(z))^{-1}\widehat{\xi}(z)E[Y_T] \ \to \ 0$$

[46] Note that it may seem that we need six bounded moments for the signals. But, in fact, the normalization by $1+C_T$ ensures all the necessary terms stay bounded.

Integrating over $S_{t_2}$ gives

$$E[Y_T] = E[P^{-1}(1+C_T)^{-2}S'_{t_1,2}\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}]$$

and Lemmas 2 and 3 imply that

$$E[Y_T] \;\to\; c_1^{-1}(1+\xi_{1,1}(z))^{-2}\xi_{2,1}(z)\,.$$

Thus, $Term23$ in (100) satisfies.

$$Term23 \;\to\; -b_*\widehat{\xi}_{1,1}(z)(1+\xi_{1,1}(z))^{-3}c_1^{-1}\xi_{2,1}(z)\,. \tag{109}$$

Finally, the last term in (100) is given by

$$\begin{aligned}
&Term24 = P^{-1}b_*\,\mathrm{tr}\,E[(1+C_T)^{-1}S'_{t_2,1}\\
&\times\left(\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}(1+C_T)^{-1}S'_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\right.\\
&\left.\times\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_2,1}(1+C_T)^{-1}S'_{t_2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)\\
&\times S_{t_1,1}(1+C_T)^{-1}S'_{t_1,2}S_{t_2,2}]\\
&=\; P^{-1}b_*\,\mathrm{tr}\,E[(1+C_T)^{-4}S_{t_2,2}S'_{t_2,1}\\
&\times\left(\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}S'_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\right.\\
&\left.\times\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_2,1}S'_{t_2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)\\
&S_{t_1,1}S'_{t_1,2}]
\end{aligned} \tag{110}$$

We will need the following lemma.

**Lemma 13** *Consider the block matrix decomposition*

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

*Then,*

$$\begin{aligned}
&E[S_{t,2}S'_{t,1}ZS_{t,1}S'_{t,1}]\\
&=\; \Psi_{2,1}(Z+Z)\Psi_{1,1} + (Q_2\,\mathrm{diag}((E[X^4]-3)Q_1ZQ_1)Q_2 + \mathrm{tr}(Z\Psi_{1,1})\Psi_{2,1})
\end{aligned} \tag{111}$$

*for any matrix $Z$. If $Z$ is uniformly bounded, then the matrices $Q_2\,\mathrm{diag}(Q_1ZQ_1)Q_2$ have uniformly bounded trace norms.*

**Proof of Lemma 13**. By linearity, it suffices to prove the formula for a rank-one matrix $A = \beta\gamma'$. Then, $S'_t = X'_t\Psi^{1/2}$ and we will decompose $\Psi^{1/2}$ into $(Q_1, Q_2)$, so that $S'_{t,k} = X'_tQ_k$.

Then,

$$E[S_{t,2}S'_{t,1}\beta\gamma' S_{t,1}S'_{t,1}] = E[Q_2\, X_t X'_t Q_1 \beta\gamma' Q'_1 X_t X'_t Q_1] \tag{112}$$

Define $\tilde{\beta} = Q_1\beta,\ \tilde{\gamma} = Q_1\gamma$. Then, if $k_1 \neq k_2$, we have

$$\begin{aligned} E[X_t X'_t \tilde{\beta}\tilde{\gamma}' X_t X'_t]_{k_1,k_2} &= E[\sum_{l_1,l_2} X_{k_1} X_{l_1} \tilde{\beta}_{l_1}\tilde{\gamma}_{l_2} X_{l_2} X_{k_2}] \\ &= E[X_{k_1}^2 X_{k_2}^2](\tilde{\beta}_{k_1}\tilde{\gamma}_{k_2} + \tilde{\beta}_{k_2}\tilde{\gamma}_{k_1}) + \sum_{\ell} \tilde{\beta}_\ell\tilde{\gamma}_\ell E[X_{k_1} X_{k_2} X_\ell^2] \\ &= \tilde{\beta}_{k_1}\tilde{\gamma}_{k_2} + \tilde{\beta}_{k_2}\tilde{\gamma}_{k_1} \end{aligned} \tag{113}$$

At the same time,

$$\begin{aligned} E[X_t X'_t \tilde{\beta}\tilde{\gamma}' X_t X'_t]_{k_1,k_1} &= E[\sum_{l_1,l_2} X_{k_1}^2 X_{l_1} \tilde{\beta}_{l_1}\tilde{\gamma}_{l_2} X_{l_2}] \\ &= \sum_{\ell} \tilde{\beta}_\ell\tilde{\gamma}_\ell E[X_{k_1}^2 X_\ell^2] \\ &= \tilde{\beta}_{k_1}\tilde{\gamma}_{k_1}(E[X_{k_1}^4] - 1) \ + \ \tilde{\beta}'\tilde{\gamma} \end{aligned} \tag{114}$$

Summarizing,

$$E[X_t X'_t \tilde{\beta}\tilde{\gamma}' X_t X'_t] = \tilde{\beta}'\tilde{\gamma} I \ + \ \tilde{\beta}\tilde{\gamma}' + \tilde{\gamma}\tilde{\beta}' + \mathrm{diag}(\tilde{\beta}\tilde{\gamma}(E[X^4] - 3))$$

Thus, by formula (112), we get

$$E[S'_{t,2}S_{t,1}\beta\gamma' S'_{t,1}S_{t,1}] = Q'_2Q_1(\beta\gamma' + \gamma\beta')Q'_1Q_1 + (Q'_2\,\mathrm{diag}((E[X^4]-3)\tilde{\beta}_{k_1}\tilde{\gamma}_{k_1})Q_2 + (\tilde{\beta}'\tilde{\gamma})Q'_2Q_1)\,, \tag{115}$$

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

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

Thus,

$$\Psi = \begin{pmatrix} Q'_1Q_1 & Q'_1Q_2 \\ Q'_2Q_1 & Q'_2Q_2 \end{pmatrix} = \begin{pmatrix} \Psi_{1,1} & \Psi_{1,2} \\ \Psi_{2,1} & \Psi_{2,2} \end{pmatrix} \tag{116}$$

and hence we get the required.

□

Since the kurtosis terms have uniformly bounded trace norms, it is straightforward to show that their contributions to asymptotic expectations get annihilated by $1/T$ and $1/P$ factors. Hence, from now on, we will be assuming in our calculations that $E[X_{i,t}^4] = 3$.

Applying Lemma 13, we can integrate over $S_{t_2}$[47]. Define

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

Then, we can rewrite (110) as

$$\begin{aligned} Term24 &= P^{-1} b_* \operatorname{tr} E[(1 + C_T)^{-4} S_{t_2,2} S'_{t_2,1} \\ &\times \left( \frac{1}{T}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_1,1} S'_{t_1,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \Psi_{1,1} \right. \\ &\left. \times \frac{1}{T}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_2,1} S'_{t_2,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \right) \\ &S_{t_1,1} S'_{t_1,2}] \\ &= P^{-1} b_* \operatorname{tr} E[(1 + C_T)^{-4} E[S_{t_2,2} S'_{t_2,1} Z S_{t_2,1} S'_{t_2,1} | S_{t_1}] (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_1,1} S'_{t_1,2}] \end{aligned} \tag{117}$$

Applying Lemma 13, we get

$$E[S_{t_2,2} S'_{t_2,1} Z S_{t_2,1} S'_{t_2,1} | S_{t_1}] = \Psi_{2,1}(Z + Z')\Psi_{1,1} + \operatorname{tr}(Z\Psi_{1,1})\Psi_{2,1})$$

Substituting this expression into (117), we get that everything reduces to computing two expectations:[48]

$$Expectation1 = P^{-1} \operatorname{tr} E[\Psi_{2,1} Z \Psi_{1,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_1,1} S'_{t_1,2}] \tag{118}$$

and

$$Expectation2 = P^{-1} \operatorname{tr} E[\Psi_{2,1} \operatorname{tr}(Z\Psi_{1,1}) (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_1,1} S'_{t_1,2}] \tag{119}$$

For $Expectation2$, we have

$$\begin{aligned} Expectation2 &= P^{-1} \operatorname{tr} E[\Psi_{2,1} \operatorname{tr}(Z\Psi_{1,1}) (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_1,1} S'_{t_1,2}] \\ &= P^{-1} \operatorname{tr} E[S'_{t_1,2} \Psi_{2,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_1,1} \operatorname{tr}(Z\Psi_{1,1})] . \end{aligned} \tag{120}$$

We know that the quantities

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

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

[48] Computing Expectation1 with $Z'$ instead of $Z$ is similar.

and

$$\begin{aligned}
&\frac{1}{T}\operatorname{tr}\left((zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}S'_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)\\
&= \frac{1}{T}\operatorname{tr}\left(S'_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}\right)
\end{aligned} \tag{121}$$

both converge to finite numbers in $L_2$ by Lemmas 2 and 3. Thus, when multiplied by $P^{-1}$, the expectation of the product of these two quantities converges to zero. Thus, $Expectation2$ converges to zero. To compute $Expectation1$, we use

$$\begin{aligned}
&Expectation1 = P^{-1}\operatorname{tr}E[\Psi_{2,1}Z\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}S'_{t_1,2}]\\
&= P^{-1}\operatorname{tr}E[S_{t_1,2}S'_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}Z'\Psi_{1,2}]\\
&= P^{-1}\operatorname{tr}E[S_{t_1,2}S'_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\\
&\times\left((zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}S'_{t_1,1}\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)\Psi_{1,2}]
\end{aligned} \tag{122}$$

We can now once again apply Lemma 13 and get

$$\begin{aligned}
&E[S'_{t_1,2}S_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\\
&\times(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S'_{t_1,1}S_{t_1,1}]\\
&= \Psi_{2,1}(\hat{Z}+\hat{Z}')\Psi_{1,1} \;+\; \operatorname{tr}(\hat{Z}\Psi_{1,1})\Psi_{2,1}
\end{aligned} \tag{123}$$

where

$$\hat{Z} = (zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1} \tag{124}$$

Therefore,

$$Expectation1 = P^{-1}\operatorname{tr}E[\left(\Psi_{2,1}(\hat{Z}+\hat{Z}')\Psi_{1,1} \;+\; \operatorname{tr}(\hat{Z}\Psi_{1,1})\Psi_{2,1}\right)\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,2}] \tag{125}$$

First, by Lemma 3 and the Vitali Theorem, $\operatorname{tr}(\hat{Z}\Psi_{1,1})$ converges to a finite, non-random number, and hence the second term in this expression converges to zero. Second, the first term also converges to zero by a similar argument, due to the $P^{-1}(T)^{-2}$ factor. Thus,

$Term24$ converges to zero. Gathering the terms, we get

$$
\begin{aligned}
&\operatorname{tr}(\Psi_{1,1}E[\hat{\beta}\hat{\beta}'])\\
&= \operatorname{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T\beta\beta'\hat{\Psi}_T \;+\; q_Tq_T')(zI+\hat{\Psi}_{T,1})^{-1}])\\
&= \operatorname{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_T(1,:)\beta\beta'\hat{\Psi}_T(1,:)'(zI+\hat{\Psi}_{T,1})^{-1}]) \;+\; \xi_{1,1}(z)+z\xi_{1,1}'(z)\\
&\overset{prob}{\to} \; P^{-1}b_*\operatorname{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}^2(zI+\hat{\Psi}_{T,1})^{-1}])\\
&+ \; P^{-1}b_*\operatorname{tr}E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2,1}\hat{\Psi}_{T,2,1}'(zI+\hat{\Psi}_{T,1})^{-1}] + \; \xi_{1,1}(z)+z\xi_{1,1}'(z)\\
&= \; \frac{c_1}{c}b_*(\psi_{*,1}(q)-2zc_1^{-1}\xi_{1,1}(z)-z^2c_1^{-1}\xi_{1,1}'(z)) \;+\; \xi_{1,1}(z)+z\xi_{1,1}'(z)\\
&+ \; P^{-1}b_*\operatorname{tr}E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2,1}\hat{\Psi}_{T,2,1}'(zI+\hat{\Psi}_{T,1})^{-1}]\\
&= \; \frac{c_1}{c}b_*(\psi_{*,1}(q)-2zc_1^{-1}\xi_{1,1}(z)-z^2c_1^{-1}\xi_{1,1}'(z)) \;+\; (1+b_*P^{-1}\operatorname{tr}\Psi_{2,2})(\xi_{1,1}(z)+z\xi_{1,1}'(z))\\
&+ Term21 + Term22 + Term23 + Term24\\
&\to \; \frac{c_1}{c}b_*(\psi_{*,1}(q)-2zc_1^{-1}\xi_{1,1}(z)-z^2c_1^{-1}\xi_{1,1}'(z)) \;+\; (1+b_*P^{-1}\operatorname{tr}\Psi_{2,2})(\xi_{1,1}(z)+z\xi_{1,1}'(z))\\
&+ b_*(1+\xi(z))^{-2}c_1^{-1}\widehat{\xi}_{2,1} - 2b_*(\xi_{1,1}(z)+z\xi_{1,1}'(z))(1+\xi_{1,1}(z))^{-1}c_1^{-1}\xi_{2,1}(z)
\end{aligned}
\tag{126}
$$

The proof of Lemma 11 is complete.

□

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

We will need the following lemma.

**Lemma 14** *Suppose that $S_t = X_t\Psi^{1/2}$. Then, under the decomposition $\Psi^{1/2} = (Q_1, Q_2)$,*

$$
\begin{aligned}
&E[S_{t,2}'S_{t,1}ZS_{t,1}'S_{t,2}]\\
&= \; \Psi_{2,1}(Z+Z')\Psi_{1,2}\\
&+ ((\kappa-2)Q_2'\operatorname{diag}(Q_1ZQ_1')Q_2 + \operatorname{tr}(Z\Psi_{1,1})\Psi_{2,2})
\end{aligned}
\tag{127}
$$

*for any matrix $Z$.*

**Proof of Lemma 14**. By linearity, it suffices to prove the formula for a rank-one matrix $A = \beta\gamma'$. Then, $S_t = X_t\Psi^{1/2}$ and we will decompose $\Psi^{1/2}$ into $(Q_1, Q_2)$, so that $S_{t,k} = \Sigma^{1/2}X_tQ_k$. Then,

$$
E[S_{t,2}'S_{t,1}\beta\gamma'S_{t,1}'S_{t,2}] \;=\; E[Q_2'\,X_t'X_tQ_1\beta\gamma'Q_1'X_t'X_tQ_2] \tag{128}
$$

Define $\tilde{\beta} = Q_1\beta$. Then,

$$\begin{aligned} &E[S'_{t,2}S_{t,1}\beta\gamma' S'_{t,1}S_{t,2}] = \\ &(Q'_2Q_1\beta\gamma' Q'_1Q_2 + Q'_2Q_1\gamma\beta' Q'_1Q_2) \\ &+ ((\kappa-2)Q'_2\,\mathrm{diag}(\tilde{\beta}_{k_1}\tilde{\gamma}_{k_1})Q_2 + (\tilde{\beta}'\tilde{\gamma})Q'_2Q_2) \end{aligned} \tag{129}$$

whereas $\tilde{\beta}'\tilde{\gamma} = \beta' Q'_1Q_1\gamma$. Now,

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

Thus,

$$\Psi \;=\; \begin{pmatrix} Q'_1Q_1 & Q'_1Q_2 \\ Q'_2Q_1 & Q'_2Q_2 \end{pmatrix} \;=\; \begin{pmatrix} \Psi_{1,1} & \Psi_{1,2} \\ \Psi_{2,1} & \Psi_{2,2} \end{pmatrix} \tag{130}$$

and hence we get the required.

□

As above, all expectations in this section are conditional on $\hat{\beta}$. We have since $\beta_1\beta'_2 \to 0$ in probability that

$$\begin{aligned} &E[(R^{\pi}_{t+1})^2] \;=\; E[\hat{\beta}'_1 S'_{t,1}R_{t+1}R'_{t+1}S_{t,1}\hat{\beta}_1] \;=\; E[\hat{\beta}'_1 S'_{t,1}((S_{t,1}\beta_1 + S_{t,2}\beta_2)(S_{t,1}\beta_1 + S_{t,2}\beta_2)' + I)S_{t,1}\hat{\beta}_1] \\ &\to\; E[\hat{\beta}'_1 S'_{t,1}R_{t+1}R'_{t+1}S_{t,1}\hat{\beta}_1] \;=\; E[\hat{\beta}'_1 S'_{t,1}(S_{t,1}\beta_1\beta'_1 S'_{t,1} + S_{t,2}\beta_2\beta'_2 S'_{t,2} + I)S_{t,1}\hat{\beta}_1] \\ &\to\; E[\hat{\beta}'_1\Psi_{1,1}\hat{\beta}_1] \;+\; E[\hat{\beta}'_1 S'_{t,1}S_{t,1}\beta_1\beta'_1 S'_{t,1}S_{t,1}\hat{\beta}_1] + \; P^{-1}b_* E[\hat{\beta}'_1 S'_{t,1}S_{t,2}S'_{t,2}S_{t,1}\hat{\beta}_1] \\ &=\; \mathcal{L} \;+\; Term2 \;+\; Term3\,. \end{aligned} \tag{131}$$

Recall that we are using the notation

$$\hat{\Psi}_T \;=\; \frac{1}{T}\sum_{t=1}^{T} S'_{t,1}S_t \;\in \mathbb{R}^{P_1\times P}\,. \tag{132}$$

and we decompose

$$\hat{\Psi}_T \;=\; \hat{\Psi}_{T,1} \;+\; \hat{\Psi}_{T,2} \tag{133}$$

### C.0.1 Term3

We have

$$\begin{aligned} Term3 &= P^{-1}b_* E[\hat\beta_1' S_{t,1}' S_{t,2} S_{t,2}' S_{t,1} \hat\beta_1] \\ &= P^{-1}b_* \operatorname{tr} E[(\hat\Psi_T\beta + q_T)'(zI+\hat\Psi_{T,1})^{-1} S_{t,1}' S_{t,2} S_{t,2}' S_{t,1} (zI+\hat\Psi_{T,1})^{-1}(\hat\Psi_T\beta + q_T)] \\ &= P^{-1}b_* \operatorname{tr} E[S_{t,2}' S_{t,1}(zI+\hat\Psi_{T,1})^{-1}(\hat\Psi_T\beta + q_T)(\hat\Psi_T\beta + q_T)'(zI+\hat\Psi_{T,1})^{-1} S_{t,1}' S_{t,2}] \\ &\to P^{-1}b_* \operatorname{tr} E[S_{t,2}' S_{t,1}(zI+\hat\Psi_{T,1})^{-1}(\hat\Psi_T\beta\beta'\hat\Psi_T + T^{-1}\hat\Psi_{T,1})(zI+\hat\Psi_{T,1})^{-1} S_{t,1}' S_{t,2}] \end{aligned} \tag{134}$$

in probability because

$$E[q_T q_T' | S] = \frac{1}{T}\hat\Psi_{T,1}\,. \tag{135}$$

By Lemma 14, we get

$$\begin{aligned} Term3 &\to P^{-1} 2\operatorname{tr}(Z_T \Psi_{1,2}\Psi_{2,1}) \\ &+ P^{-1}\operatorname{tr}((\kappa-2)Q_2' \operatorname{diag}(Q_1 Z_T Q_1') Q_2) + P^{-1}\operatorname{tr}(Z_T\Psi_{1,1})\operatorname{tr}(\Psi_{2,2}) \end{aligned} \tag{136}$$

with

$$Z_T = (zI+\hat\Psi_{T,1})^{-1}(\hat\Psi_T\beta\beta'\hat\Psi_T' + T^{-1}\hat\Psi_{T,1})(zI+\hat\Psi_{T,1})^{-1}\,. \tag{137}$$

Thus,

$$\begin{aligned} \operatorname{tr}(Z_T\Psi_{1,1}) &= \operatorname{tr}((zI+\hat\Psi_{T,1})^{-1}(\hat\Psi_T\beta\beta'\hat\Psi_T' + T^{-1}\hat\Psi_{T,1})(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,1}) \\ &= \mathcal{L} \end{aligned} \tag{138}$$

by (86). At the same time,

$$\begin{aligned} P^{-1}\operatorname{tr}(Z_T\Psi_{1,2}\Psi_{2,1}) &\to b_* P^{-2}\operatorname{tr}(\hat\Psi_T'(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,2}\Psi_{2,1}(zI+\hat\Psi_{T,1})^{-1}\hat\Psi_T) \\ &+ P^{-1}T^{-1}\operatorname{tr}((zI+\hat\Psi_{T,1})^{-1}(\hat\Psi_{T,1}+zI-zI)(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,2}\Psi_{2,1}) \\ &\to P^{-1}b_* c^{-1}\widetilde{\xi}_{2,1}(z) + P^{-1}(\xi_{2,1}(z) - z\xi_{2,1}'(z)) \to 0\,. \end{aligned} \tag{139}$$

where

$$\widetilde{\xi}_{2,1}(z) = \operatorname{tr}(\hat\Psi_T'(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,2}\Psi_{2,1}(zI+\hat\Psi_{T,1})^{-1}\hat\Psi_T) = O(P)\,.$$

Similarly,

$$\begin{aligned} &P^{-1}\operatorname{tr}((\kappa-2)Q_2'\operatorname{diag}(Q_1((zI+\hat\Psi_{T,1})^{-1}(P^{-1}\hat\Psi_T\hat\Psi_T' + T^{-1}\hat\Psi_{T,1})(zI+\hat\Psi_{T,1})^{-1})Q_1')Q_2) \\ &\le K P^{-1}\operatorname{tr}(Z_T) \to 0 \end{aligned} \tag{140}$$

Thus, we get

$$Term3 \to b_*\operatorname{tr}(\Psi_{2,2})P^{-1}\mathcal{L} \tag{141}$$

### C.0.2 Term2

By a slight modification of Lemma 14,

$$\begin{aligned}
&E[S'_{t,1}S_{t,1}\beta_1\beta'_1S'_{t,1}S_{t,1}] \\
&= 2\Psi_{1,1}\beta_1\beta'_1\Psi_{1,1} + ((\kappa-2)\Psi_{1,1}^{1/2}\operatorname{diag}(\Psi_{1,1}^{1/2}\beta_1\beta'_1\Psi_{1,1}^{1/2})\Psi_{1,1}^{1/2} + \operatorname{tr}(\beta_1\beta'_1\Psi_{1,1})\Psi_{1,1}) \\
&\approx 2\Psi_{1,1}\beta_1\beta'_1\Psi_{1,1} + ((\kappa-2)\Psi_{1,1}^{1/2}\operatorname{diag}(\Psi_{1,1}^{1/2}\beta_1\beta'_1\Psi_{1,1}^{1/2})\Psi_{1,1}^{1/2} + b_*c^{-1}c_1\psi_{*,1}(q)\Psi_{1,1})
\end{aligned} \tag{142}$$

and therefore

$$\begin{aligned}
&E[\hat\beta'_1S'_{t,1}S_{t,1}\beta_1\beta'_1S'_{t,1}S_{t,1}\hat\beta_1] \\
&= \operatorname{tr} E[2\Psi_{1,1}\beta_1\beta'_1\Psi_{1,1} \\
&+ ((\kappa-2)\Psi_{1,1}^{1/2}\operatorname{diag}(\Psi_{1,1}^{1/2}\beta_1\beta'_1\Psi_{1,1}^{1/2})\Psi_{1,1}^{1/2} + b_*c^{-1}c_1\psi_{*,1}(q)\Psi_{1,1})\hat\beta_1\hat\beta'_1] \\
&= \operatorname{tr} E[\Big(2\Psi_{1,1}\beta_1\beta'_1\Psi_{1,1} \\
&+ ((\kappa-2)\Psi_{1,1}^{1/2}\operatorname{diag}(\Psi_{1,1}^{1/2}\beta_1\beta'_1\Psi_{1,1}^{1/2})\Psi_{1,1}^{1/2} + b_*c^{-1}c_1\psi_{*,1}(q)\Psi_{1,1})\Big) \\
&\times (zI+\hat\Psi_{T,1})^{-1}(\hat\Psi_T\beta + q_T)(\hat\Psi_T\beta + q_T)'(zI+\hat\Psi_{T,1})^{-1}] \\
&\to \operatorname{tr} E[\Big(2\Psi_{1,1}\beta_1\beta'_1\Psi_{1,1} + b_*c^{-1}c_1\psi_{*,1}(q)\Psi_{1,1}\Big) \\
&\times (zI+\hat\Psi_{T,1})^{-1}(\hat\Psi_{T,1}\beta_1\beta'_1\hat\Psi_{T,1} + \hat\Psi_{T,2}\beta_2\beta'_2\hat\Psi'_{T,2} + T^{-1}\hat\Psi_{T,1})(zI+\hat\Psi_{T,1})^{-1}] \\
&= 2T1 + 2T2 + 2T3 \\
&+ b_*c^{-1}c_1\psi_{*,1}(q)T4 + b_*c^{-1}c_1\psi_{*,1}(q)T5 + b_*c^{-1}c_1\psi_{*,1}(q)T6
\end{aligned} \tag{143}$$

where we have used the fact that the cross-terms involving $\beta_1\beta'_2$ converge to zero and that the kurtosis terms (those involving $\kappa-2$) also converge to zero.[49]

We now analyze each term separately. The first term in (143) gives

$$\begin{aligned}
T1 &= \operatorname{tr} E[(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,1}\beta_1\beta'_1\Psi_{1,1}(zI+\hat\Psi_{T,1})^{-1}\hat\Psi_{T,1}\beta\beta'\hat\Psi_{T,1}] \\
&= E[\beta'_1\hat\Psi_{T,1}(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,1}\beta_1\beta'_1\Psi(zI+\hat\Psi_{T,1})^{-1}\hat\Psi_{T,1}\beta_1] \\
&= E[(\beta'_1\hat\Psi_{T,1}(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,1}i\beta_1)^2] \to (b_*c^{-1}c_1(\psi_{*,1}(q) - zc_1^{-1}\xi_{1,1}(z)))^2
\end{aligned} \tag{145}$$

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

$$\begin{aligned}
&\operatorname{tr} E[\beta'_1\hat\Psi_{T,1}(zI+\hat\Psi_{T,1})^{-1}\Psi_{1,1}^{1/2}\operatorname{diag}(\Psi_{1,1}^{1/2}\beta_1\beta'_1\Psi_{1,1}^{1/2})\Psi_{1,1}^{1/2}\Psi(zI+\hat\Psi_{T,1})^{-1}\hat\Psi_{T,1}\beta_1] \\
&= E[\tilde\beta'_1A\operatorname{diag}(\tilde\beta_1\tilde\beta'_1)A'\tilde\beta_1] \\
&= E[\sum_{i,j,k}\tilde\beta_iA_{i,j}\tilde\beta_j^2A_{k,j}\tilde\beta_k] = E[\sum_{i,j}\tilde\beta_i^2A_{i,j}^2\tilde\beta_j^2] \approx P^{-2}\operatorname{tr} E[AA'] \to 0
\end{aligned} \tag{144}$$

because $A$ is bounded.

in probability by Proposition 2 because all variables are uniformly bounded because $\|\beta\|$ stays bounded almost surely. Namely, first we notice that

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

is probability. And second,

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

almost surely by Proposition 2.

Then,

$$\begin{aligned} T2 &= \operatorname{tr} E[(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}\beta_1\beta_1'\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\beta_2\beta_2'\hat{\Psi}_{T,2}'] \\ &= P^{-1}b_*E[\beta_1'\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\hat{\Psi}_{T,2}'(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}\beta_1] \\ &\to P^{-2}b_*^2\operatorname{tr} E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\hat{\Psi}_{T,2}'(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}] \\ &\le \|\Psi_{1,1}\|P^{-2}b_*^2\operatorname{tr} E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\hat{\Psi}_{T,2}'(zI+\hat{\Psi}_{T,1})^{-1}] \;\to\; 0 \end{aligned} \tag{146}$$

by the proof of Lemma 11. Then,

$$\begin{aligned} T3 &= \operatorname{tr} E[(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}\beta_1\beta_1'\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\frac{1}{T}\hat{\Psi}_{T,1}] \\ &= E[\beta_1'\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\frac{1}{T}\hat{\Psi}_{T,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}\beta_1] \\ &\le b_*K/T \;\to\; 0 \end{aligned} \tag{147}$$

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

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

is the sense of positive semi-definite order.

Then,

$$\begin{aligned} T4 &= \operatorname{tr} E[(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}\beta_1\beta_1'\hat{\Psi}_{T,1}] \\ &= E[\beta'\hat{\Psi}_{T,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}\beta] \end{aligned} \tag{148}$$

where

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

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

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

Using the identity

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

we get

$$\begin{aligned}&\frac{1}{P}\operatorname{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-2}((\hat{\Psi}_{T,1}+zI)^2 - 2z(\hat{\Psi}_{T,1}+zI)+z^2))\\ &=\; \frac{1}{P}\operatorname{tr}(\Psi_{1,1}\; - 2z\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1} + z^2\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-2})\,.\end{aligned} \tag{149}$$

By Proposition 2,

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

Then, standard arguments for analytic functions ($\xi_{1,1}(z)$ is analytic for $z$ for $\Re z > 0$)[50] imply that

$$\frac{1}{P_1}\operatorname{tr}((zI+\hat{\Psi}_{T,1})^{-2}\Psi_{1,1}) \;\to\; -c^{-1}\xi'_{1,1}(z)\,. \tag{150}$$

Thus,

$$\begin{aligned}&T4 = \operatorname{tr} E[(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}\beta_1\beta_1'\hat{\Psi}_{T,1}]\\ &=\; E[\beta_1'\hat{\Psi}_{T,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}\beta_1]\\ &\sim\; b_* c^{-1}c_1 P_1^{-1}\operatorname{tr}(\Psi_{1,1}\; - 2z\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1} + z^2\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-2})\\ &=\; b_* c^{-1}c_1\psi_{*,1}(q)\;-\; b_* 2zc^{-1}c_1P_1^{-1}\operatorname{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1})\;+\; b_* c^{-1}c_1P_1^{-1}z^2\operatorname{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-2})\\ &\sim\; b_* c^{-1}c_1(\psi_{*,1}(q) - 2zc_1^{-1}\xi_{1,1}(z) - z^2c_1^{-1}\xi'_{1,1}(z))\end{aligned} \tag{151}$$

because

$$\frac{1}{P_1}\operatorname{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}) \;\to\; c_1^{-1}\xi_{1,1}(z) \tag{152}$$

implies

$$\frac{1}{P_1}\operatorname{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-2}) \;\to\; -c_1^{-1}\xi'_{1,1}(z)\,. \tag{153}$$

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

Now, since $\beta_2\beta_2' \sim P^{-1} I_{P_2 \times P_2}$, we get by the proof of Lemma 11 that

$$\begin{aligned} T5 &= \operatorname{tr} E[(zI + \hat{\Psi}_{T,1})^{-1} \Psi_{1,1} (zI + \hat{\Psi}_{T,1})^{-1} \hat{\Psi}_{T,2} \beta_2 \beta_2' \hat{\Psi}_{T,2}'] \\ &\to P^{-1} \operatorname{tr} E[(zI + \hat{\Psi}_{T,1})^{-1} \Psi_{1,1} (zI + \hat{\Psi}_{T,1})^{-1} \hat{\Psi}_{T,2} \hat{\Psi}_{T,2}'] \\ &\to \frac{\operatorname{tr}(\Psi_{2,2})}{P} b_* \widehat{\xi}_{1,1}(z) (I + \xi_{1,1}(z))^{-2} \\ &+ b_* ((I + \xi_{1,1}(z))^{-1})^2 c^{-1} \widehat{\xi}_{2,1}(z) \\ &- 2 b_* \widehat{\xi}_{1,1}(z) ((I + \xi_{1,1}(z))^{-1}) c^{-1} \xi_{2,1}(z) (I + \xi_{1,1}(z))^{-2} \end{aligned} \tag{154}$$

Finally,

$$\begin{aligned} T6 &= \frac{1}{T} \operatorname{tr} E[(zI + \hat{\Psi}_{T,1})^{-1} \Psi (zI + \hat{\Psi}_{T,1})^{-1} \hat{\Psi}_{T,1}] \\ &= \frac{1}{T} \operatorname{tr} E[(zI + \hat{\Psi}_{T,1})^{-1} \Psi (zI + \hat{\Psi}_{T,1})^{-1} (zI + \hat{\Psi}_{T,1} - zI)] \\ &\sim (\xi_{1,1}(z) + z\xi_{1,1}'(z)) \end{aligned} \tag{155}$$

where we have used that

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

Putting all the terms together, we finally get from (131) that

$$\begin{aligned} E[(R_{t+1}^{\pi})^2] &= \mathcal{L} + Term2 + Term3 \\ &= \mathcal{L} + \underbrace{\operatorname{tr}(\Psi_{2,2}) P^{-1}}_{Term3 \text{ by } (141)} \\ &+ 2T1 + 2T2 + 2T3 \\ &+ \underbrace{b_* c^{-1} c_1 \psi_{*,1}(q) T4 + b_* c^{-1} c_1 \psi_{*,1}(q) T5 + b_* c^{-1} c_1 \psi_{*,1}(q) T6}_{by \ (143)} \\ &= \mathcal{L} + \operatorname{tr}(\Psi_{2,2}) P^{-1} \mathcal{L} \\ &+ 2\Big( (b_* c^{-1} c_1 (\psi_{*,1}(q) - z c_1^{-1} \xi_{1,1}(z)))^2 \Big) \\ &+ b_* c^{-1} c_1 \psi_{*,1}(q) \Big( b_* c^{-1} c_1 (\psi_{*,1}(q) - 2z c_1^{-1} \xi_{1,1}(z) - z^2 c_1^{-1} \xi_{1,1}'(z)) \Big) \\ &+ b_* c^{-1} c_1 \psi_{*,1}(q) \Delta(z) \\ &+ b_* c^{-1} c_1 \psi_{*,1}(q) \Big( \xi_{1,1}(z) + z\xi_{1,1}'(z) \Big) \\ &= \mathcal{L}(z) + \operatorname{tr}(\Psi_{2,2}) P^{-1} \mathcal{L}(z) + 2\mathcal{E}(z)^2 \\ &+ b_* c^{-1} c_1 \psi_{*,1}(q) \mathcal{L}(z) \\ &= \mathcal{L}(z) \Big( 1 + b_* \frac{\operatorname{tr}(\Psi)}{P} \Big) + 2\mathcal{E}(z)^2 \end{aligned} \tag{156}$$

## D Proof of Theorem 7 and Optimal Shrinkage

Defining $k = 1 + b_* \frac{\mathrm{tr}(\Psi_{2,2})}{P}$, we get that

$$\begin{aligned} 2\mathcal{E} - \mathcal{L} &= 2b_* q\nu - qb_*\hat{\nu} + cq\nu' k \\ &= \lim P^{-1} \mathrm{tr}\left( (2b_* q\hat{\Psi}(zI+\hat{\Psi})^{-1} - cq(1+k)\hat{\Psi}(zI+\hat{\Psi})^{-2} - qb_*\hat{\Psi}^2(zI+\hat{\Psi})^{-2})\Psi \right) \end{aligned} \tag{157}$$

Consider the function

$$f(z) = 2b_* x(z+x)^{-1} - c(1+k)x(z+x)^{-2} - x^2(z+x)^{-2}. \tag{158}$$

for any $x > 0$. Then,

$$f'(z) = -2b_* x(z+x)^{-2} + 2c(1+k)x(z+x)^{-3} + 2b_* x^2(z+x)^{-3}. \tag{159}$$

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

$$z_* = c(1+k)/b_* \tag{160}$$

Furthermore,

$$f(z_*) = \frac{b_* x^2 + 2b_* x z_* - c(1+k)x}{(z_*+x)^2} = \frac{b_* x^2 + b_* x z_*}{(z+x)^2} = b_* \frac{x}{x+z} \tag{161}$$

implying that

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

Similarly,

$$\frac{\mathcal{E}^2(z)}{\mathcal{L}(z)} = \lim \frac{(b_* q)^2 (\mathrm{tr}(\hat{\Psi}(zI+\hat{\Psi})^{-1}\Psi))^2}{q\,\mathrm{tr}((c(1+k)\hat{\Psi} + b_*\hat{\Psi}^2)(zI+\hat{\Psi})^{-2}\Psi)} \tag{162}$$

Define

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

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

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

Then,

$$f'(z) \;=\; \Big( -2E[X(z+X)^{-1}]E[X(z+X)^{-2}]E[(aX+bX^2)(z+X)^{-2}] \\ +2E[X(z+X)^{-1}]^2E[(aX+bX^2)(z+X)^{-3}] \Big) / E[(aX+bX^2)(z+X)^{-2}]^2 \,. \tag{163}$$

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

$$E[X(z+X)^{-2}]E[(aX+bX^2)(z+X)^{-2}] \;\leq\; E[X(z+X)^{-1}]E[(aX+bX^2)(z+X)^{-3}] \,. \tag{164}$$

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

$$E[(z+X)^{-1}]E[(a+bX)(z+X)^{-1}] \;\leq\; E[(a+bX)(z+X)^{-2}] \,. \tag{165}$$

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

To prove the virtue of complexity, it remains to consider

$$\nu(z_*) \;=\; q\,\psi_{*,1} - z_* c^{-1} \xi(z_*; cq) \tag{166}$$

where

$$\xi(z, cq) \;=\; \frac{1 - zm(-z; cq)}{(cq)^{-1} - 1 + zm(-z; cq)} \;=\; -1 \;+\; \frac{(cq)^{-1}}{(cq)^{-1} - 1 + zm(-z; cq)} \,. \tag{167}$$

Theorem 8 implies

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

and, hence,

$$\tilde{m}(-z; c) = (1-c)z^{-1} \;+\; cm(-z; c) \,, \tag{168}$$

is the unique positive solution to

$$z = \int \frac{(1 - (c-1)\tilde{m}x)\, dH(x)}{\tilde{m}(1 + \tilde{m}\, x)} \tag{169}$$

Furthermore,

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

Thus, our goal is to show that

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

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

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

and hence we need

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

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

$$f'(c) \;=\; \psi_{*,1} + \frac{\tilde{m}'(c)}{\tilde{m}^2} \;\geq\; 0\,. \tag{170}$$

We have

$$0 \;=\; \int \frac{(-\tilde{m}x - (c-1)\tilde{m}'(c)x)\tilde{m}(1+\tilde{m}x) - (1-(c-1)\tilde{m}x)(\tilde{m}'(1+\tilde{m}x) + \tilde{m}\tilde{m}'x)}{\tilde{m}^2(1+\tilde{m}x)^2} dH(x) \tag{171}$$

so that

$$\tilde{m}'(c) \;=\; \frac{-\int \frac{xdH(x)}{1+\tilde{m}x}}{\int \frac{(c-1)x\tilde{m}(1+\tilde{m}x)+(1-(c-1)\tilde{m}x)(1+2\tilde{m}x)}{\tilde{m}^2(1+\tilde{m}x)^2} dH(x)}\,, \tag{172}$$

We start with the observation that the denominator in this fraction is always non-negative. Indeed, (169) implies that

$$(c-1)\int \frac{\tilde{m}xdH(x)}{\tilde{m}(1+\tilde{m}x)} \;=\; \int \frac{dH(x)}{\tilde{m}(1+\tilde{m}x)} - z\,. \tag{173}$$

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

$$(c-1)\int \frac{\tilde{m}xdH(x)}{(1+\tilde{m}x)} \;= \int \frac{dH(x)}{(1+\tilde{m}x)} - z\tilde{m} \tag{174}$$

which implies

$$\begin{aligned}(c-1)\int \frac{(\tilde{m}x)^2 dH(x)}{(1+\tilde{m}x)^2} &= (c-1)\int \frac{\tilde{m}x(-1+1+\tilde{m}x)dH(x)}{(1+\tilde{m}x)^2} \\ &= \int \frac{dH(x)}{(1+\tilde{m}x)} - z\tilde{m} - (c-1)\int \frac{\tilde{m}x dH(x)}{(1+\tilde{m}x)^2} \\ &= -z\tilde{m} + \int \frac{(1+(2-c)\tilde{m}x)dH(x)}{(1+\tilde{m}x)^2}\end{aligned} \tag{175}$$

Thus,

$$\begin{aligned}&\int \frac{(c-1)x\tilde{m}(1+\tilde{m}x)+(1-(c-1)\tilde{m}x)(1+2\tilde{m}x)}{\tilde{m}^2(1+\tilde{m}x)^2} dH(x) \\ &= \int \frac{-(c-1)(x\tilde{m})^2+1+2\tilde{m}x}{(1+\tilde{m}x)^2} dH(x) \\ &= z\tilde{m} + \int \frac{c\tilde{m}x dH(x)}{(1+\tilde{m}x)^2}\end{aligned} \tag{176}$$

so that (7) is equivalent to

$$\int x dH(x) \left( z\tilde{m} + \int \frac{c\tilde{m}x}{(1+\tilde{m}x)^2} \right) \geq \int \frac{x dH(x)}{1+\tilde{m}x} \tag{177}$$

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

We can rewrite

$$(c-1)\int \frac{\tilde{m}x dH(x)}{(1+\tilde{m}x)} = \int \frac{dH(x)}{(1+\tilde{m}x)} - z\tilde{m}\,. \tag{178}$$

as

$$c - 1 + z\tilde{m} - c\int \frac{dH(x)}{(1+\tilde{m}x)} = 0\,. \tag{179}$$

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

$$1 - c^{-1} + c^{-1} z\tilde{m} - \int (1 - \tilde{m}x + \tilde{m}^2 x^2) dH(x) \approx 0\,. \tag{180}$$

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

$$-c^{-1} + c^{-1} y c a c^{-1} + E[x](ac^{-1} + bc^{-2}) - E[x^2]a^2 c^{-2} = 0 \tag{181}$$

implying

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

and

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

so that

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

Thus,

$$\begin{aligned}
&\int xdH(x)\left(z\tilde{m} \;+\; \int \frac{c\tilde{m}x}{(1+\tilde{m}x)^2}\right) \;-\int \frac{xdH(x)}{1+\tilde{m}x} \\
&=\; E[x]\left(z(ac^{-1}+bc^{-2}) \;+\; \int c(ac^{-1}+bc^{-2})x(1-2\tilde{m}x-(\tilde{m}x)^2+4(\tilde{m}x)^2)\right) \\
&-\int xdH(x)(1-\tilde{m}x+(\tilde{m}x)^2) \;+\; O(c^{-3}) \\
&=\; E[x]\left(cy(ac^{-1}+bc^{-2}) \;+\; \int (a+bc^{-1})(x-2(ac^{-1}+bc^{-2})x^2+3(ac^{-1})^2x^3)\right) \\
&-\int dH(x)(x-(ac^{-1}+bc^{-2})x^2+(ac^{-1})^2x^3) \;+\; O(c^{-3}) \\
&=\; E[x]byc^{-1}+bE[x]^2c^{-1}-2E[x]a^2c^{-1}E[x^2]+ac^{-1}E[x^2] \;+\; O(c^{-2}) \\
&=\; E[x]E[x^2]E[x]^{-1}(E[x]+y)^{-2}yc^{-1}+E[x^2]E[x]^{-1}(E[x]+y)^{-2}E[x]^2c^{-1} \\
&-2E[x](E[x]+y)^{-2}c^{-1}E[x^2]+(E[x]+y)^{-1}c^{-1}E[x^2] \;+\; O(c^{-2}) \\
&=\; \frac{2E[x^2]y}{(E[x]+y)^2} \;>\; 0\,.
\end{aligned} \tag{182}$$

Thus complete the proof, it suffices to show that the function $-1/\tilde{m}$ is concave in $c$.

Differentiating (179) with respect to $c$, we get

$$\tilde{m}'(c) \;=\; -\int \frac{\tilde{m}x}{1+\tilde{m}x}dH(x)/(z+c\int \frac{xdH(x)}{(1+\tilde{m}x)^2})\,, \tag{183}$$

and

$$\tilde{m}''(c) \;=\; \frac{2c\int \frac{x^2(\tilde{m}')^2dH(x)}{(1+\tilde{m}x)^3}-2\int \frac{\tilde{m}'xdH(x)}{(1+\tilde{m}x)^2}}{z+c\int \frac{xdH(x)}{(1+\tilde{m}x)^2}}\,. \tag{184}$$

Our goal is to show that $-1/\tilde{m}$ is concave, which is equivalent to the inequality

$$\tilde{m}''(c)\tilde{m}(c) \;<\; 2\,(\tilde{m}')^2\,. \tag{185}$$

Substituting (183) and (184), we can rewrite the desired inequality as

$$c(-\tilde{m}')\int \frac{x^2dH(x)}{(1+\tilde{m}x)^3} \;+\; \frac{xdH(x)}{(1+\tilde{m}x)^2} < \; \frac{x}{(1+\tilde{m}x)} \tag{186}$$

which is equivalent to

$$c(-\tilde{m}')\int \frac{x^2 dH(x)}{(1+\tilde{m}x)^3} \;<\; \int \frac{\tilde{m}x^2}{(1+\tilde{m}x)^2} \tag{187}$$

From (183) we get

$$-c\tilde{m}'(c) \;\leq\; \int \frac{\tilde{m}x}{1+\tilde{m}x} dH(x) / \int \frac{x dH(x)}{(1+\tilde{m}x)^2}\,. \tag{188}$$

and hence the desired inequality (187) holds if

$$\int \frac{x}{1+\tilde{m}x} dH(x) \int \frac{x^2 dH(x)}{(1+\tilde{m}x)^3} \;<\; \int \frac{x^2}{(1+\tilde{m}x)^2} \int \frac{x dH(x)}{(1+\tilde{m}x)^2} \tag{189}$$

Changing the measure to $\frac{x}{1+\tilde{m}x} dH(x) / \int \frac{x}{1+\tilde{m}x} dH(x)$, we can rewrite it as

$$E[\frac{x}{(1+\tilde{m}x)^2}] \;\leq\; E[\frac{x}{(1+\tilde{m}x)}] E[\frac{1}{(1+\tilde{m}x)}]\,,$$

which follows because the function $\frac{x}{(1+\tilde{m}x)}$ is monotone increasing while $\frac{1}{(1+\tilde{m}x)}$ is monotone decreasing in $x$.

# E Linear Kitchen Sink

**Proposition 9** *Let*

$$\hat{\pi}_t^G(z) \;=\; G_t'(zI + \hat{\Psi}_G)^{-1} \frac{1}{T}\sum_t G_t R_{t+1} \quad \textit{with} \quad \hat{\Psi}_G \;=\; \frac{1}{T}\sum_t G_t G_t' \tag{190}$$

*be the prediction of the linear kitchen sink regression on $G_t$. Then, in the limit as $P \to \infty$, $\hat{\pi}_t = S_t'\hat{\beta}(z)$ based on random linear features (**??**) converges almost surely to $\hat{\pi}_t^G(z)$.*

Proposition 9 shows that when $\gamma$ is sufficiently small, the random feature regression is equivalent to a standard linear regression.

**Proof of Proposition 9**. In this case,

$$S_t S_t' \;=\; \Omega' G_t G_t' \Omega \tag{191}$$

and hence, defining

$$\hat{\Psi}_G \;=\; \frac{1}{T}\sum_t G_t G_t'\,, \tag{192}$$

we get that

$$\hat{\Psi} \;=\; T^{-1}\sum_t S_t S_t' \;=\; \Omega' \hat{\Psi}_G \Omega \tag{193}$$

and hence

$$\hat{\beta}(z) \;=\; (zI + \Omega'\hat{\Psi}_G\,\Omega)^{-1}\Omega' X_t,\; X_t \;=\; \frac{1}{T}\sum_t G_t R_{t+1}\,. \tag{194}$$

while the portfolio strategy is given by

$$\hat{\pi}_t(z) \;=\; \hat{\beta}(z)'\Omega' G_t \;=\; X_t'\Omega(zI + \Omega'\hat{\Psi}_G\,\Omega)^{-1}\Omega' G_t \tag{195}$$

We now make the following observation.

**Lemma 15** *When $P \to \infty$, we have with probability one*

$$\Omega(zI + \Omega'\hat{\Psi}_G\,\Omega)^{-1}\Omega' \;\to\; (zI + \hat{\Psi}_G)^{-1} \tag{196}$$

**Proof**. Without loss of generality, we assume that $\hat{\Psi}_G$ is non-degenerate. Let $\tilde{\Omega} \;=\; (\hat{\Psi}_G)^{1/2}\Omega$. Then, the columns of $\tilde{\Omega}$ are independent, identically distributed 15-dimensional Gaussian vectors, $\tilde{\omega}_i \sim N(0, \hat{\Psi}_G)$. Therefore,

$$(\hat{\Psi}_G)^{1/2}\Omega(zI + \Omega'\hat{\Psi}_G\,\Omega)^{-1}\Omega'(\hat{\Psi}_G)^{1/2} \;=\; \tilde{\Omega}(zI + \tilde{\Omega}'\tilde{\Omega})^{-1}\tilde{\Omega}' \tag{197}$$

By the Woodbury matrix identity,

$$(zI + \tilde{\Omega}'\tilde{\Omega})^{-1} \;=\; z^{-1}I \;-\; z^{-2}\tilde{\Omega}'(I + z^{-1}\tilde{\Omega}\tilde{\Omega}')^{-1}\tilde{\Omega} \tag{198}$$

We have by the law of large numbers that

$$Q = \tilde{\Omega}\tilde{\Omega}' \;\to\; \hat{\Psi}_G \tag{199}$$

almost surely. Therefore,

$$\tilde{\Omega}(zI+\tilde{\Omega}'\tilde{\Omega})^{-1}\tilde{\Omega}' = z^{-1}Q - z^{-2}Q(I+z^{-1}Q)^{-1}Q \;\to\; z^{-1}\hat{\Psi}_G - z^{-2}\hat{\Psi}_G(I+z^{-1}\hat{\Psi}_G)^{-1}\hat{\Psi}_G \tag{200}$$

almost surely. By direct calculation, this expression coincides with $\hat{\Psi}_G^{1/2}(zI + \hat{\Psi}_G)^{-1}\hat{\Psi}_G^{1/2}$. The proof is complete. □

The proof is complete. □

# F Additional Exhibits

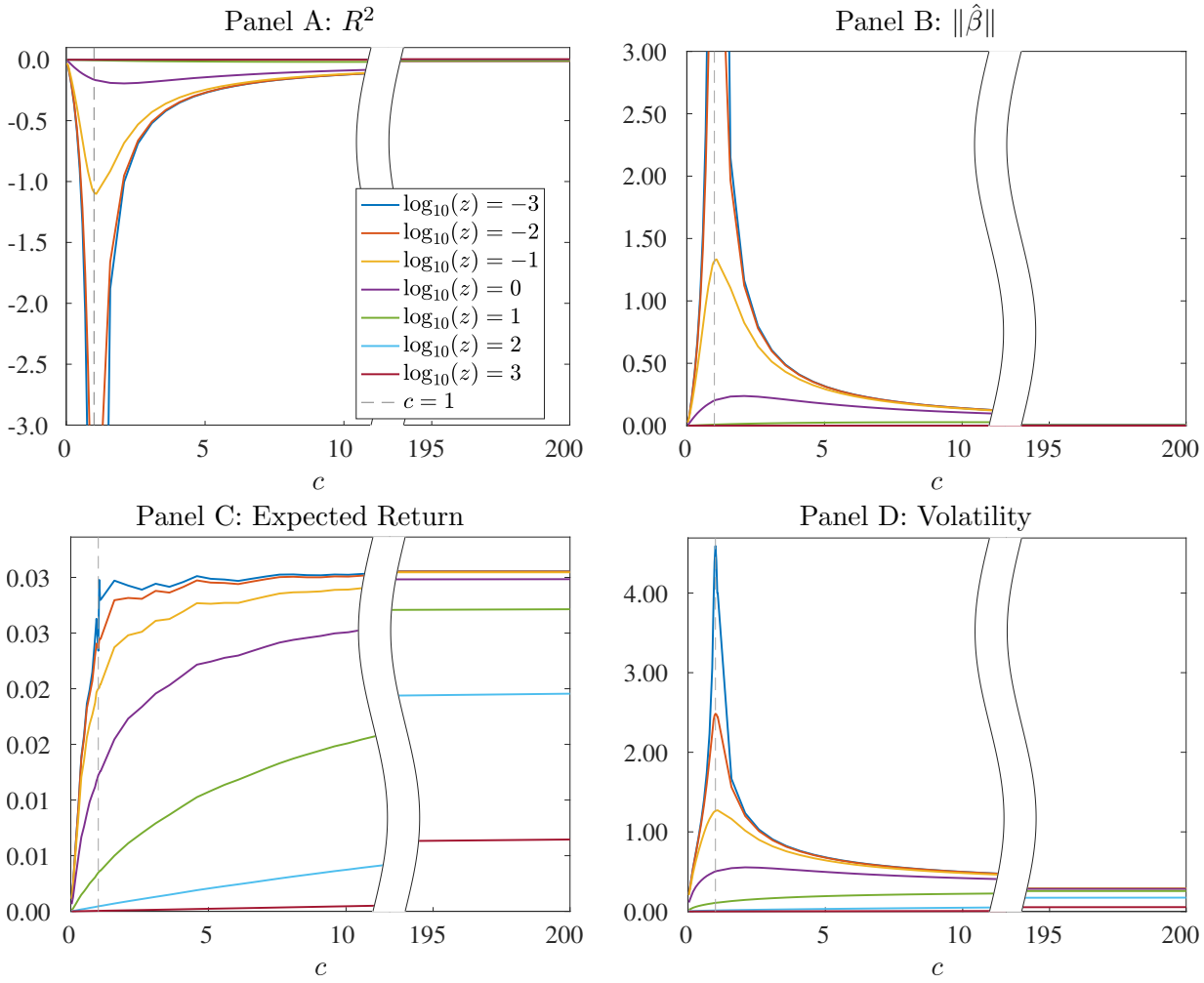



**Figure 12:** Out-of-sample Market Timing Performance With 60-month Training Window

**Note.** Out-of-sample prediction accuracy and portfolio performance estimates for empirical analysis described in Section 6.3. Training window is $T = 60$ months and predictor count $P$ (or $cT$) ranges from 2 to 12,000. Predictors are RFFs generated from 15 Goyal and Welch (2008) predictors with $\gamma = 2$.

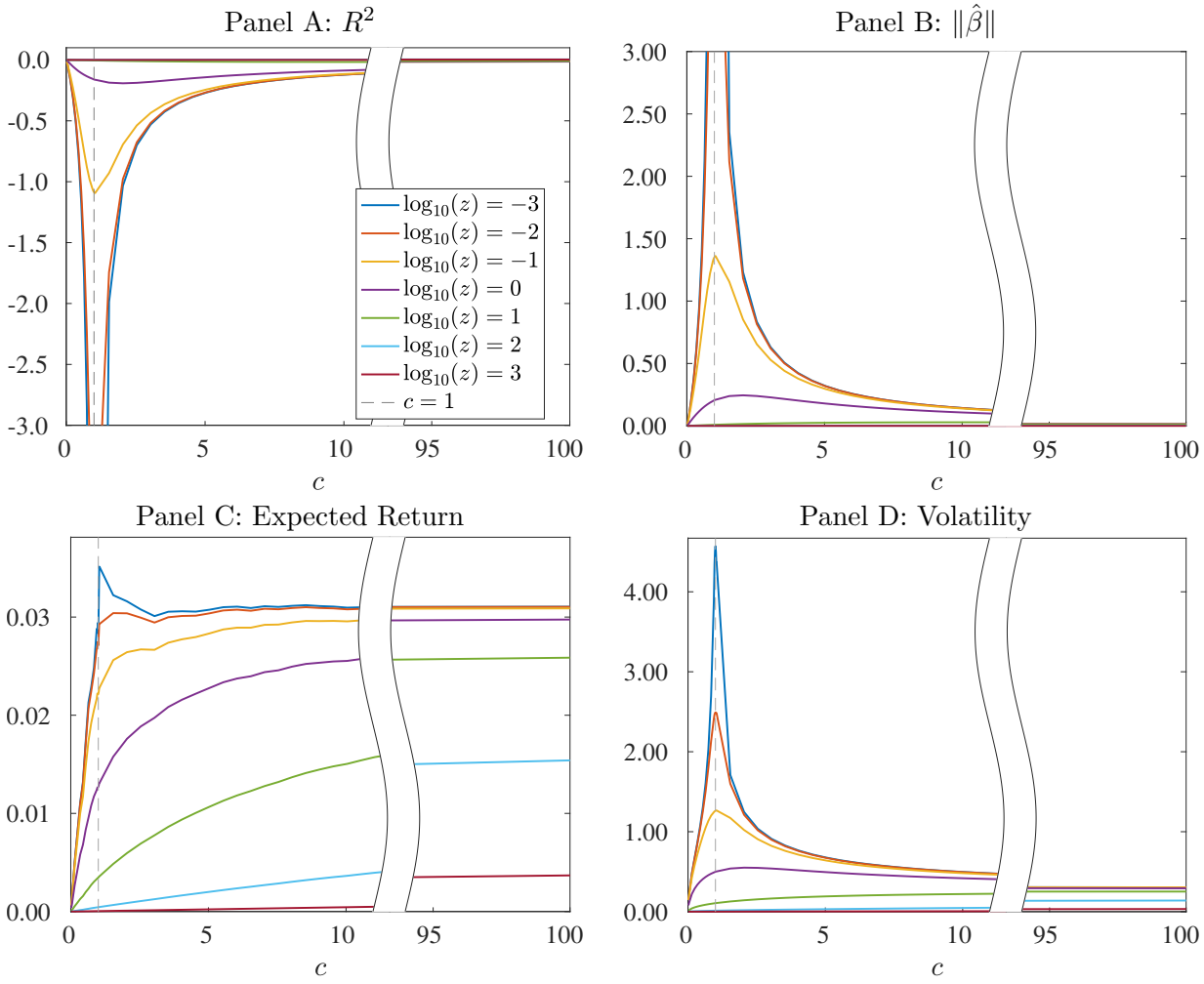



**Figure 13:** Out-of-sample Market Timing Performance With 120-month Training Window

**Note.** Out-of-sample prediction accuracy and portfolio performance estimates for empirical analysis described in Section 6.3. Training window is $T = 120$ months and predictor count $P$ (or $cT$) ranges from 2 to 12,000. Predictors are RFFs generated from 15 Goyal and Welch (2008) predictors with $\gamma = 2$.

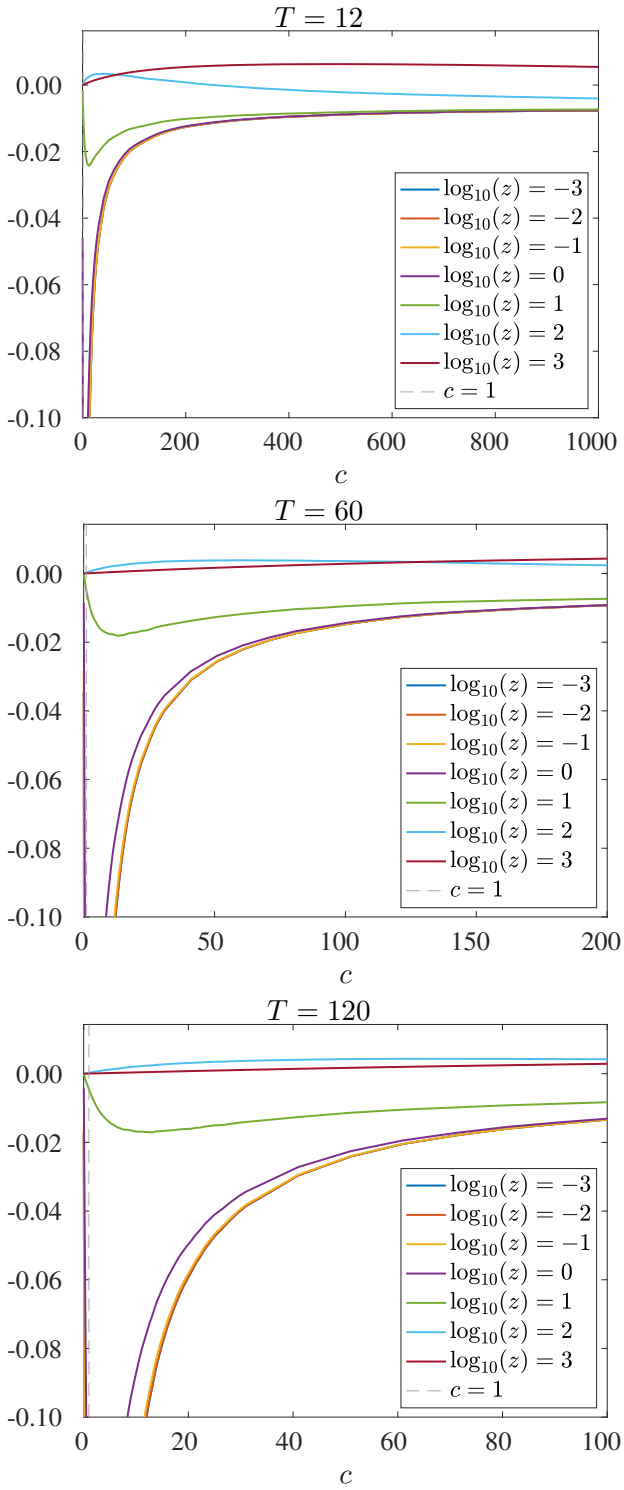



**Figure 14:** Out-of-Sample $R^2$ Detail

**Note.** Out-of-sample prediction accuracy for empirical analysis described in Section 6.3. Training window is $T = 12$, 60, or 120 months and predictor count $P$ (or $cT$) ranges from 2 to 12,000. Predictors are RFFs generated from 15 Goyal and Welch (2008) predictors with $\gamma = 2$.

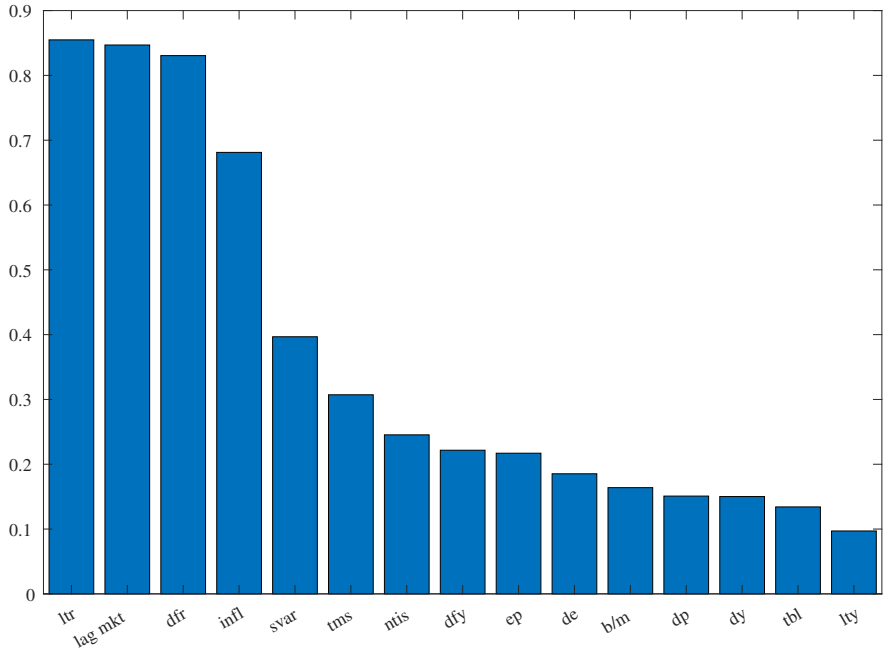



**Figure 15:** Normalized Volatility of Predictors in 12-Month Windows

**Note.** Bars show (normalized) average predictor volatility in 12-month windows, calculated by first scaling each predictor by its full sample volatility, then calculating the rolling 12-month volatility of each predictor, and finally calculating the time series average of those 12-month volatilities. In this calculation, an iid variable will have average 12-month volatility of about one because its volatility over any window should, on average, equal its unconditional volatility. For persistent variables, this value will be closer to 0 since it takes longer for such a variable to realize its total unconditional volatility.

**Table 2:** Comparison With Univariate Timing Strategies

**Note.** Information ratio (and associated alpha $t$-statistic) of the high-complexity strategy on the linear univariate timing strategy of each predictor. The univariate timing strategy is defined as the product of a predictor at time $t$ with the market return at $t+1$. We also report the information ratio versus all 15 univariate strategies simultaneously ("All"), based on the out-of-sample tangency portfolio of the 15 timing strategies scaled to have an expected volatility of 20%.

| | dfy | infl | svar | de | lty | tms | tbl | dfr |
|---|---|---|---|---|---|---|---|---|
| IR | 0.41 | 0.46 | 0.46 | 0.33 | 0.41 | 0.33 | 0.44 | 0.47 |
| $t$ | 3.9 | 4.4 | 4.3 | 3.1 | 3.9 | 3.1 | 4.2 | 4.5 |
| $R^2$ | 3.6% | 0.9% | 0.3% | 11.3% | 6.0% | 8.5% | 3.5% | 0.0% |
| | dp | dy | ltr | ep | bm | ntis | lag-mkt | All |
| IR | 0.31 | 0.31 | 0.46 | 0.35 | 0.38 | 0.45 | 0.48 | 0.32 |
| $t$ | 2.9 | 2.9 | 4.3 | 3.3 | 3.6 | 4.3 | 4.5 | 2.9 |
| $R^2$ | 13.6% | 13.4% | 0.6% | 11.0% | 6.4% | 3.4% | 5.9% | 11.4% |

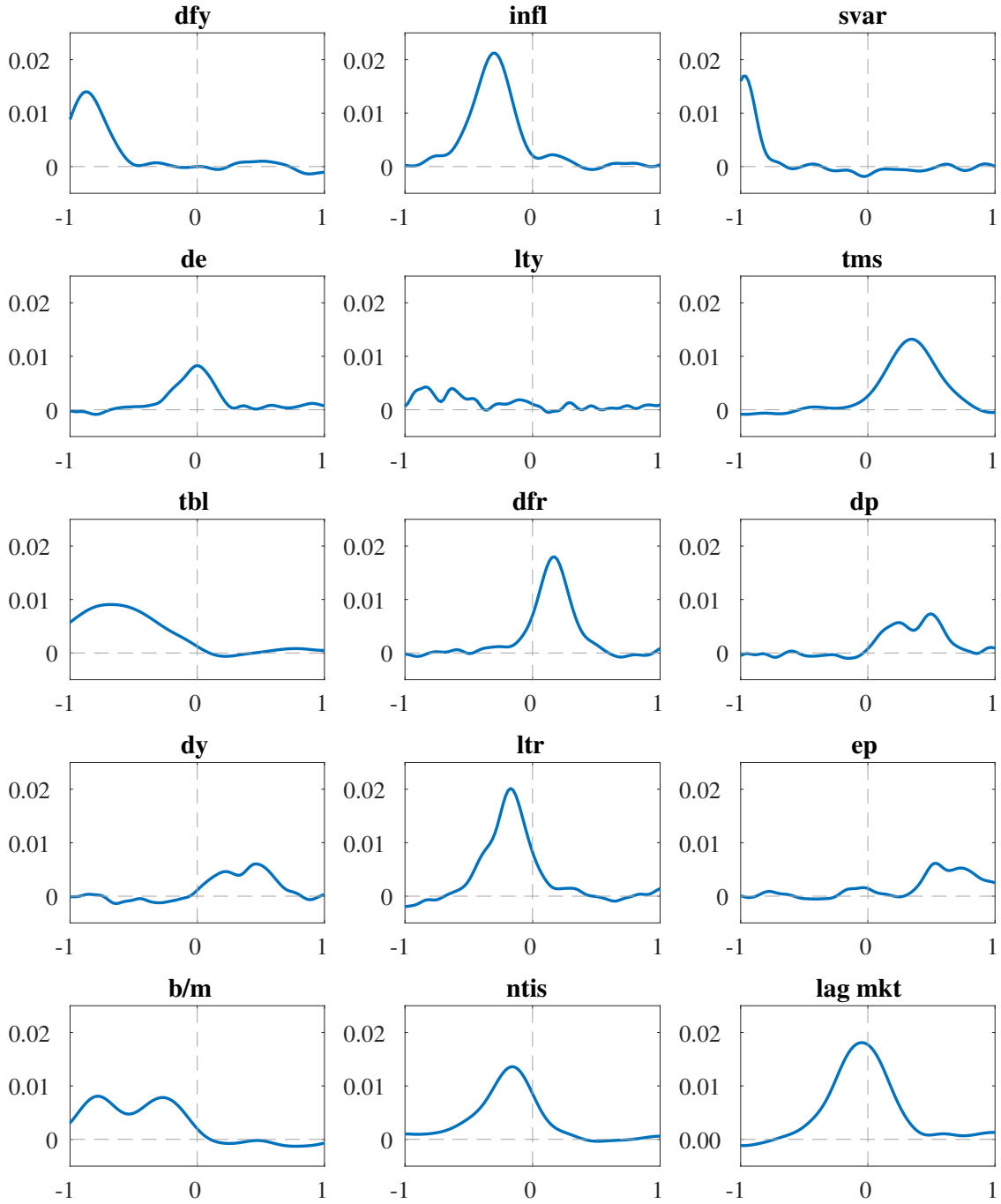



**Figure 16:** Nonlinear Prediction Effects

**Note.** Panels show marginal nonlinear return prediction patterns associated with each of the 15 predictors. To trace the impact of predictor $i$ on expected returns, we fix the prediction model estimated from a given training sample and fix the values of all variables other than $i$ at their values at the time of the forecast. Next, we vary the value of the $i^{th}$ predictor from its full sample min (corresponding to $-1$ on the plots) to its full sample max (corresponding to $+1$) and record how the return prediction varies. Then we average this prediction response function across all training windows and plot the result. Training window is $T = 12$ with $P = 12,000$, $z = 10^3$, and $\gamma = 2$.

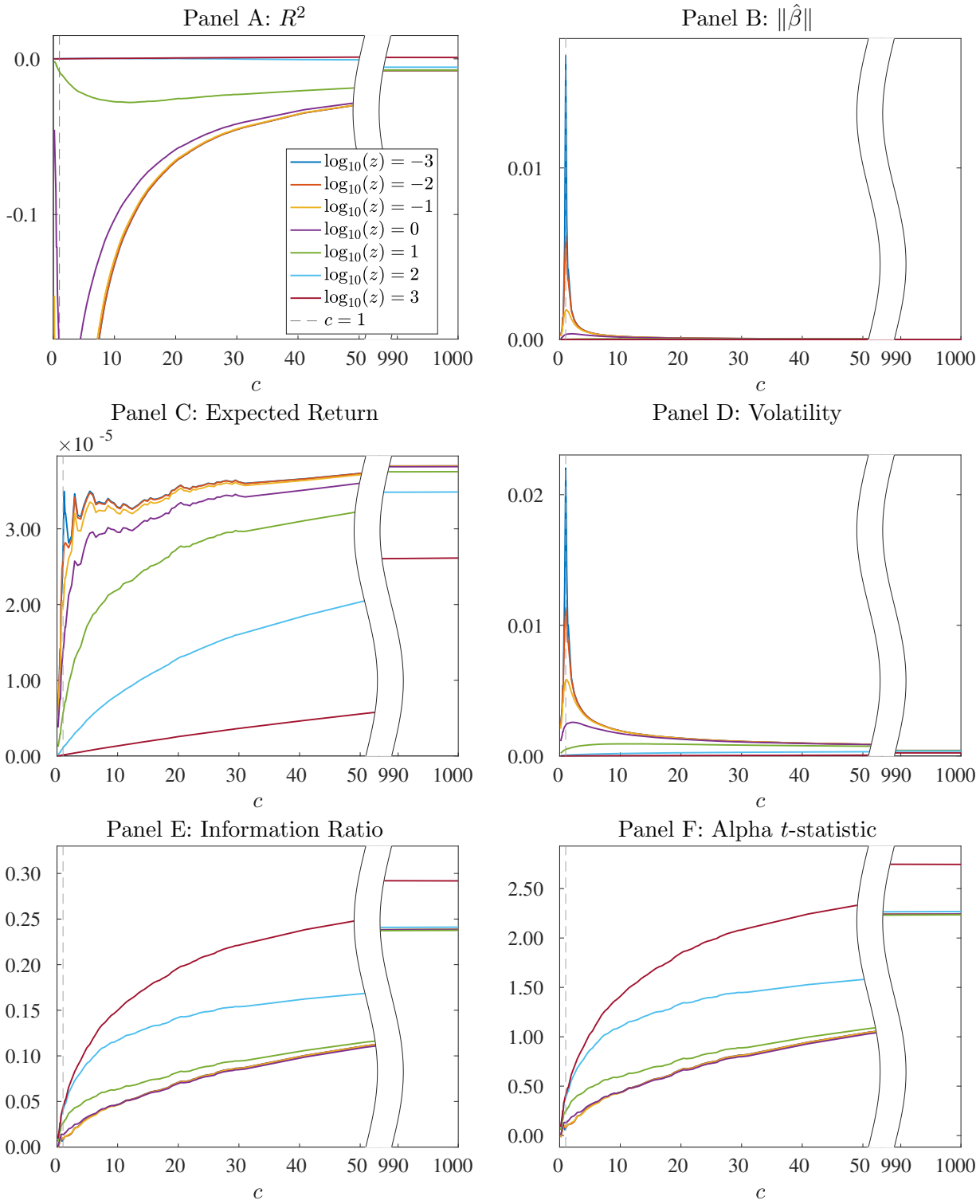



**Figure 17:** Out-of-sample Market Timing Performance With Un-standardized Returns

**Note.** Out-of-sample prediction accuracy and portfolio performance estimates for empirical analysis described in Section 6.3. Training window is $T = 12$ months and predictor count $P$ (or $cT$) ranges from 2 to 12,000. Predictors are RFFs generated from 15 Goyal and Welch (2008) predictors with $\gamma = 2$. In contrast to our main analysis, returns are not volatility-standardized in this figure.

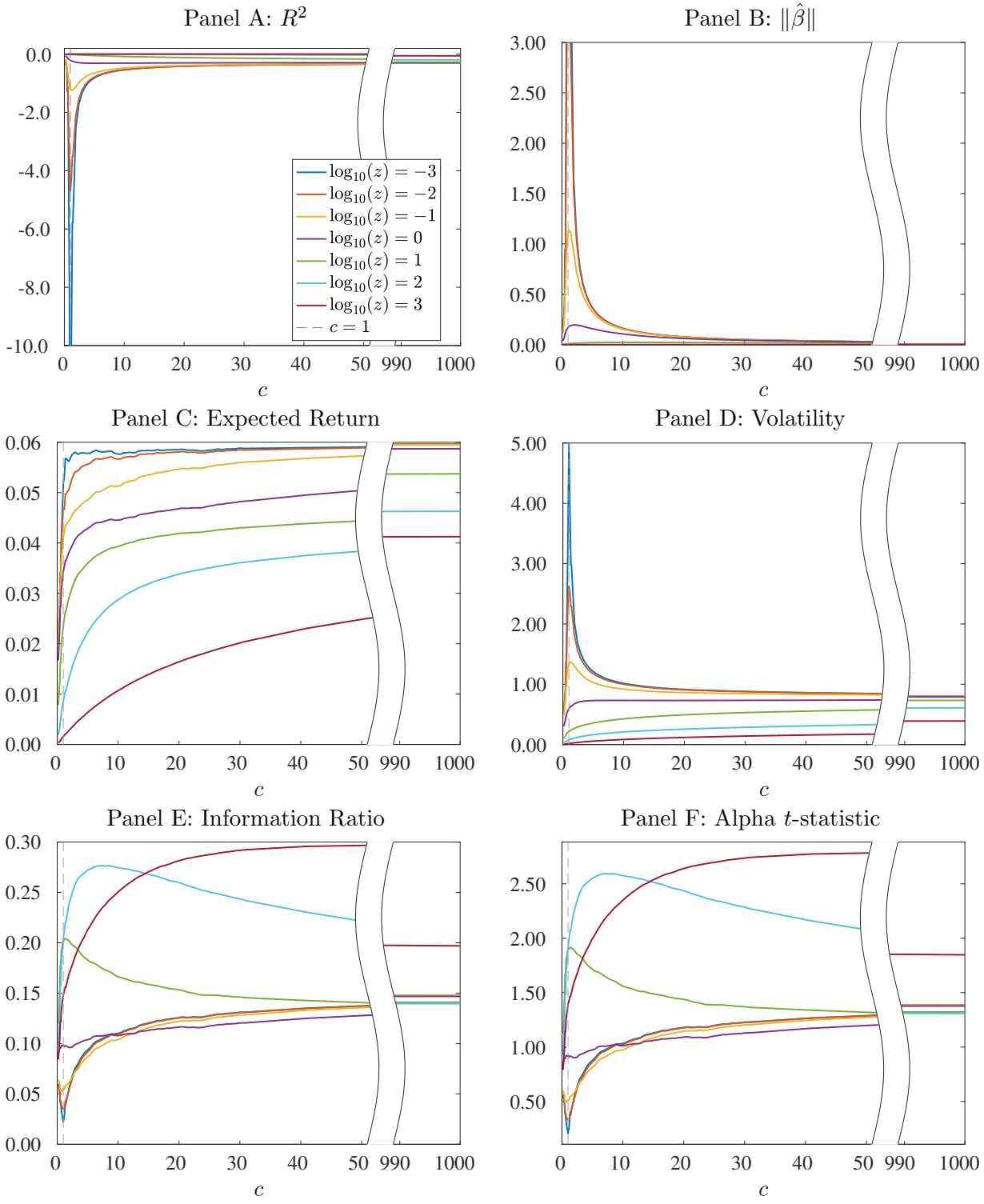



**Figure 18:** Out-of-sample Market Timing Performance With Bandwidth $\gamma = 1$

**Note.** Out-of-sample prediction accuracy and portfolio performance estimates for empirical analysis described in Section 6.3. Training window is $T = 12$ months and predictor count $P$ (or $cT$) ranges from 2 to 12,000. Predictors are RFFs generated from 15 Goyal and Welch (2008) predictors with $\gamma = 1$.

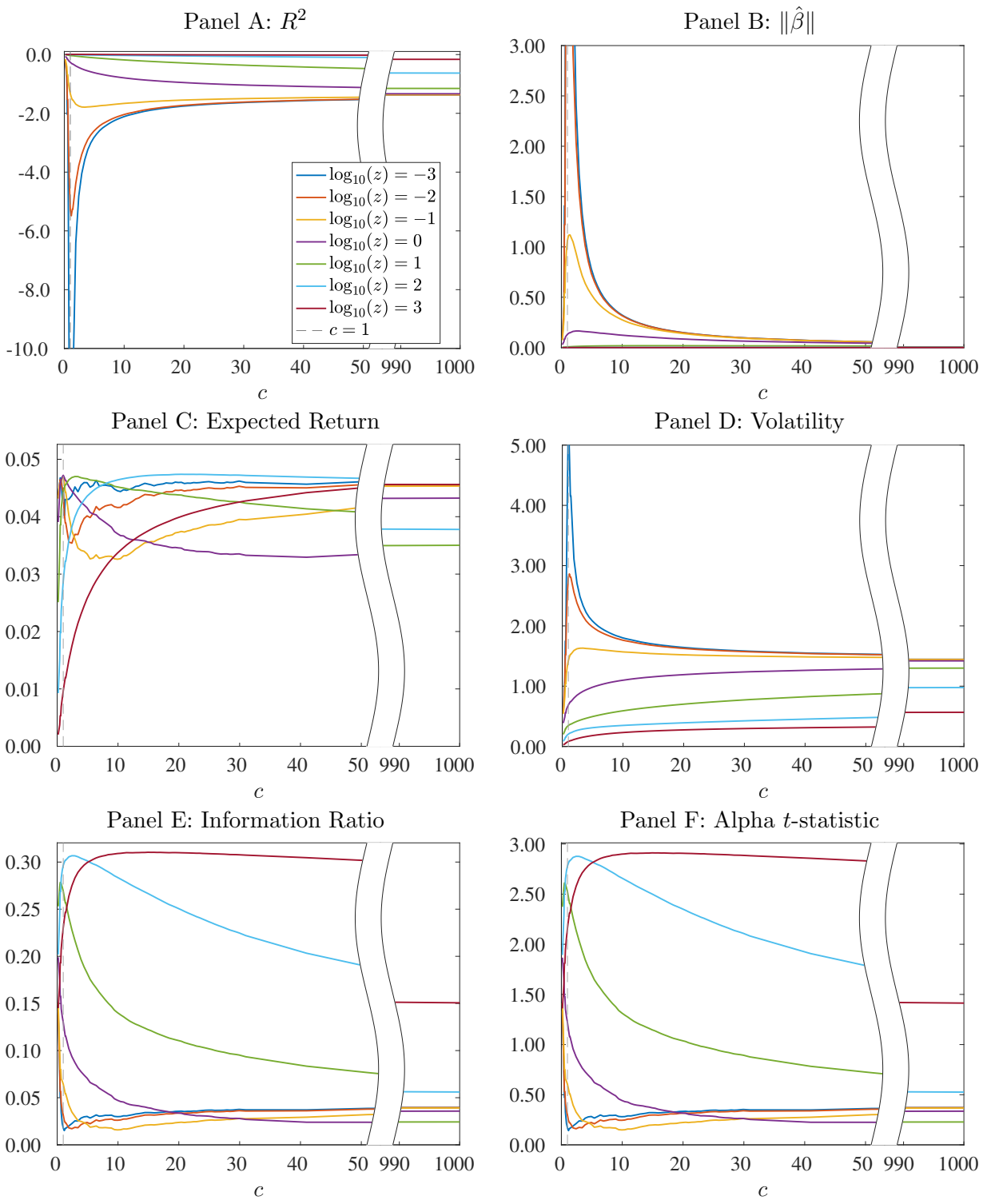



**Figure 19:** Out-of-sample Market Timing Performance With Bandwidth $\gamma = 0.5$

**Note.** Out-of-sample prediction accuracy and portfolio performance estimates for empirical analysis described in Section 6.3. Training window is $T = 12$ months and predictor count $P$ (or $cT$) ranges from 2 to 12,000 using a range of $P$. Predictors are RFFs generated from 15 Goyal and Welch (2008) predictors with $\gamma = 0.5$.

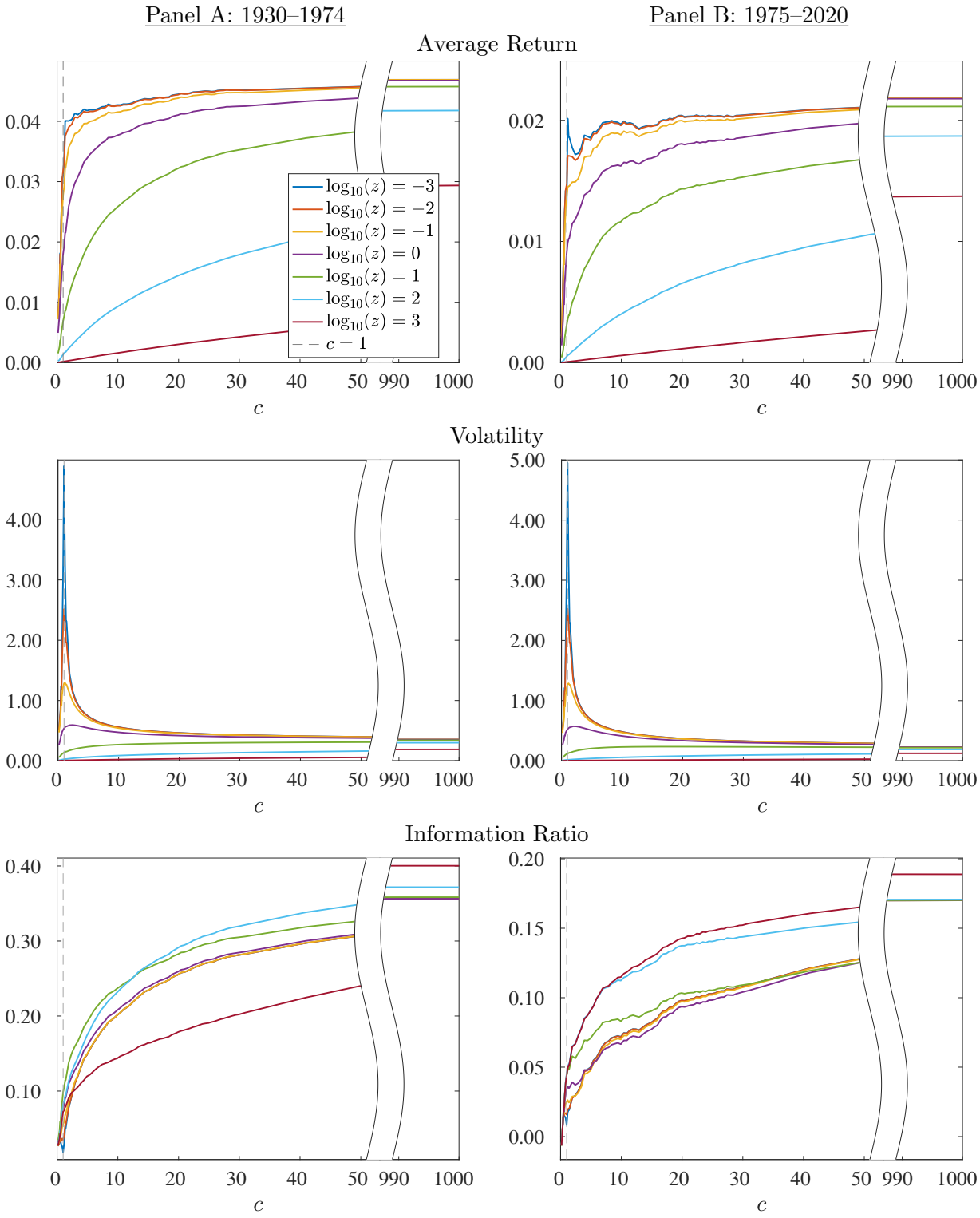



**Figure 20:** Out-of-sample Performance by Subsample ($T = 12$)

**Note.** Subsample analysis of 1930–1974 and 1975–2020. See notes in Figures 7 and 8.

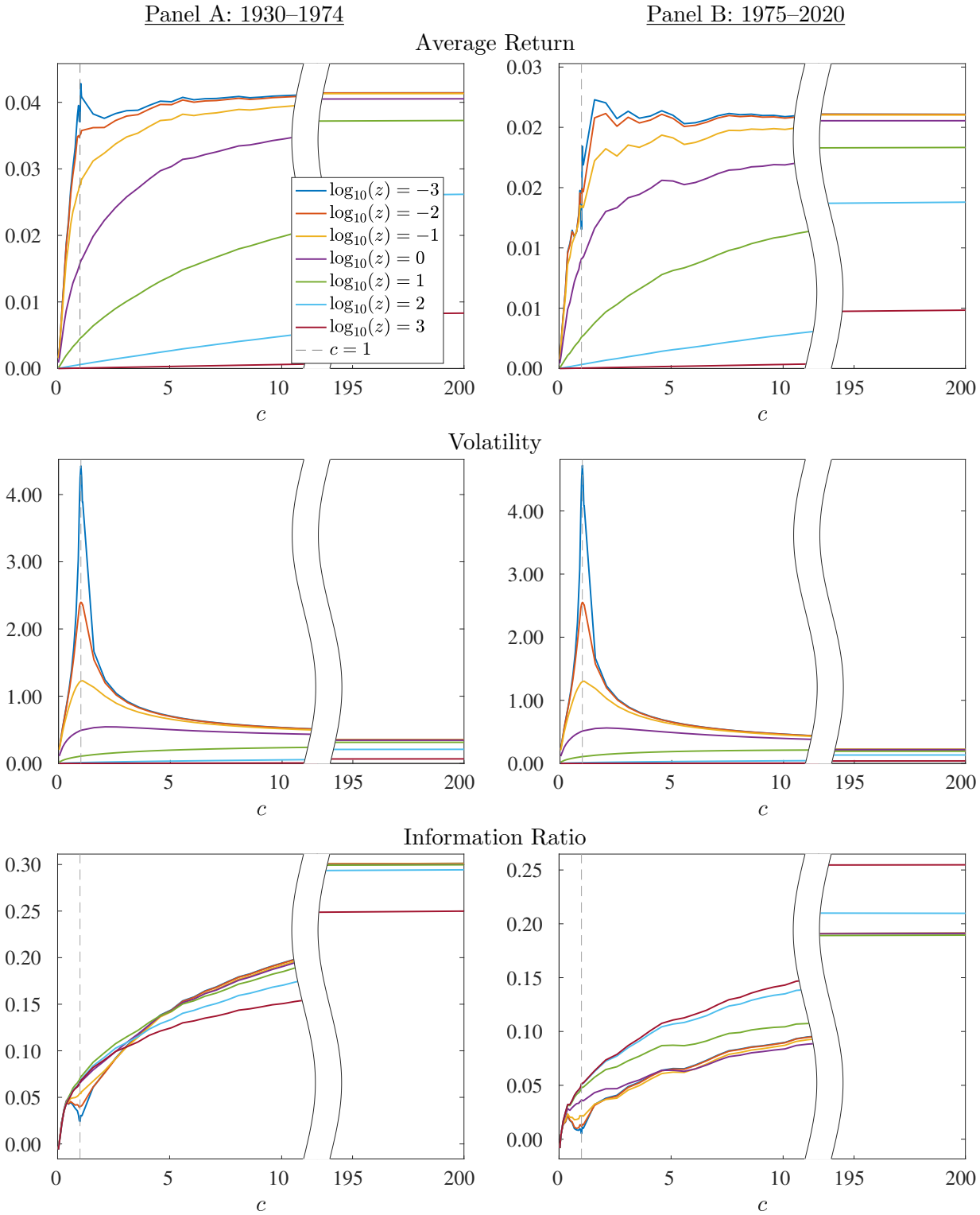



**Figure 21:** Out-of-sample Performance by Subsample for ($T = 60$)

**Note.** Subsample analysis of 1930–1974 and 1975–2020. See notes in Figures 7 and 8.

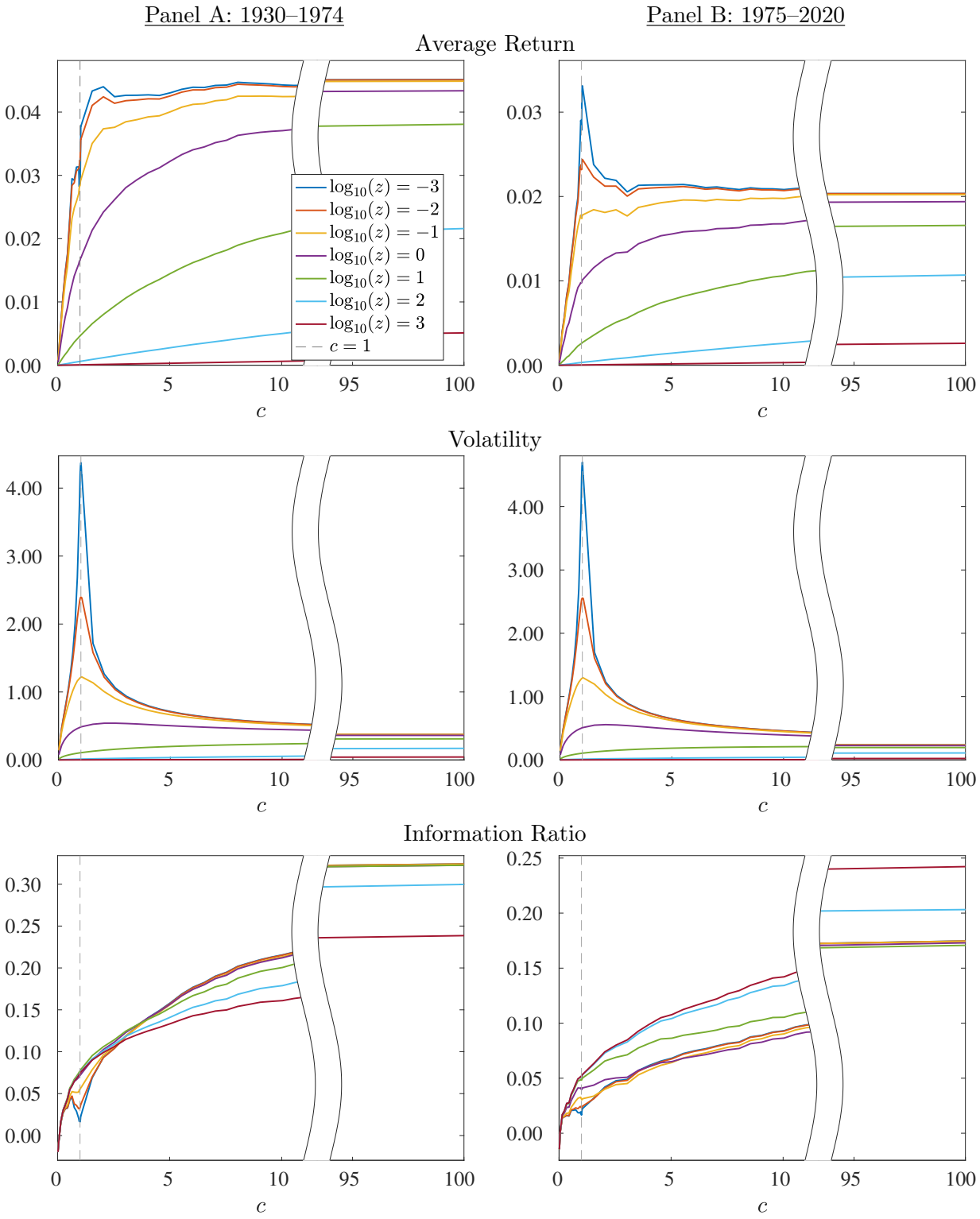



**Figure 22:** Out-of-sample Performance by Subsample for ($T = 120$)

**Note.** Subsample analysis of 1930–1974 and 1975–2020. See notes in Figures 7 and 8.

## G Comparison With Momentum

We begin by rewriting the forecast from a linear model as a weighted sum of training sample returns. In particular, consider making a forecast at time $T$ for the return at time $T+1$. Estimation is performed in a 12-month training window ending at date $T$. The regression forecast is:

$$\begin{aligned} S_T'\beta &= S_T'\left(\sum_{t=T-11}^{T} S_{t-1}S_{t-1}'\right)^{-1}\sum_{t=T-11}^{T} r_t S_{t-1} \\ &= \sum_{t=T-11}^{T} r_t \underbrace{\left[S_T'\left(\sum_{t=T-11}^{T} S_{t-1}S_{t-1}'\right)^{-1} S_{t-1}\right]}_{w_t} \\ &= \sum_{t=T-11}^{T} r_t w_t \end{aligned}$$

The equation states that, in general, rolling out-of-sample return predictions amount to a sequence of moving averages of past returns. It follows that if the predictors are static in the training window, $w_t$ will equally weight the training returns—in other words, it will behave like time series momentum Moskowitz et al. (2012). Since some of the Goyal and Welch (2008) predictors are highly persistent, we wish to rule out the possibility that the high-complexity model simply captures a time-series momentum effect.

The most direct approach to this question is to put the high-complexity market timing strategy with a 12-month training window ("VoC" for short) head-to-head with a timing strategy based on 12-month momentum. Define the high-complexity strategy as

$$R_{voc,t+1} = \pi_{voc,t} R_{m,t+1},$$

where $R_m$ is the market return and $\pi_{voc,t}$ is the out-of-sample prediction $\hat{E}_t[R_{m,t+1}]$ from our rolling 12-month random features model.[51] Define the momentum strategy as

$$R_{mom,t+1} = \pi_{mom,t} R_{m,t+1},$$

where

$$\pi_{mom,t} = \frac{\frac{1}{12}\sum_{j=0}^{11} R_{m,t-j}}{\sqrt{\frac{1}{12}\sum_{j=0}^{11}\left(R_{m,t-j} - \left(\frac{1}{12}\sum_{k=0}^{11} R_{m,t-k}\right)\right)^2}}$$

After construction, we normalize both of these to have identical sample volatility of 20% so that means/alphas are directly comparable.

[51] For this analysis, we use the raw market return without volatility standardization, to remain directly comparable with the momentum strategy. However, $\pi_{voc,t}$ is the position from the main paper, and is trained on volatility-standardized returns.

The annualized out-of-sample performance of these two strategies from the sample 1931–2020 is:

| | Mean | Vol | SR | IR vs. $R_m$ |
|---|---|---|---|---|
| $R_{voc}$ | 8.40% | 20% | 0.42 | 0.32 |
| $R_{mom}$ | 6.39% | 20% | 0.32 | 0.33 |

Both strategies have similar gross performance and information ratios in a spanning regression against the market.

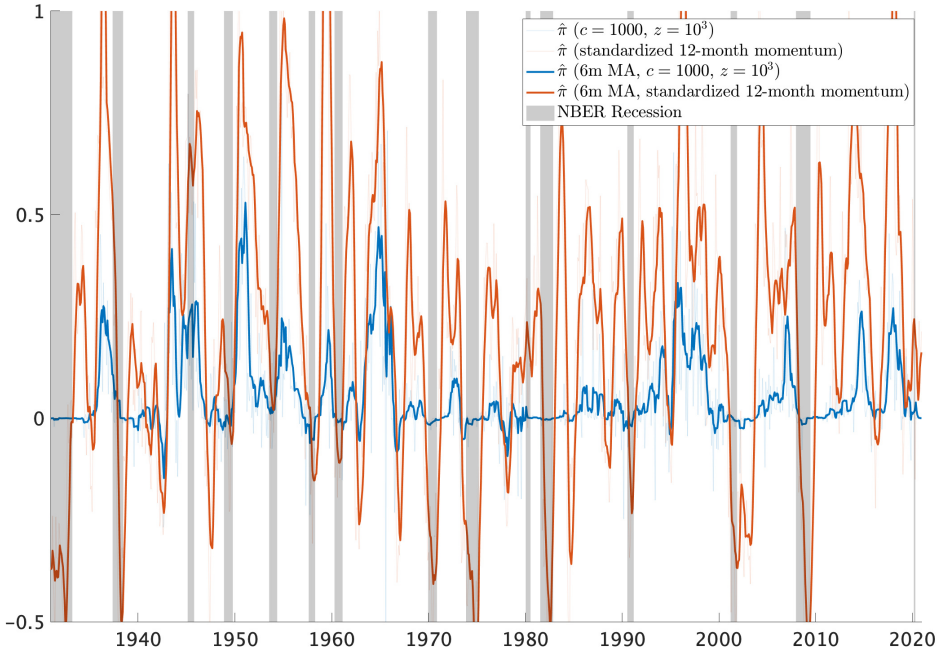
Next, we run spanning regressions of each strategy on the other:

| | $R_{voc}$ on $R_{mom}$ | $R_{mom}$ on $R_{voc}$ |
|---|---|---|
| $\alpha$ | 5.77% | 2.93% |
| $t(\alpha)$ | 2.99 | 1.51 |
| $R^2$ | 0.17 | 0.17 |

These results tease apart the differences in the strategies. First, from the 17% $R^2$, we see the strategies are 42% correlated. Second, VoC has a significant monthly alpha of 5.8% per annum versus momentum ($t = 3.0$). Momentum has positive alpha versus our strategy, but it is smaller (2.9%) and insignificant ($t = 1.5$).

Next, we see that the strategies are only partially correlated because they take different bets. The clearest way to illustrate their difference is by plotting their timing weights:

Timing Bets: 12-Month Window




The two sets of bets have a 49% correlation. We see from the plot that they indeed share some common episodes of large positive bets. But momentum is clearly a different strategy, taking

some big bets when VoC does not. Furthermore, momentum bets on market downturns in times when VoC does not.

The differences between the two strategies are perhaps unsurprising in light of Section 6.5, which shows that many of the Goyal and Welch (2008) predictors have very low persistence, and these low persistence predictors are, in fact, the main drivers of VoC performance.

Other important evidence comes from our comparison of VoC and the Goyal-Welch linear "kitchen sink" regression in Table 1. The rationale that a 12-month regression with persistent regressors will recover 12-month momentum also applies to the linear regression (it too is conducted in rolling 12-month training windows). The best-performing linear strategy (that with penalty $z = 10^3$) is highly correlated with 12-month momentum (92.4%), and its annualized alpha versus 12-month momentum is 0.63% per annum with a $t$-stat of 0.86 (an information ratio of 0.09). In other words, the linear regression turns out to be 12-month momentum, more or less. Yet VoC has a significant alpha of 4.40% ($t = 2.5$) over this linear model (an information ratio of 0.26, shown in Table 1). The comparison between the complex nonlinear VoC model and the linear kitchen sink is apples-to-apples because the momentum effect due to short training with persistent regressors will affect both strategies. Nonetheless, VoC significantly improves over the linear model. We interpret this evidence as saying that the linear model cannot find a good way to use the variation in fast-moving predictors, so it leans more heavily on the static predictors. Meanwhile, VoC learns to take advantage of the fast predictors in nonlinear ways.

To sum up, momentum explains only 17% of the variation in the VoC strategy and 31% of its performance ($1 - 5.77\%/8.40\%$). The evidence suggests that momentum is not the primary driver of VoC performance.

**Swiss Finance Institute**

Swiss Finance Institute (SFI) is the national center for fundamental research, doctoral training, knowledge exchange, and continuing education in the fields of banking and finance. SFI's mission is to grow knowledge capital for the Swiss financial marketplace. Created in 2006 as a public–private partnership, SFI is a common initiative of the Swiss finance industry, leading Swiss universities, and the Swiss Confederation.

swiss:finance:institute

c/o University of Geneva, Bd. Du Pont d'Arve 42, CH-1211 Geneva 4
T +41 22 379 84 71, rps@sfi.ch, www.sfi.ch