The Virtue of Complexity in Machine Learning Portfolios

Bryan Kelly, Semyon Malamud, and Kangying Zhou*

February 24, 2022

Abstract

We theoretically characterize the behavior of machine learning portfolios in the high complexity regime, i.e. when the number of parameters exceeds the number of observations. We demonstrate a surprising "virtue of complexity": Sharpe ratios of machine learning portfolios generally increase with model parameterization, *even* with minimal regularization. Empirically, we document the virtue of complexity in US equity market timing strategies. High complexity models deliver economically large and statistically significant out-of-sample portfolio gains relative to simpler models, due in large part to their remarkable ability to predict recessions.

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, Paul Goldsmith-Pinkham, Amit Goyal, and seminar and conference participants at AQR and Yale. 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.

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 the behavior of portfolios formed from heavily parameterized models.

We provide a theoretical analysis of such "machine learning portfolios." To provide transparent and intuitive characterizations, our theoretical environment has two simplifying aspects. First, we study linear least squares prediction models. This is without loss of generality as a number of recent papers have established an equivalence between sophisticated models such as deep neural networks and high-dimension linear models (Jacot et al., 2018; Hastie et al., 2019; Allen-Zhu et al., 2019). Second, we focus on a single risky asset so that the portfolio problem boils down to optimal market timing. While our analysis can be applied to a panel of many assets, the role of covariances in optimal machine learning portfolios introduces an unnecessary layer of complexity to our analysis. In other words, the two simplifying aspects of our setting make our key findings more accessible, and neither are critical to our conclusions.

As a backdrop to our analysis, we begin with a well understood deficiency of ordinary least squares (OLS) prediction. 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, and in turn its expected out-of-sample Sharpe ratio collapses to zero. The intuition behind this is simple: When 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. A common interpretation of

this is insidious overfit: With $P = T$, the regression exactly fits the training data and the model does not generalize out-of-sample.

In this paper, we study behavior of portfolios in the *high model complexity* regime where the number of predictors *exceeds* the number of observations ($P > T$).[1] 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. This seemingly counter-intuitive phenomenon is sometimes called "benign overfit" (Bartlett et al., 2020; Tsigler and Bartlett, 2020).

We analyze related phenomena in the context of machine learning portfolios. We establish the striking theoretical result that market timing strategies based on ridgeless least squares predictions generate positive Sharpe ratio improvements for arbitrarily high levels of model complexity. Stated more plainly, when the true data generating process (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 show that strategies based on extremely high-dimensional models can thrive out-of-sample, even with minimal ridge regularization.

Our theoretical analysis delivers a number of additional conclusions. First, it shows that the out-of-sample R^2 from a prediction model is a poor measure of its economic value. A market timer can generate potentially large economic profits even when the R^2 is large and

[1] The statistics and machine learning community often refer to $P > T$ as the "over-parameterized" regime. We avoid this terminology as it suggests that too many parameters are being used, and this is not necessarily the case. For example, the true data generating process may be highly complex (i.e., P is large relative to T), and thus a correctly specified model would require $P > T$. We would normally think of that when an empirical model has the same specification as the true model, it is correctly parameterized as opposed to over-parameterized.

negative. The reason is that the R^2 is heavily influenced by the variance of forecasts.[2] 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 economically large.

Second, we study two theoretical cases, one for correctly specified models and one for mis-specified models. Again surprisingly, even mis-specified models achieve positive expected out-of-sample Sharpe ratio gains from market timing (though, naturally, correctly specified models do even better). In the correctly specified case, we learn how timing portfolios behave 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 a variety of DGPs. However, assuming correct model specification is unrealistic in a few key respects. To begin with, 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. Furthermore, the comparative statics in our analysis of correctly specified models (which simultaneously vary the complexity of the true DGP and the model) are not relevant for any specific empirical application because the DGP for that data set is of course fixed. In practice, when we vary the empirical model specification from simple to complex, we change how accurately the model approximates a fixed DGP.

We structure our theoretical analysis of mis-specified models with these real-world considerations in mind. We show that the performance of mis-specified machine learning portfolios tend to *continually improve* as model complexity increases. This is true even when we push the empirical model to extremely large parameterizations (holding the number of

[2]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.

observations constant). The intuition is that when the true DGP is unknown, the gains from improved approximations of the DGP dominate statistical costs of heavy parameterization.

Third, while the results discussed so far refer primarily to the case of ridgeless regression, we show that machine learning portfolios often benefit further by moving away from the ridgeless limit and 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 at its worst.

From a technical standpoint, we characterize the behavior of portfolios in the high complexity regime using asymptotic analysis as the size of the model grows with the number of observations at a fixed rate ($T \to \infty$ and $P/T \to c > 0$). Such analysis requires the apparatus of random matrix theory, on which we draw heavily to derive our results. Conceptually, this delivers an approximation for 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. In particular, we study market timing strategies based on predictions from very simple models with a single parameter to extremely complex models with over 1,000 parameters (applied to training samples with as few as 12 months). The data inputs to our models are 15 standard predictor variables from the finance literature compiled by Welch and Goyal (2008). To map our data analysis to the theory, we require a method that smoothly transitions from low to high complexity models while holding 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 Welch and Goyal (2008) predictors as inputs and maintain the core ridge regression structure of our theory.

We find extraordinary agreement between the 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 look essentially like 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 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 an 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.[3] In particular, wide neural networks (many nodes in each layer) are effectively kernel regressions, and early stopping 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.[4]

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 ridge(less) regression under very general conditions.[5] They are also the first to introduce mis-specified models where some of 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

[3]See, for example, Jacot et al. (2018); Hastie et al. (2019); Du et al. (2018, 2019); Allen-Zhu et al. (2019).
[4]See, for example, Spigler et al. (2019); Belkin et al. (2018, 2019, 2020); Bartlett et al. (2020).
[5]See also Richards et al. (2021) who obtain less general results in an asymptotic setting (as in our paper).

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. Finally, in the finance literature, our paper is tangentially related to Martin and Nagel (2021) who examine equilibrium market efficiency implications of machine learning models.

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 theory and empirical robustness analysis. 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*

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

with ε_{t+1} *i.i.d.,* $E[\varepsilon_{t+1}] = E[\varepsilon_{t+1}^3] = 0, E[\varepsilon_{t+1}^2] = \sigma^2$, $E[\varepsilon_{t+1}^4] < \infty$, *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 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* Ψ *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 Lindenberg condition*

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

As we show below, the theoretical properties of machine learning portfolios depend heavily on the *distribution of eigenvalues* of Ψ. 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 Ψ are well-behaved in this limit.

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

$$F^\Psi(x) = \frac{1}{P} \sum_{k=1}^{P} \mathbf{1}_{\lambda_k(\Psi) \leq x} \tag{2}$$

converges to a non-random probability distribution H supported on $(0, +\infty)$.[6] Furthermore, Ψ is uniformly bounded as $P \to \infty$. We will use

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

to denote asymptotic moments of the eigenvalues of Ψ.

Our last assumption governs the behavior of the true predictive coefficient, β.

Assumption 4 *We assume $\beta = \beta_P$ is random, $\beta = (\beta_i)_{i=1}^P \in \mathbb{R}^P$, independent of S and R, and satisfies $E[\beta] = 0$, and $E[\beta\beta'] = P^{-1} b_{*,P} I$ for some constant $b_{*,P} = E[\|\beta\|^2]$,[7] and satisfies $b_{*,P} \to b_*$ almost surely, for some $b_* > 0$. Furthermore, $E[\beta_i^4] \leq cP^{-2}$ for some $c > 0$, and β satisfy the same Lindenberg condition as X.*

Randomness of β in Assumption 4 is a device that allows us to characterize the prediction and portfolio problem for generic predictive coefficients. The assumption that β is mean zero is inconsequential; we could allow for 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. From a technical standpoint, it is possible to derive explicit expressions for

[6] If 0 is in the support of H, then Ψ is strictly degenerate, meaning that some signals are redundant.
[7] This identity follows because $b_* = \operatorname{tr} E[\beta\beta'] = E[\operatorname{tr}(\beta\beta')] = E[b_*]$.

portfolio performance without this assumption, but the expressions become more complex. In this case, the asymptotic behavior depend on the distribution of projections of β on the eigenvectors of Ψ (the signal principal components).[8] We leave this important generalization for future research.

When β 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 ℓ_2 norm of β, controls the "true" Sharpe ratio. The assumption ensures that Sharpe ratios of timing strategies remain bounded as the number of predictors grows. In other words, our setting is one with many predictors, 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 β. This is due to a law of large numbers in the $P \to \infty$ limit, and is guaranteed by the following lemma.[9]

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

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

[8] See, Hastie et al. (2019). In particular, when β 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.

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

where π_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,

$$SR = \frac{E[R^{\pi}_{t+1}]}{\sqrt{E[(R^{\pi}_{t+1})^2]}}. \tag{3}$$

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.[10]

Our analysis centers on the following timing strategy functional form:

$$\pi_t(\beta) = \beta' S_t. \tag{4}$$

This strategy takes positions equal to the asset's conditional expected return. Note that this timing strategy optimizes the *conditional* Sharpe ratio. That is, it achieves the same Sharpe ratio as the conditional Markowitz solution, $\pi_t^{\text{Cond. MV}} = E_t[R_{t+1}]/\text{Var}_t[R^2_{t+1}] = \beta' S_t$, according to equation (1). While strategy π_t is conditionally mean-variance efficient, it is not the optimizer of the unconditional objective in (3), which takes the form $\pi_t^{\text{Uncond. MV}} = \beta' S_t/(1 + (\beta' S_t)^2)$.[11] In the proof of Proposition 1 in the Appendix, we show that π_t in equation (4) and $\pi_t^{\text{Uncond. MV}}$ are equal up to third order terms.[12] We study $\pi_t = \beta' S_t$ for the simplicity of its linearity in both β and S_t, but note that our conclusions are identical for

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

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

[12] In particular, the Sharpe ratio in equation (3) is less than one due to the Cauchy-Schwarz inequality. We show that the difference in Sharpe ratios for π_t versus $\pi_t^{\text{Uncond. MV}}$ is on the order of the Sharpe ratio cubed.

$\pi_t^{\text{Uncond. MV}}$ because, in the limit as $P \to \infty$, the normalization factor $1 + (\beta' S_t)^2$ converges to a constant.[13]

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

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

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

The infeasible market timing Sharpe ratio is

$$SR \to \frac{1}{\sqrt{3 + (b_* \psi_{*,1})^{-1}}} < \left(\frac{1}{3}\right)^{1/2}. \tag{5}$$

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

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

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

$$MSE(\hat{\beta}) = E\left[\left(R_{t+1} - S_t'\hat{\beta}\right)^2 |\hat{\beta}\right] = E[R_{t+1}^2] - 2\underbrace{E[\hat{\pi}_t R_{t+1}|\hat{\beta}]}_{\substack{\text{Timing} \\ \text{Expected Return}}} + \underbrace{E[\hat{\pi}_t^2|\hat{\beta}]}_{\substack{\text{Timing} \\ \text{Leverage}}}. \tag{6}$$

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 (6) 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 β satisfies $E[\pi_t R_{t+1}] = E[\beta'\Psi\beta] = E[\pi_t^2]$.[14] 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 (1) 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 (5)).

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

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 as the prediction model becomes increasing 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 $P/T < 1$.

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 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 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, $(T^{-1}\sum_t S_t S_t')^+ \frac{1}{T}\sum_t S_t R_{t+1}$.[15] This solution is equivalent 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 ℓ_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

[15] 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.

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 random matrix theory that we build from.

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, Ψ. For any symmetric matrix Ψ, a convenient matrix identity states

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

where $\lambda_i(\Psi)$ are the eigenvalues of Ψ. Using formula (2), we can rewrite this identity as

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

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

$$m_{\Psi}(z) \coloneqq \int \frac{1}{x - z} dH(x) = \lim_{P\to\infty} \frac{1}{P}\operatorname{tr}\left((\Psi - zI)^{-1}\right). \tag{7}$$

The function $m_{\Psi}(z)$ is the *limiting* Stieltjes transform of the eigenvalue distribution of Ψ. Equation (7) is a powerful step towards understanding the least squares estimator in the machine learning regime (and hence machine learning predictions and portfolios). It states that key properties of the limiting inverse of the ridge-regularized signal covariance matrix can be completely characterized if we just know Ψ's eigenvalue distribution.

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

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

In traditional finite P statistics, we would have convergence between the sample covariance $\hat{\Psi}$ and the true covariance Ψ 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 Ψ 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 7), 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 $c > 0$.[16] The next result shows that, quite remarkably, if we constrain ourselves to linear ridge regression estimators, all asymptotic expressions depend only on $m(z;c)$ and do not require m_Ψ.[17]

Proposition 2 *We have*

$$\lim_{T\to\infty} \frac{1}{T} \operatorname{tr}((zI + \hat{\Psi})^{-1}\Psi) \;\to\; \xi(z;c) \tag{9}$$

[16] Theorem 7 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

$$m(-z;c) = \frac{-((1-c)+z) + \sqrt{((1-c)+z)^2 + 4cz}}{2cz}. \tag{8}$$

By direct calculation, (8) is indeed the unique positive solution to (24) when $m_\Psi(z) = (1-z)^{-1}$. While the eigenvalue distributions of the sample and true covariance matrices do not coincide, Theorem 7 describes the precise non-linear 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, $P^{-1}(P-T)z^{-1} = (1 - c^{-1})z^{-1}$.

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

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 the timing strategy (see equation (6)). One would imagine, then, that we need to know the limiting eigenvalue distribution of both matrices (or their Stieltjes transforms, m and m_Ψ) in order to describe $\operatorname{tr} E[(zI + \hat{\Psi})^{-1}\Psi]$. Proposition 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 Ψ.[18]

We refer to the constant c as "model complexity," which (as preceding results show) plays a critical role in understanding model behavior. It describes the limiting ratio of predictors to data points: $P/T \to c$. 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 Ψ 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 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 an abundance of 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

[18]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 non-linear.

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.

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

$$MSE(z;c) = \lim_{T,P\to\infty,\ P/T\to c} E\left[\left(R_{t+1} - S_t'\hat{\beta}(z)\right)^2 |\hat{\beta}(z)\right]. \tag{10}$$

Notably, while $\hat{\beta}(z)$ is random and depends on the sample realization, we show below that the limit in (10) 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*

$$\begin{aligned} \mathcal{E}(z;c) &= \lim\ E[\hat{\pi}_t R_{t+1} | \hat{\beta}(z)] = b_* \nu(z;c) \\ \mathcal{L}(z;c) &= \lim\ E[\hat{\pi}_t^2 | \hat{\beta}(z)] = b_* \hat{\nu}(z;c) - c\nu'(z;c) \\ R^2(z;c) &= \ \frac{2\mathcal{E}(z;c) - \mathcal{L}(z;c)}{1 + b_* \psi_{*,1}} \end{aligned} \tag{11}$$

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, we have

$$R^2(0,c) = R^2(0,0) \ - \ (1 \ + \ b_* \psi_{*,1})^{-1} \begin{cases} (c^{-1} - 1)^{-1}, & c < 1 \\ \mu(c), & c > 1. \end{cases} \tag{12}$$

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

$$\lim_{c\to\infty} R^2(0,c) = 0 \ > \ \lim_{c\to 1} R^2(0,c) = -\infty\,. \tag{13}$$

When the prediction model is complex ($c > 0$), the limiting eigenvalues of $\hat{\Psi}$ and Ψ diverge, and this unambiguously reduces the predictive R^2 relative to the infeasible best, $R^2(0,0)$. 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. In fact, 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 in the midst of complexity. Shrinkage controls variance but introduces bias. Proposition 3 points out that the amount of shrinkage that optimizes the bias-variance tradeoff is $z_* = c/b_*$. More complex settings benefit from heavier shrinkage, while setting 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, and 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 Ψ to the identity matrix and fix $b_* = 0.2$ (recall σ^2 is normalized to one). The upper 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 using the true parameter values) is the dotted red line and provides a point of reference.

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 $c = 1$, we see that predictive accuracy deteriorates rapidly as model complexity increases. This captures the well known

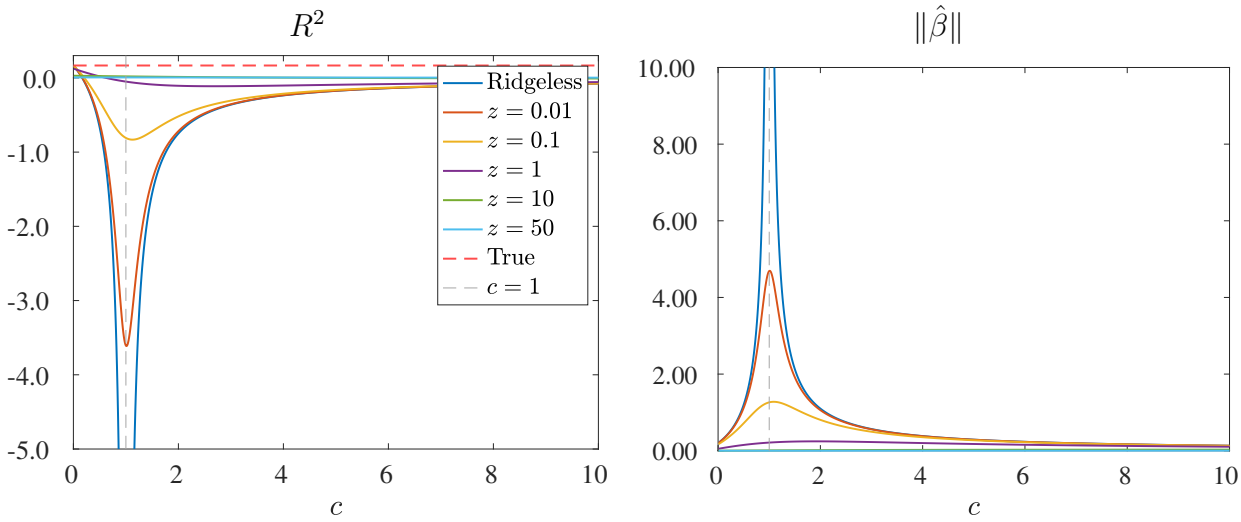

Figure 1: Expected Out-of-sample R^2 and Norm of Least Squares Coefficient

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

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 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.

The reason is that, while there are many equivalent β solutions that exactly fit[19] 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 a form of shrinkage, biasing the beta estimate toward zero. Due to this bias, the forecast variance drops, and this improves the R^2. In other words, despite $z \to 0$, the ridgeless

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

solution still regularizes the least squares estimator, and moreso 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.[20] This result challenges standard financial economics doctrine that places heavy emphasis on model parsimony. It shows that even in very simple data generating 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.

Our results regarding R^2 and MSE are similar to those in Hastie et al. (2019) and Richards et al. (2021), though we impose weaker technical conditions on X_t and Ψ (see Appendix C for a comparison of our theoretical approach versus prior literature). 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 (11) 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.

[20] See Spigler et al. (2019), Belkin et al. (2018), Belkin et al. (2019), Belkin et al. (2020), and Hastie et al. (2019).

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* (11). *As a result, the Sharpe ratio satisfies*

$$SR(z;c) = \frac{\mathcal{E}(z;c)}{\sqrt{\mathcal{V}(z;c)}} = \frac{1}{\sqrt{2 + (1 + b_*\psi_{*,1})\,\frac{\mathcal{L}(z;c)}{(\mathcal{E}(z;c))^2}}}\,. \tag{14}$$

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*

$$SR(z_*;c) = \frac{1}{\sqrt{2 + (1 + b_*\psi_{*,1})\,\frac{1}{b_*\nu(z_*;c)}}} < SR(0;0)\,, \tag{15}$$

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 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 despite the fact that 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.

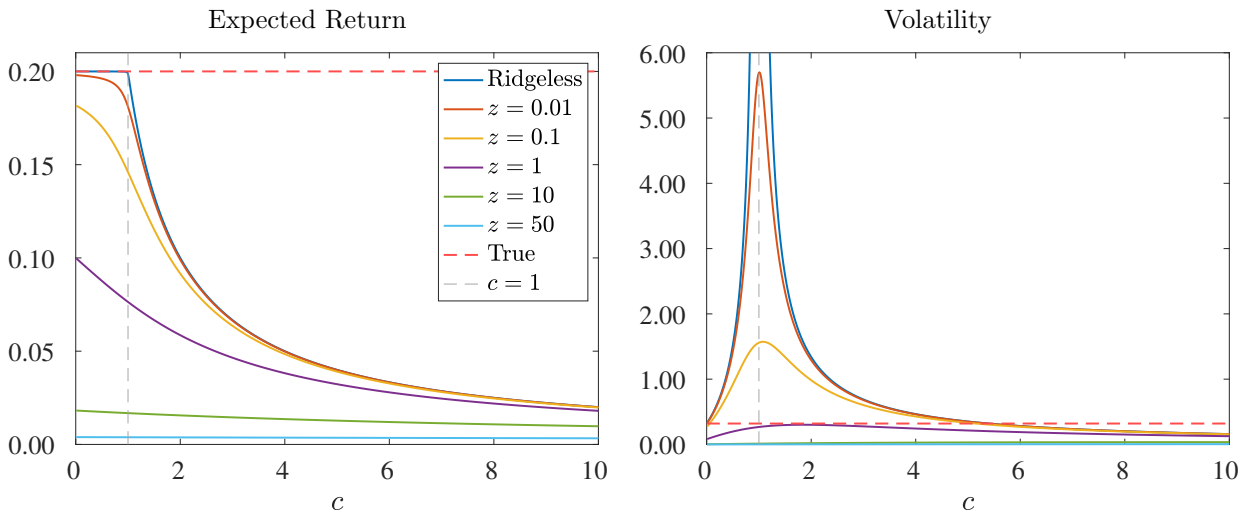

Figure 2: Expected Out-of-sample Risk and Return of Market Timing

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 Ψ is the identity matrix and $b_* = 0.2$.

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. The reason for 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 this attenuation is a reduction in the strategy's riskiness. For fairly 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

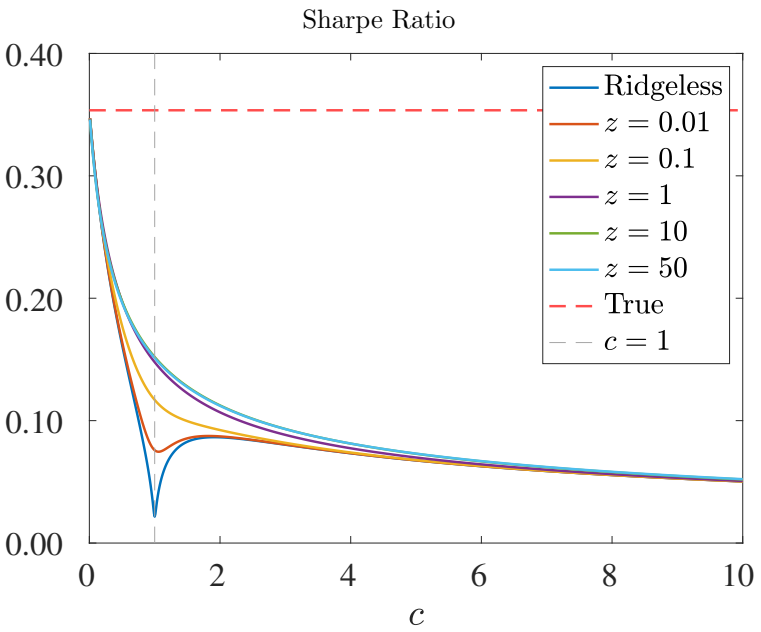

Figure 3: Expected Out-of-sample Sharpe Ratio of Market Timing

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

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 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 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 for 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 always prefers ridgeless regression despite its comparatively poor performance in terms of 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 timing strategies with negative R^2 can have high average out-of-sample returns, and thus positive out-of-sample Sharpe ratios.[21] More plainly, positivity of out-of-sample R^2 is *not* a necessary condition for an economically valuable timing strategy. In fact, the least squares timing strategies in our framework all have strictly positive out-of-sample expected return and Sharpe ratio regardless of the extent of shrinkage or model complexity (despite having enormously negative R^2 in many cases).

Much of the empirical literature in return prediction and market timing focuses its evaluations on out-of-sample predictive R^2 (see, e.g. Welch and Goyal, 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.

[21] 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 true Sharpe ratio of the DGP.

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 for cannot really 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 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. This correlation structure is captured by the off-diagonal blocks of Ψ.

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

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

$$\begin{aligned} \xi_{2,1}(z;cq;q) &= \lim_{T \to \infty} T^{-1} \operatorname{tr} E[(zI + \hat{\Psi}_{1,1})^{-1} \Psi_{1,2} \Psi_{1,2}'] \ \geq \ 0 \\ \widehat{\xi}_{2,1}(z;cq;q) &= \lim_{T \to \infty} T^{-1} \operatorname{tr} E[(zI + \hat{\Psi}_{1,1})^{-1} \Psi_{1,1} (zI + \hat{\Psi}_{1,1})^{-1} \Psi_{1,2} \Psi_{1,2}'] \ \geq \ 0 \, . \end{aligned} \tag{16}$$

While the existence of the limits in (16) cannot be guaranteed in general, the expectations are uniformly bounded for $z > 0$ (since so are the Ψ matrices). Hence, by passing to a subsequence of T, P, we can always assume the limits in (16) 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,

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

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*

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

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

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

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 (16). When both quantities (16) are zero, expressions significantly simplify. It is straightforward to show that both quantities in (16) are zero if the matrices $\Psi_{1,2}, \Psi_{2,1}$ have uniformly bounded traces. For example, this is the case 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* $\Psi_P = D_P + Q_P$ *where* $\limsup_{P\to\infty} \operatorname{rank} Q_P < \infty$, *while* D_P *are diagonal matrices, and* D_P, Q_P *are uniformly bounded. Then,* $\xi_{1,2} = \xi_{2,1} = 0$, *whereas all quantities in Proposition 5 coincide with those for* Ψ_P *replaced by* D_P. *Furthermore,*

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

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

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

$$\begin{aligned}
&\mathcal{E}(0; cq; c) = b_* q \left(\psi_{*,1}(q) - (cq)^{-2} m_*(cq; q)^{-1} \mathbf{1}_{q>1/c}\right) \\
&\mathcal{L}(0; cq; q) = \mathcal{E}(0; cq; c) + \left(1 + b_*(\psi_{*,1}(1) - q\psi_{*,1}(q))\right) \begin{cases} ((cq)^{-1} - 1)^{-1}, & q < 1/c \\ \tilde{\mu}(cq; c), & q > 1/c \end{cases} \\
&\mathcal{V}(0; cq; q) = 2(\mathcal{E}(0; cq; q))^2 + (1 + b_*\psi_{*,1})\, \mathcal{L}(0; cq; q) \\
&SR(0; cq; c) = \frac{\mathcal{E}(0; cq; c)}{\sqrt{\mathcal{V}(0; cq; c)}}
\end{aligned} \tag{18}$$

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

$$\mathcal{E}(0; cq; c) = b_* \psi_{*,1} \min\{q, c^{-1}\} \tag{19}$$

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 because there is not enough data to support the model's heavy parameterization. That analysis shows

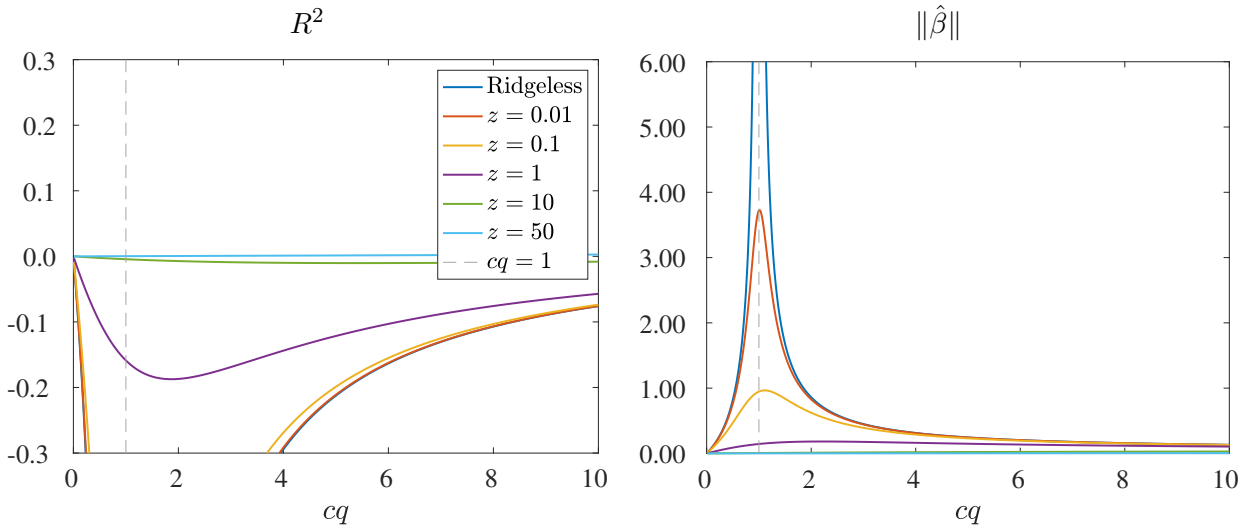

Figure 4: Expected Out-of-sample Prediction Accuracy From Mis-specified Models

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

that when models are correctly specified, the best performance (both 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.

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

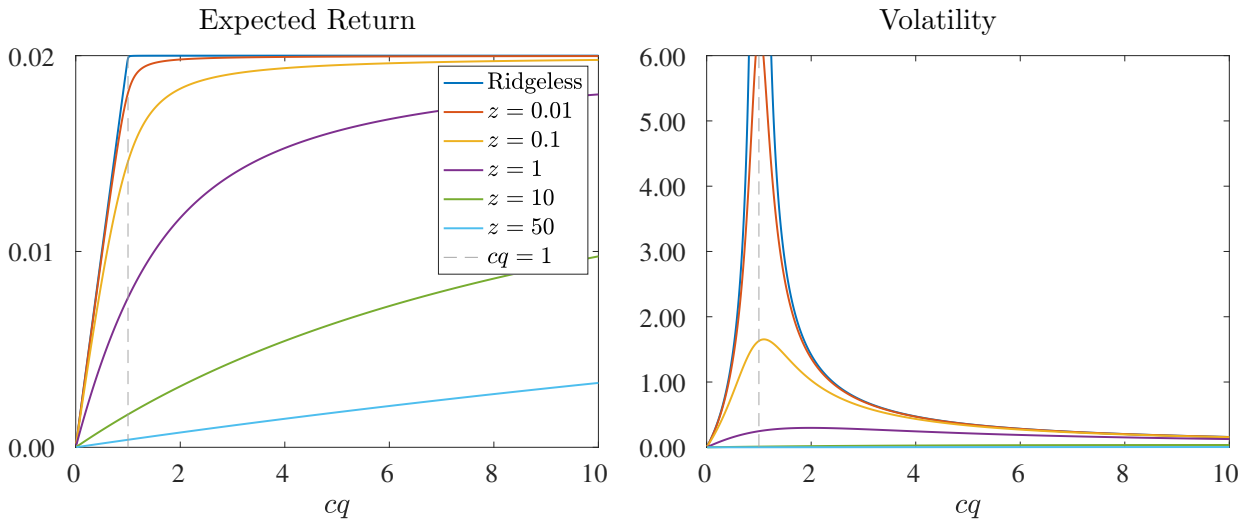

Figure 5: Expected Out-of-sample Timing Strategy Risk and Return From Mis-specified Models

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 Ψ is the identity matrix, $b_* = 0.2$, and the complexity of the true model is $c = 10$.

assumed to have a complexity of $c = 10$. We continue to calibrate Ψ 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 ($cq = 10$ and thus correctly specified). The left panel of Figures 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, 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 qualitatively patterns to those in Figure 1.

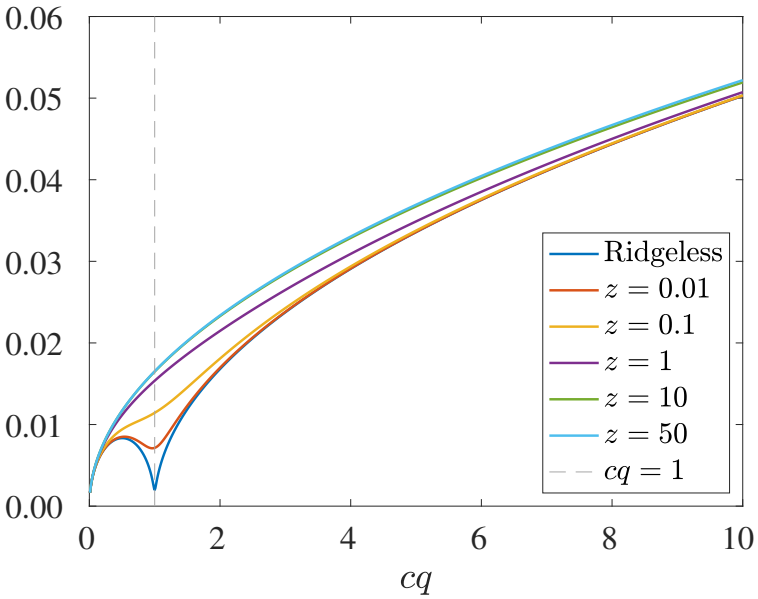

Figure 6: Expected Out-of-sample Timing Strategy Sharpe Ratio From Mis-specified Models

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

The most important difference versus Figure 1 is the pattern for 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 $\mathcal{E}(0;0)c^{-1} = b_*\psi_{*,1}c^{-1}$ 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 (19).

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 like we saw in the correctly specified setting, the expected return improvement from additional complexity leads the Sharpe ratio to also

increase with complexity. 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 performance of market timing in terms of both expected return and Sharpe ratio.

6 Virtue of Complexity: Empirical Evidence From Market Timing

In this section we present empirical analyses that are exact empirical analogues 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 Welch and Goyal (2008) that are available at the monthly frequency over the sample 1926–2020.[22]

We volatility standardize returns and predictors using backward-looking standard deviations that preserve the out-of-sample nature of our forecasts. Returns are standardized by their trailing 12-month return standard deviation (to capture their comparatively fast-moving conditional volatility), while 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 that we have enough stability in our initial predictor standardization, so the final sample that we bring to our analysis begins in 1930. We perform this standardization to align the empirical analysis with our homoskedastic theoretical

[22]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.

setting, but 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 (1). 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 a methodology from the machine learning literature known as random Fourier features, or RFF (Rahimi and Recht, 2007, 2008). Let G_t denote our 15×1 vector of predictors. The RFF methodology converts G_t into a pair of new signals

$$S_{i,t} = [\sin(\omega_i' G_t), \cos(\omega_i' G_t)]', \quad \omega_i \sim iidN(0, \gamma I). \tag{20}$$

$S_{i,t}$ uses the vector ω_i to form a random linear combination of G_t, which is then fed through the trigonometric functions.[23] 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 through the non-linear transformation in (20). If one desires a very low-dimensional model in (1), say $P = 2$, one can generate a single pair of random Fourier features. For a very high-dimensional model, say $P = 10{,}000$, one can instead draw many random weight vectors ω_i, $i = 1, ..., 5{,}000$. The larger the number of random features, the richer the approximation (1) provides to the general functional form $E[R_{t+1}|G_t] = f(G_t)$ where f is some smooth non-linear function. Indeed, the RFF approach is a wide two-layer neural network with fixed weights in the first layer (in the form of ω_i) and optimized weights in the second layer (in the form the regression estimates for β).

[23] Random features can be generated in a number of ways (for a survey see Liu et al., 2020). Our choice of functional form in (20) 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.

6.3 Complexity, Shrinkage, and Out-of-sample Market Timing

To conduct the empirical analogue of the theoretical analysis in Figure 4, 5, and 6, we consider a one-year rolling training window ($T = 12$) and a large set of random Fourier features (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 a training sample of a mere $T = 12$ observations 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 shockingly small samples. None of our conclusions are sensitive to the choice of training window (see robustness discussion below).

To draw plots 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 (20) with bandwidth parameter γ.[24]

(ii) Fix a model defined by the number of features $P \in \{2, ..., 12{,}000\}$ and a ridge shrinkage parameter $\log_{10}(z) \in \{-3, ..., 3\}$. The set of predictors S_t for regression (1) correspond to the first P RFFs from (i).

(iii) Given the model in (ii), conduct a recursive out-of-sample prediction and market timing strategy. For each $t \in \{12, ..., 1{,}091\}$, estimate (1) using training observations $\{(R_t, S_{t-1}), ..., (R_{t-11}, S_{t-12})$. From estimated regression coefficient, construct the out-of-sample return forecast $\hat{\beta}' S_t$ and the 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.

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

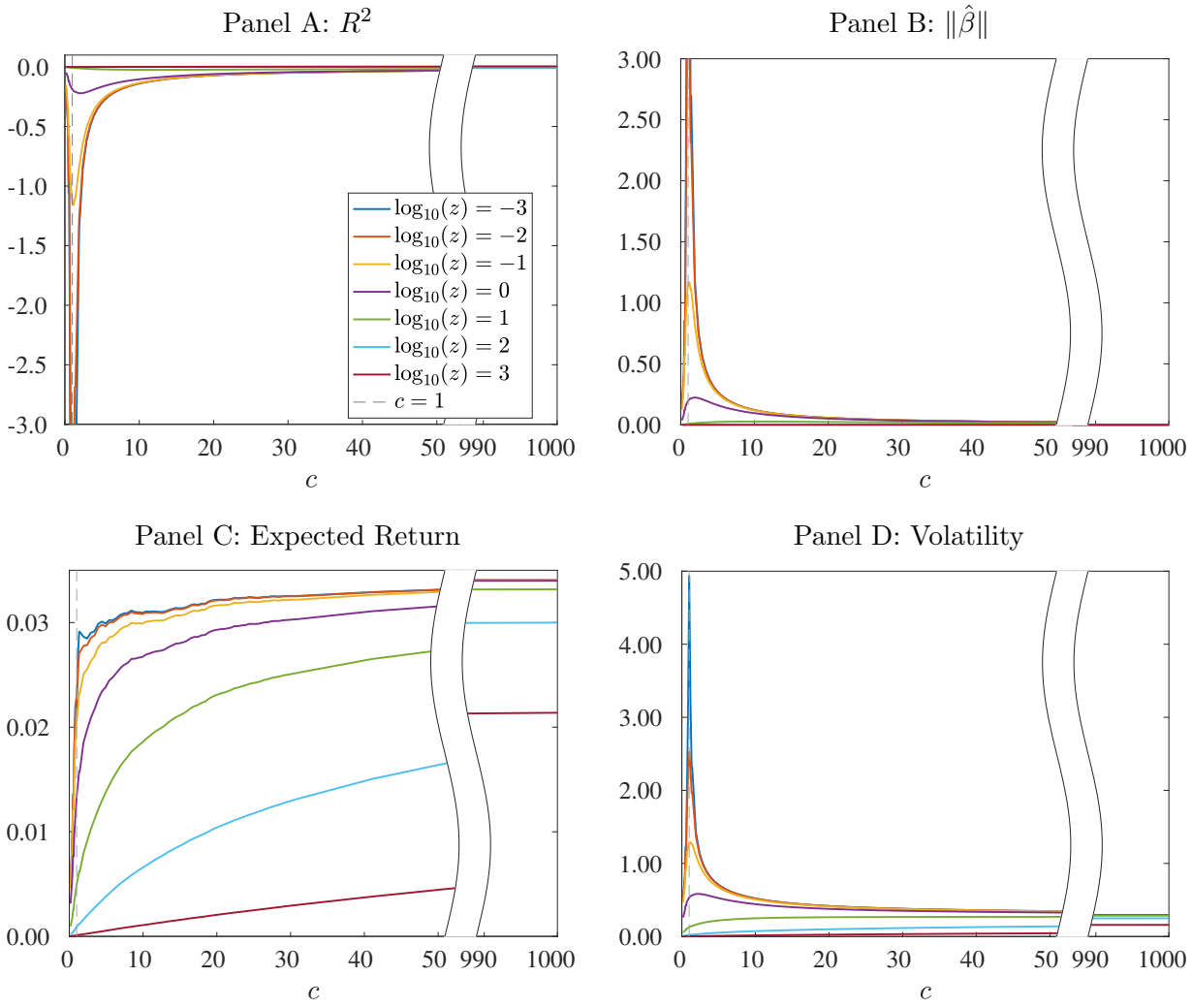

Figure 7: Out-of-sample Market Timing Performance

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 Welch and Goyal (2008) predictors with $\gamma = 2$.

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.

Figures 7 and 8 plot the out-of-sample prediction and market timing performance as a

function of model complexity and ridge shrinkage. The wide range of complexity that we consider, $c \in [0, 1000]$, can make it difficult to read plots. To better visualize the results while emphasizing both behavior near the interpolation boundary and behavior for extreme complexity, we break the x-axis between $c = 50$ and $c = 990$.

The first conclusion from these figures is that out-of-sample behavior of machine learning market return predictions is a strikingly close match to the patterns 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 characteristic recovery in the high complexity regime $c >> 1$.

By far 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 flat, just as predicted by equation (19) 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.

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, and 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

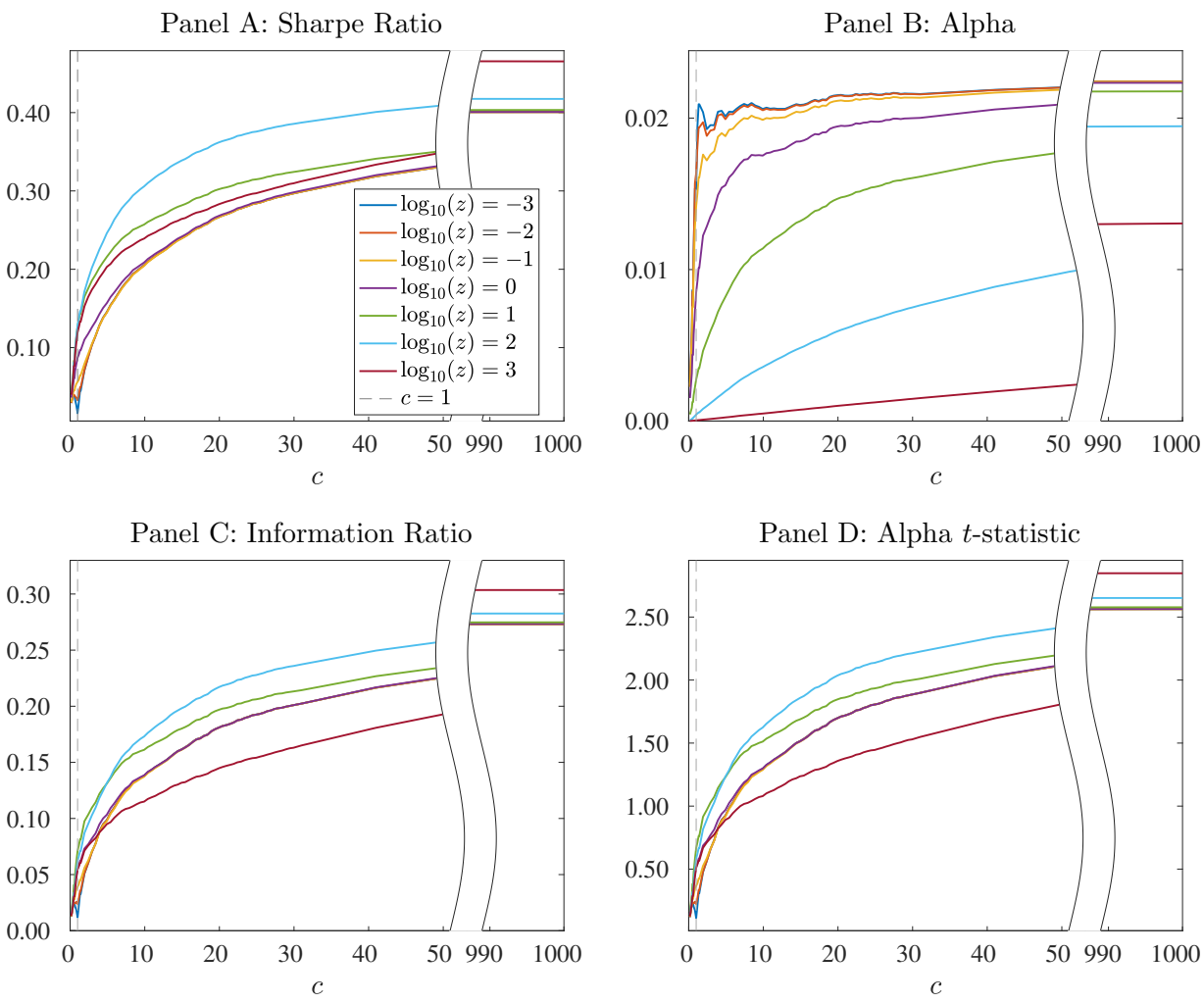

Figure 8: Out-of-sample Market Timing Performance

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 Welch and Goyal (2008) predictors with $\gamma = 2$. Alphas are versus a static position in the volatility-standardized market portfolio.

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.

Extreme behavior at the interpolation boundary makes it difficult to fully appreciate the

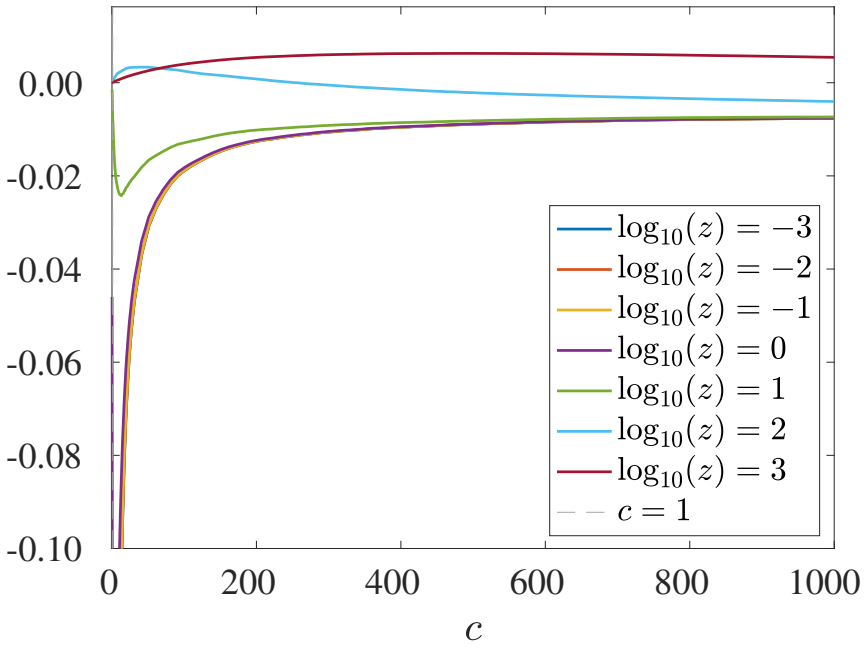

Figure 9: Out-of-Sample R^2 Detail

Note: Out-of-sample prediction accuracy for empirical analysis described in Section 6.3. Training window is $T = 12$ months and predictor count P (or cT) ranges from 2 to 12,000 using a range of P. Predictors are RFFs generated from 15 Welch and Goyal (2008) predictors with $\gamma = 2$

patterns in R^2. Figure 9 provides more detail by plotting the out-of-sample R^2 zooming-in on the range $[-10\%, 10\%]$. 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.

What do market timing strategies look like in the high complexity regime? Figure 10 plots $\hat{\pi}(z, c)$ for two empirical configurations. We show raw positions and six-month moving averages of the raw positions for better readability (our trading results are based on the raw positions and not the moving averages). The blue line corresponds to the the highest complexity and highest shrinkage configurations of our empirical model ($c = 1000$ and $z =$

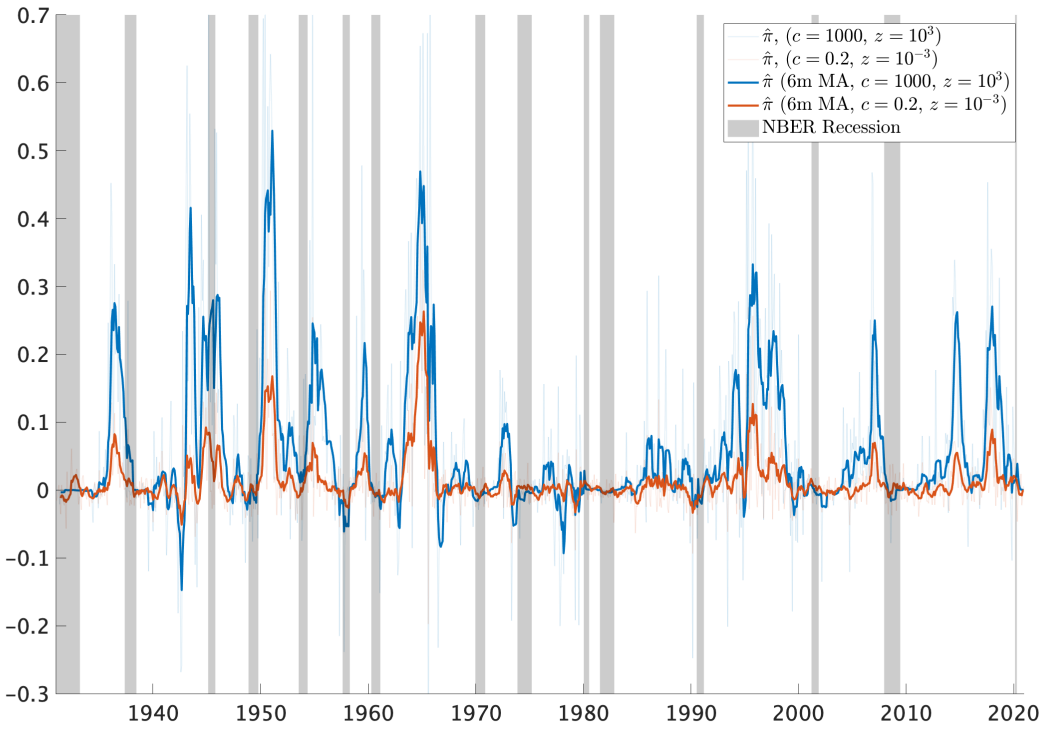

Figure 10: Market Timing Positions

Note: Out-of-sample market timing positions for empirical analysis described in Section 6.3. Training window is $T = 12$ months. Predictors are RFFs generated from 15 Welch and Goyal (2008) predictors with $\gamma = 2$. Blue lines show results for the $c = 1000$ and $z = 10^3$ model and red lines for the $c = 0.2$ and $z = 10^{-3}$ model (heavy lines show 6-month moving averages).

10^3, averaged over 1,000 RFF repetitions). The red line shows the lowest complexity and lowest shrinkage case we analyze ($c = 0.2$ and $z = 10^{-3}$). The basic timing patterns from these two models are representative of the results from other model configurations. The positions advocated by these two models have a time series correlation of 84.5% (87.2% for their moving average).

The timing positions in Figure 10 are remarkable. First, they show that the high complexity strategy is a long-only strategy at heart. Positions (or, equivalently, expected

market returns) from the machine learning models tend to be positive or zero. They almost never bet on a market downturn. 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 learns this rule without being given an explicit constraint!

Second, the machine learning strategy learns to divest leading up to recessions. NBER recession dates are shown in the gray shaded regions. For 14 out of 15 recessions in our test sample, the timing strategy essentially zeros out its position in the market prior to the recession (the exception is the eight-month recession of 1945). And it does this on a purely out-of-sample basis.

Figure 11 shows the robustness of our main findings in subsamples, splitting the test sample into halves. The left side of the figure reports machine learning timing strategy out-of-sample performance from 1930–1974, and the right side from 1975–2020. The figure shows 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, and 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 also about half as large in the second sample. This is consistent with the machine's trading patterns plotted in Figure 10. Starting around 1968,

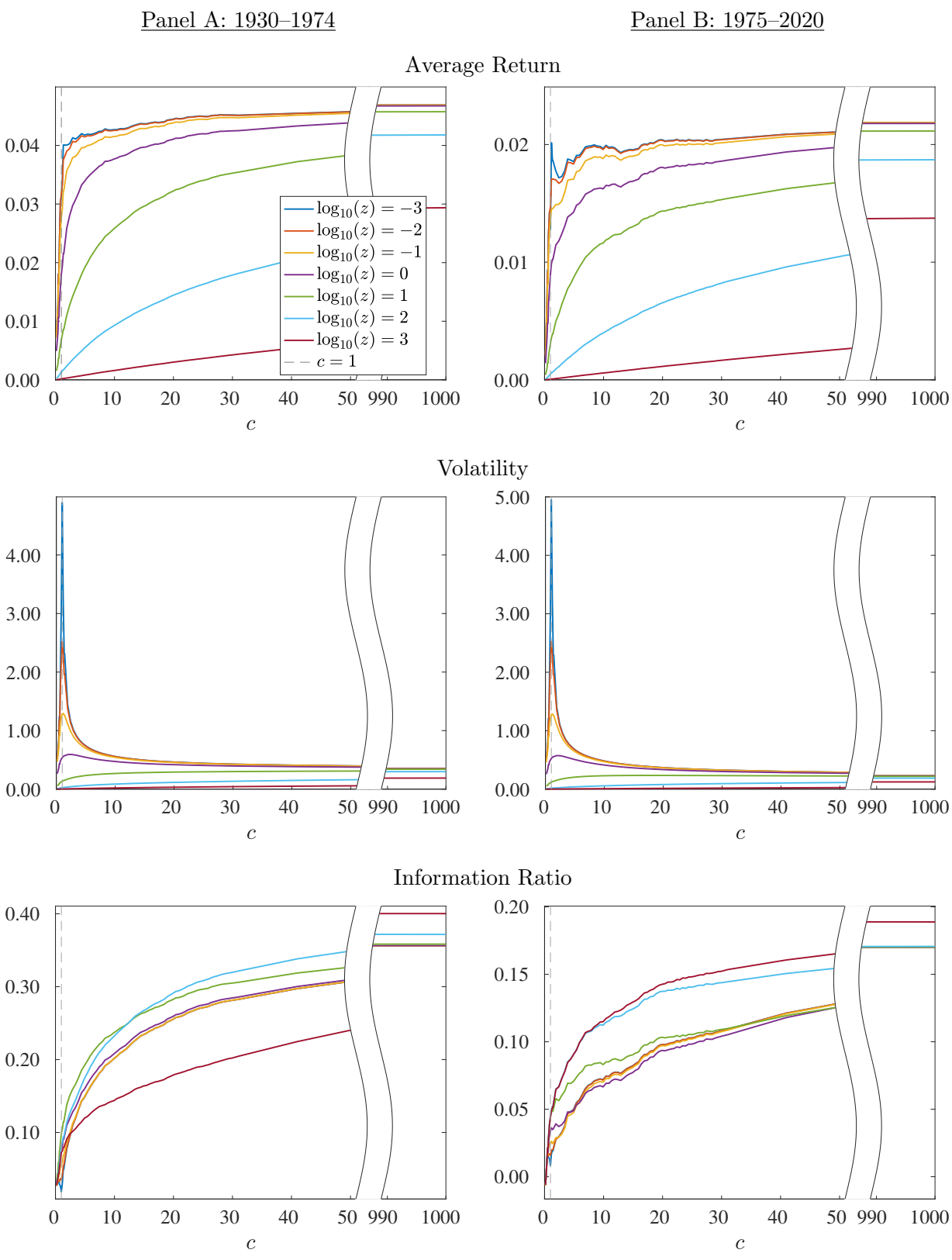

Figure 11: Out-of-sample Market Timing Performance by Subsample

Note: Subsample analysis of 1930–1974 and 1975–2020. See notes in Figures 7 and 8.

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

Our results seem at odds with the primary conclusion of Welch and Goyal (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 Welch and Goyal (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. 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. In particular, we use a rolling 12-month training window and forecast out-of-sample. We focus on a version of what Welch and Goyal (2008) call the "kitchen sink" regression. Our implementation of this uses all 15 monthly predictors in a simple, linear ridgeless regression.

Table 1 reports the results. To set the stage, we report summary stats of the buy-and-hold strategy in the first column.[25] The market return has a full sample Sharpe ratio of 0.50 per annum, a maximum one-month loss of -4.48 standard deviations,[26] and skewness of -0.41.

[25] More specifically, the first column reports summary statistics for the market return with rolling 12-month volatility standardization. Thus, the buy-and-hold version of this asset is itself a basic timing strategy, where timing is inversely proportional to rolling volatility. We do this simply because the standardized market is the target in our forecasting analysis. 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. As noted earlier, we prefer to use the volatility standardized market because it aligns more directly with our theoretical framework.

[26] Because returns are volatility-standardized using rolling 12-month standard deviation, the max loss is in monthly conditional standard deviation units.

Table 1: Comparison With Welch and Goyal (2008)

Note. Out-of-sample prediction accuracy and portfolio performance estimates for empirical analysis described in Section 6.3 and compared with simple models based on Welch and Goyal (2008).

		WG Kitchen Sink					High Complexity Machine		
	Market	$z = 0^+$	$z = 10^{-3}$	$z = 10$	$z = 10^3$	v.\Mkt		v.\Mkt	v.\WG
R^2	-	-97.64	-34.18	-0.10	-0.04	-	0.01	-	-
SR (IR)	0.50	-0.11	-0.12	0.29	0.46	0.33	0.47	0.31	0.26
t	4.74	-1.02	-1.11	2.74	4.37	3.06	4.46	2.89	2.47
Max Loss	-4.48	-98.49	-71.76	-4.66	-2.43	-2.70	-1.23	-1.14	-0.94
Skewness	-0.41	-0.86	-2.96	-1.11	-0.05	-0.04	2.48	2.29	1.97

The first finding of Table 1 is that we confirm the conclusions of Welch and Goyal (2008). Monthly return forecasts using the usual suspect predictors in ridgeless regression behave egregiously. The monthly out-of-sample R^2 from ridgeless regression ($z = 0^+$) is large and negative at -9764%. The timing strategy based on these predictions is also poor. The Sharpe ratio is -0.11 and is insignificantly different from zero. This seems not so terrible given the wildness of the forecasts, but it is due to the fact that the strategy's volatility is so high. It's maximum loss is 98 standard deviations. In light of our theoretical analysis, this agreement with the conclusions of Welch and Goyal (2008) is perhaps unsurprising. With $P = 15$ and $T = 12$, this analysis takes place close to the interpolation boundary, thus forecasts and timing strategy returns are expected to be highly volatile, as our estimates confirm. In Table 2, we repeat the same analysis as Table 1 but use a longer training window of five years. The conclusions are the same as those from Table 1.

Our theoretical analysis also suggests that, in circumstances like these, the benefits from additional ridge shrinkage are potentially large. Therefore, we re-estimate the Welch and Goyal (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

Table 2: Comparison With Welch and Goyal (2008) With 60-Month Training Window

Note. This table repeats the analysis of Table 1 using a 60-month training window.

		WG Kitchen Sink					High Complexity Machine		
	Market	$z = 0^+$	$z = 10^{-3}$	$z = 10$	$z = 10^3$	v.\Mkt		v.\Mkt	v.\WG
R^2	-	-0.97	-0.66	-0.01	0.00	-	0.00	-	-
SR (IR)	0.50	0.00	-0.02	0.49	0.44	0.10	0.42	0.25	0.27
t	4.74	0.00	-0.14	4.51	4.09	0.93	3.92	2.30	2.51
Max Loss	-4.48	-35.82	-25.51	-1.66	-1.38	-0.95	-0.46	-0.42	-0.43
Skewness	-0.41	-11.06	-8.45	-0.25	-0.30	-0.09	1.66	1.50	1.33

negative, as seen in the out-of-sample R^2 of -10% when $z = 10$. However, with this much shrinkage, the benefits of market timing become large. The annualized out-of-sample Sharpe ratio of the strategy is 0.29 and statistically significant ($t = 2.7$). Larger ridge shrinkage yields larger benefits still. When $z = 10^3$, the out-of-sample R^2 becomes -4% per month, while the annualized Sharpe ratio is 0.46 with a t-statistic of 4.4. This performance is not due to static market exposure. In the sixth column ("v. Mkt") we report performance after regressing the out-of-sample strategy from the fifth column on the market. This has an information ratio of 0.33 ($t = 3.1$). Also note that for the highly shrunken WG regression the maximum loss and skewness become more attractive.

These patterns align with the behavior predicted by our theoretical analysis. Near the interpolation boundary, models can seem defective in terms of R^2 despite shrinkage, yet they can nonetheless confer large economic benefits for investors. But much higher complexity models have further benefits yet. The last three columns of Table 1 show that the machine learning timing strategy (with $c = 1000$ and $z = 10^3$ in this example) further enhances out-of-sample performance. The average out-of-sample R^2 is 1% per month, and it has 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 heavily shrunken ($z = 10^3$) WG strategy. 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 timing strategy are not reflected in the Sharpe ratio but would be an important utility boost for investors that care about non-Gaussian risks. Note that the machine learning strategy accomplishes this using the identical information set as the WG strategy; it just exploits this information in a high-dimensional, non-linear way.

6.4 Robustness

In Appendix D, we report a number of robustness analyses around our main empirical results. We investigate the effect of longer estimation windows (120 months, versus 12 in our main analysis), different kernel bandwidths in RFF feature generation ($\gamma = 1$, versus $\gamma = 2$ in our main analysis), and 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.

7 Conclusion

The field of asset pricing is in the midst of a boom in research applications using machine learning. The asset management is experiencing a parallel boom in its adoption of 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 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 at play.

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 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 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 strategies that emerge have some remarkable attributes, behaving as long-only strategies that divest the market leading up to recessions. Our high complexity models learn this behavior with no guidance from researcher priors or modeling constraints.

Our results are *not* a license to add arbitrary predictors to a model—one cannot spin straw into gold. Instead, we encourage i) including all plausibly relevant predictors and ii) using rich non-linear models rather than simple linear specifications. Doing so confers

prediction and portfolio benefits, even when training data is scarce, and particularly when accompanied by prudent shrinkage.

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., 2018; 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, M, D Hsu, S Ma, and S Mandal, "Reconciling modern machine learning and the biasvariance trade-off. arXiv e-prints," 2018.

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

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

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.

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

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.

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

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

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

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

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.

Liu, Fanghui, Xiaolin Huang, Yudong Chen, and Johan AK Suykens, "Random Features for Kernel Approximation: A Survey on Algorithms, Theory, and Beyond," *arXiv preprint arXiv:2004.11154*, 2020.

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

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

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.

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.

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," *arXiv preprint arXiv:1506.02785*, 2015.

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

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

Wu, Denny and Ji Xu, "On the Optimal Weighted l2 Regularization in Overparameterized Linear Regression," *arXiv preprint arXiv:2006.05800*, 2020.

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

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 π_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 π_t portfolio, we have

$$E[\pi_t R_{t+1}] = E[(\beta' S_t)^2] = E[\beta' \Psi \beta] = \beta' \Psi \beta \tag{21}$$

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

$$\begin{aligned} &\sigma^4 E[(\pi_t)^2 R_{t+1}^2] = \sigma^4 E[(\pi_t)^2 E_t[R_{t+1}^2]] = E[((S_t'\beta)^2)^2(\sigma^2 + (S_t'\beta)^2)] \\ &= \sigma^2 \beta' \Psi \beta \ + \ E[(S_t'\beta)^4] = \sigma^2 \beta' \Psi \beta \ + \ E[(X_t'\tilde{\beta})^4] \\ &= \sigma^2 \beta' \Psi \beta \ + \ E[\sum_{i_1,i_2,i_3,i_4} X_{i_1} X_{i_2} X_{i_3} X_{i_4} \tilde{\beta}_{i_1} \tilde{\beta}_{i_2} \tilde{\beta}_{i_3} \tilde{\beta}_{i_4}] \end{aligned} \tag{22}$$

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

$$\sigma^4 E[(\pi_t)^2 R_{t+1}^2] = \sigma^2 \beta' \Psi \beta \ + \ 3(\beta' \Psi \beta)^2 \ + \ \sum_i (E[X_{i,t}^4] - 3)\tilde{\beta}_i^4 \tag{23}$$

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 Ψ.

Theorem 7 *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*

$$m(z;c) = \frac{1}{1 \;-\; c \;-\; c\, z\, m(z;c)}\, m_{\Psi}\left(\frac{z}{1 \;-\; c \;-\; c\, z\, m(z;c)}\right). \tag{24}$$

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

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

We will need an auxiliary

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

$$P^{-1} S_t' A_P S_t - P^{-1} \operatorname{tr}(A_P \Psi) \;\to\; 0 \tag{25}$$

is probability.

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

□

Lemma 3 *We have*

$$P^{-1} \operatorname{tr}(Q_P(zI + \hat{\Psi}_T)^{-1}) \;-\; E[P^{-1} \operatorname{tr}(Q_P(zI + \hat{\Psi}_T)^{-1})] \;\to\; 0 \tag{26}$$

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)),

$$(zI+\hat{\Psi}_T)^{-1} = (zI+\hat{\Psi}_{T,t})^{-1} - \frac{1}{T}(zI+\hat{\Psi}_{T,t})^{-1}S_tS_t'(zI+\hat{\Psi}_{T,t})^{-1}\frac{1}{1+(T)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t} \tag{27}$$

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

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

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

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

and therefore

$$\begin{aligned} q_T(z) - E[q_T(z)] &= \sum_{t=1}^{T}(E_{\tau-1}[q_T(z)] - E_T[q_T(z)]) \\ &= \frac{1}{M}\sum_{t=1}^{T}(E_{t-1}-E_t)[\operatorname{tr}(zI+\hat{\Psi}_T)^{-1}Q_P - \operatorname{tr}(zI+\Psi_{T,T}^{-1)}Q_P] \\ &= -\frac{1}{M}\sum_{\tau=1}^{T}(E_{t-1}-E_t)[\gamma_t]\,, \end{aligned} \tag{28}$$

where we have used (27) and defined

$$\gamma_t = \operatorname{tr}\left(\frac{1}{T}(zI+\hat{\Psi}_{T,t})^{-1}S_t(I+\frac{1}{T}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}Q_P\right) \tag{29}$$

We will need the following known properties of the trace:

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

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

and

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

Thus,

$$\begin{aligned}
\|\gamma_t\| \ &\leq \ \|Q_P\|\,\mathrm{tr}\left(\frac{1}{T}(zI+\hat{\Psi}_{T,t})^{-1}S_t(I+\frac{1}{T}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}\right) \\
&\leq \ z^{-1}\,\mathrm{tr}\left(\frac{1}{T}(zI+\hat{\Psi}_{T,t})^{-1/2}S_t(I+\frac{1}{T}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t)^{-1}S_t(zI+\hat{\Psi}_{T,t})^{-1/2}\right) \qquad (30) \\
&= \ z^{-1}\,\mathrm{tr}(B(zI+B)^{-1}) \ \leq \ Nz^{-1}\,,
\end{aligned}$$

where

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

Thus,

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

forms a bounded martingale difference sequence. Applying the Burkholder-Davis-Gundy

inequality (see, e.g., Burkholder (1966)), we get

$$\begin{aligned} E[|q_T(z) - E[q_T(z)]|^q] &\leq K_q P^{-q} E\left(\sum_{t=1}^{T} |(E_{t-1} - E_t)[\gamma_t]|^2\right)^{q/2} \\ &\leq K_q (2N/z)^q P^{-q/2} \left(\frac{P}{T}\right)^{-q/2} . \end{aligned} \tag{31}$$

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 several steps.

- Let

$$\hat{\Psi}_{T,t} = \frac{1}{T} \sum_{\tau \neq t} S_\tau S_\tau' . \tag{32}$$

Then, by the Sherman-Morrison formula (27),

$$\begin{aligned}(zI+\hat{\Psi}_T)^{-1}S_t &= (zI+\hat{\Psi}_{T,t})^{-1}S_t \\ &- \frac{1}{T}(zI+\hat{\Psi}_{T,t})^{-1}S_tS_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t\frac{1}{1+(T)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t} \\ &= (zI+\hat{\Psi}_{T,t})^{-1}S_t\frac{1}{1+(T)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}\,.\end{aligned} \tag{33}$$

- By Lemma 2,

$$P^{-1}\,S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t\;-P^{-1}\,\mathrm{tr}(\Psi(zI+\hat{\Psi}_{T,t})^{-1})\;\to\;0 \tag{34}$$

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

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

alsmost surely. Thus,

$$P^{-1}\,S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t\;-E[P^{-1}\,\mathrm{tr}(\Psi(zI+\hat{\Psi}_{T,t})^{-1})]\;\to\;0 \tag{35}$$

is probability.

- Theorem 7 implies that

$$P^{-1}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_T)^{-1}]\;\to\;m(-z;c) \tag{36}$$

- Now, we have

$$
\begin{aligned}
1 &= P^{-1}\operatorname{tr} E[(zI+\hat{\Psi}_T)^{-1}(zI+\hat{\Psi}_T)] = P^{-1}\operatorname{tr} E[(zI+\hat{\Psi}_T)^{-1}]z \;+\; P^{-1}\operatorname{tr} E[(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T] \\
&= zm(-z,c) \;+\; P^{-1}\operatorname{tr} E[(zI+\hat{\Psi}_T)^{-1}\frac{1}{T}\sum_t S_t S_t'] \\
&= \{symmetry\ across\ t\} \;=\; zm(-z,c) \;+\; P^{-1}\operatorname{tr} E[(zI+\hat{\Psi}_T)^{-1}\frac{1}{N}S_t S_t'] \\
&= \{using\ Sherman-Morrison\ (33)\} \\
&= zm(-z,c) \;+\; P^{-1}\operatorname{tr} E[(zI+\hat{\Psi}_{T,t})^{-1}S_t\frac{1}{1+(T)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}S_t'] \\
&= zm(-z,c) \;+\; E[\frac{P^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}{1+(T)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}]
\end{aligned}
\tag{37}
$$

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 (34),

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

in probability and this sequence is uniformly bounded. Hence,

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

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

$$\beta'\Psi\hat{\beta} \;\to\; b_*(\psi_{*,1} - c^{-1}z\xi(z)) \tag{38}$$

in probability, and then we establish the identity

$$\mathrm{tr}(\Psi\hat{\beta}\hat{\beta}') \to\; b_*(\psi_{*,1} - 2zc^{-1}\xi(z) - z^2c^{-1}\xi'(z)) + \xi(z) + z\xi'(z) \tag{39}$$

in probability. We start with the observation that

$$\frac{1}{T}\sum_{t=1}^{T} S_t R_{t+1} = \frac{1}{T}\sum_{t=1}^{T} S_t(S_t'\beta + \varepsilon_{t+1}) = \hat{\Psi}_T\beta \;+\; q_T\,, \tag{40}$$

where we have defined

$$q_T = \frac{1}{T}\sum_{t=1}^{T} S_t'\varepsilon_{t+1}\,. \tag{41}$$

Therefore,

$$\hat{\beta} = (zI + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T\beta \;+\; q_T) \tag{42}$$

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 (76) and Assumption 4, we have (using that ε_t are independent of S_t and have zero

means) that

$$
\begin{aligned}
&\beta' E[S_t S_t'] \hat{\beta} \\
&= \beta' E[S_t S_t'](zI + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T \beta \ + \ q_T) \\
&\approx \beta' \Psi \, (zI + \hat{\Psi}_T)^{-1} \hat{\Psi}_T \beta \\
&= \{by \ Lemma \ 1\}\} \\
&\overset{prob}{\to} \ b_* \, P^{-1} \operatorname{tr} E[\Psi \, (zI + \hat{\Psi}_T)^{-1} \hat{\Psi}_T] \\
&= b_* P^{-1} \operatorname{tr} E[\Psi \, (zI + \hat{\Psi}_T)^{-1}(zI + \hat{\Psi}_T - zI)] \\
&= b_* P^{-1} \operatorname{tr} E[\Psi - z\Psi \, (zI + \hat{\Psi}_T)^{-1}] \\
&= b_* P^{-1} \left(\operatorname{tr} \Psi \ - \ z \operatorname{tr} E[(zI + \hat{\Psi}_T)^{-1} \Psi] \right) \\
&= \{by \ Proposition \ 2\} \\
&\to_{T \to \infty} \ b_* \nu(z) \, .
\end{aligned}
\tag{43}
$$

At the same time,

$$
\begin{aligned}
&\operatorname{tr}(\Psi \hat{\beta} \hat{\beta}') \\
&= \operatorname{tr}(\Psi (zI + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T \beta \ + \ q_T)(\hat{\Psi}_T \beta \ + \ q_T)'(zI + \hat{\Psi}_T)^{-1}) \\
&= \operatorname{tr}(\Psi (zI + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T \beta \ + \ q_T)(\beta' \hat{\Psi}_T \ + \ q_T')(zI + \hat{\Psi}_T)^{-1}) \\
&\approx \operatorname{tr}(\Psi (zI + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T \beta \beta' \hat{\Psi}_T \ + \ q_T q_T')(zI + \hat{\Psi}_T)^{-1})
\end{aligned}
\tag{44}
$$

where we have used the fact that the terms that are linear in q_T converge to zero in

probability. Now,

$$
\begin{aligned}
E[q_T q_T'|S] &= \frac{1}{T^2} E[\sum_t S_t \varepsilon_{t+1} \sum_{t_1} \varepsilon_{t_1+1} S_{t_1}'|S] \\
&= \frac{1}{T^2} E[\sum_{t,t_1} S_t \varepsilon_{t+1} \varepsilon_{t_1+1}' S_{t_1}'|S] \\
&= \frac{1}{T^2} E[\sum_t S_t \varepsilon_{t+1} \varepsilon_{t+1}' S_t'|S] \\
&= \frac{1}{T^2} \sum_t S_t E[\varepsilon_{t+1} \varepsilon_{t+1}'|S] S_t' \\
&= \frac{1}{T^2} \sum_t S_t \sigma^2 S_t' \\
&= \frac{1}{T^2} \sum_t S_t S_t' = \frac{1}{T} \hat{\Psi}_T \,,
\end{aligned}
\tag{45}
$$

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, (44) takes the form

$$
\begin{aligned}
&\mathrm{tr}(\Psi\hat{\beta}\hat{\beta}') \\
&= \mathrm{tr}(\Psi(zI+\hat{\Psi}_T)^{-1}(\hat{\Psi}_T\beta\beta'\hat{\Psi}_T \;+\; q_T q_T')(zI+\hat{\Psi}_T)^{-1}) \\
&= \mathrm{tr}(\Psi(zI+\hat{\Psi}_T)^{-1}(\hat{\Psi}_T\beta\beta'\hat{\Psi}_T \;+\; \frac{1}{T}\hat{\Psi}_T)(zI+\hat{\Psi}_T)^{-1}) \\
&= \mathrm{tr}(\Psi(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T\beta\beta'\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}) \\
&+ \mathrm{tr}(\Psi(zI+\hat{\Psi}_T)^{-1}\frac{1}{T}\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}) \\
&= \mathrm{tr}(\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}\Psi(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T\beta\beta') \\
&+ \mathrm{tr}(\Psi(zI+\hat{\Psi}_T)^{-1}\frac{1}{T}(zI+\hat{\Psi}_T-zI)(zI+\hat{\Psi}_T)^{-1}) \\
&= \{by\ Lemmas\ 1,\ 3\ and\ Vitali's\ thorem\} \qquad (46) \\
&\overset{prob}{\to}\; b_* P^{-1}\,\mathrm{tr}\,E[\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}\Psi(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T] \\
&+ \frac{1}{T}\,\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}(zI+\hat{\Psi}_T)(zI+\hat{\Psi}_T)^{-1}]) \\
&- z\frac{1}{T}\,\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}(zI+\hat{\Psi}_T)^{-1}]) \\
&= b_* P^{-1}\,\mathrm{tr}\,E[\Psi(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}] \\
&+ \frac{1}{T}\,\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}]) \\
&- z\frac{1}{T}\,\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-2}]) \\
&= Term1 \;+\; Term2 \;+\; Term3\,.
\end{aligned}
$$

We now proceed with each term:

$$
\begin{aligned}
&(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1} \;= \{\text{all matrices commute}\} = (zI+\hat{\Psi}_T)^{-2}\hat{\Psi}_T^2 \\
&= \; (zI+\hat{\Psi}_T)^{-2}(\hat{\Psi}_T^2+2z\hat{\Psi}_T+z^2I-2z\hat{\Psi}_T-z^2I) \\
&= \; (zI+\hat{\Psi}_T)^{-2}(\hat{\Psi}_T^2+2z\hat{\Psi}_T+z^2I-2z(\hat{\Psi}_T+zI)+z^2I) \\
&= \; (zI+\hat{\Psi}_T)^{-2}\Big((zI+\hat{\Psi}_T)^2-2z(\hat{\Psi}_T+zI)+z^2I\Big) \\
&= \; I \;-\; 2z(zI+\hat{\Psi}_T)^{-1}+z^2(zI+\hat{\Psi}_T)^{-2}\,.
\end{aligned}
\tag{47}
$$

Therefore,

$$
\begin{aligned}
&Term1 = b_*P^{-1}\,\mathrm{tr}\,E[\Psi(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T\hat{\Psi}_T(zI+\hat{\Psi}_T)^{-1}] \\
&= \; b_*P^{-1}\,\mathrm{tr}\,E[\Psi(I \;-\; 2z(zI+\hat{\Psi}_T)^{-1}+z^2(zI+\hat{\Psi}_T)^{-2})]\,,
\end{aligned}
\tag{48}
$$

and

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

by Proposition 2, and hence

$$
\frac{d}{dz}\frac{1}{T}\,\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}]) \;\to\; \frac{d}{dz}\xi(z) \tag{49}
$$

by the Vitali theorem. However,

$$
\frac{d}{dz}\,\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-1}]) = -\,\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-2}]) \tag{50}
$$

and hence

$$
\frac{1}{T}\,\mathrm{tr}(\Psi E[(zI+\hat{\Psi}_T)^{-2}]) \;\to\; -\frac{d}{dz}\xi(z)\,. \tag{51}
$$

Summarizing, we get

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

whereas

$$Term1 \;\to\; b_* P^{-1}\,\mathrm{tr}\,E[\Psi(I-2z(zI+\hat{\Psi}_T)^{-1}+z^2(zI+\hat{\Psi}_T)^{-2})] \;\to\; b_*(\psi_{*,1}-2zc^{-1}\xi(z)-z^2c^{-1}\xi'(z)) \tag{52}$$

and hence

$$\begin{aligned}
&\mathrm{tr}(\Psi E[\hat{\beta}\hat{\beta}']) = \; Term1 \;+\; Term2 \;+\; Term3 \\
&\overset{prob}{\to} b_*(\psi_{*,1} - 2zc^{-1}\xi(z) - z^2c^{-1}\xi'(z)) + \xi(z) + z\frac{d}{dz}\xi(z) \\
&= \; b_*\,\hat{\nu}(z;c) \;-\; c\,\nu'(z;c)
\end{aligned} \tag{53}$$

Now, by (6), we have

$$MSE \;\to\; E[R_{t+1}^2] \;-\; 2\,E[\hat{\beta}' S_t S_t' \beta] \;+\; \mathrm{tr}\,E[\hat{\beta}\beta'\Psi] \tag{54}$$

and therefore equations (43) and (53) imply that

$$MSE \;\to\; E[R_{t+1}^2] \;-\; 2\,\mathcal{E}(z;c) + \;\mathcal{L}(z;c) \tag{55}$$

and hence

$$R^2(z;c) \;=\; 1 - \frac{MSE}{E[R_{t+1}^2]} \;\to\; \frac{2\,\mathcal{E}(z;c) + \;\mathcal{L}(z;c)}{E[R_{t+1}^2]} \tag{56}$$

whereas

$$\begin{aligned} 2\mathcal{E}(z;c) - \mathcal{L}(z;c) &= \lim P^{-1} \operatorname{tr}(2b_*\hat{\Psi}(zI+\hat{\Psi}) - b_*\hat{\Psi}^2 - c\hat{\Psi})(zI+\hat{\Psi})^{-2}\Psi) \\ &= \lim P^{-1} \operatorname{tr}(\hat{\Psi}(2b_*z - c) + b_*\hat{\Psi}^2)(zI+\hat{\Psi})^{-2}\Psi) \end{aligned} \tag{57}$$

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,*

$$m(z;c) = (1-c^{-1})z^{-1} + m_*(c) + n_*(c)z \ + \ O(z^2), \ z \to \ 0\,. \tag{58}$$

Furthermore,

$$\begin{aligned} m_*(c) &= c^{-1}((\sigma_*\psi_{*,1})^{-1}c^{-1} + \sigma_*^{-1}\psi_{*,2}\psi_{*,1}^{-3}c^{-2}) \ + \ O(c^{-4}) \\ n_*(c) &= c^{-1}\left(-(\sigma_*\psi_{*,1})^2c^2 + 3\sigma_*^2\psi_{*,2}c\right)^{-1} \ + \ O(c^{-5})\,. \end{aligned} \tag{59}$$

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

Case 1: $c > 1$ Substituting

$$\tilde{m}(-z;c) = (1-c)z^{-1} \ + \ cm(-z;c)\,, \tag{60}$$

into the equation of Theorem 7, 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

$$\frac{\int \frac{dH(x)}{\tilde{m}(1+\tilde{m}\,x)}}{\int \frac{x\,dH(x)}{(1+\tilde{m}\,x)}} = c-1\,. \tag{61}$$

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

$$\int \frac{dH(x)}{(ac^{-1}+bc^{-2})(1+(ac^{-1}+bc^{-2})\,x)} = (c-1)\int \frac{x\,dH(x)}{(1+(ac^{-1}+bc^{-2})\,x)} \tag{62}$$

that is (up to negligible terms)

$$\begin{aligned} & a^{-1}c\int (1-bc^{-1}/a+(bc^{-1}/a)^2)(1-(ac^{-1}+bc^{-2})\,x+(ac^{-1}+bc^{-1})^2\,x^2)dH(x) \\ &= \ (c-1)\int x(1-(ac^{-1}+bc^{-2})\,x+(ac^{-1}+bc^{-2})^2\,x^2)dH(x) \end{aligned} \tag{63}$$

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

$$m_*(c) = c^{-1} \tilde{m}_*(c) \underbrace{\sim}_{c \to \infty} c^{-1}((\sigma_* \psi_{*,1})^{-1} c^{-1} + \sigma_*^{-1} \psi_{*,2} \psi_{*,1}^{-3} c^{-2}) \tag{64}$$

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

$$\begin{aligned}(c-1)\int \frac{x^2\,dH(x)}{(1+\tilde{m}_*\,x)^2} \;&<\; (c-1)\int \frac{x\,dH(x)}{\tilde{m}_*(1+\tilde{m}_*\,x)} = \int \frac{dH(x)}{\tilde{m}_*^2(1+\tilde{m}_*\,x)} \\ = \int \frac{(1+\tilde{m}_*\,x)dH(x)}{\tilde{m}_*^2(1+\tilde{m}_*\,x)^2} \;&<\; \int \frac{(1+2\tilde{m}_*\,x)dH(x)}{\tilde{m}_*^2(1+\tilde{m}_*\,x)^2}\end{aligned} \tag{65}$$

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,

$$\begin{aligned} cn_*(c) \;&\sim\; \left(-a^{-2}c^2\int \frac{(1+2(ac^{-1}+bc^{-2})x)dH(x)}{((1+bc^{-1}/a)(1+(ac^{-1}+bc^{-2})\,x))^2} + (c-1)\int \frac{x^2\,dH(x)}{(1+(ac^{-1}+bc^{-2})\,x)^2}\right)^{-1} \\ &\sim\; \left(-a^{-2}c^2\int (1+2ac^{-1}x - 2bc^{-1}/a - 2ac^{-1}x)dH(x) + c\int x^2(1-2ac^{-1}x)\,dH(x)\right)^{-1} \\ &\sim\; \left(-a^{-2}c^2 + (2a^{-3}b + \sigma_*^2\psi_{*,2})c\right)^{-1} \\ &=\; \left(-(\sigma_*\psi_{*,1})^2c^2 + (2(\sigma_*\psi_{*,1})^3\sigma_*^{-1}\psi_{*,2}/\psi_{*,1}^3 + \sigma_*^2\psi_{*,2})c\right)^{-1} \\ &=\; \left(-(\sigma_*\psi_{*,1})^2c^2 + 3\sigma_*^2\psi_{*,2}c\right)^{-1} \end{aligned} \tag{66}$$

when $c \to \infty$. □

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

$$\begin{aligned}
\xi(z) &= -1 + c^{-1}/(c^{-1} - 1 + zm(-z;c)) \\
&= -1 + c^{-1}/(zm_*(c) + z^2 n_*(c) + O(z^3)) \\
&= -1 + c^{-1}(zm_*(c))^{-1}/(1 + zn_*(c)/m_*(c) + O(z^2)) \\
&= -1 + c^{-1}(zm_*(c))^{-1}(1 - zn_*(c)/m_*(c) + O(z^2)) \\
&= -1 + c^{-1}(zm_*(c))^{-1} \ - \ c^{-1}n_*(c)m_*(c)^{-2} \ + \ O(z)
\end{aligned} \tag{67}$$

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*

$$1 - zm(z) = \psi_{*,1}az^{-1} - z^{-2}(\psi_{*,2} + c\psi_{*,1}^2)a^2 + z^{-3}a^3(\psi_{*,3} + 3c\psi_{*,2}\psi_{*,1} + c^2\psi_{*,1}^3) \ + \ O(z^{-4}) \tag{68}$$

for $z \to \infty$.

Proof of Lemma 7. Then, Theorem 7 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

$$\begin{aligned}
&1-zm(-z)-\psi_{*,1}az^{-1}\\
&=\; (1-c+czm(z))\int \frac{xdH(x)}{x(1-c+czm(-z))+z} - \psi_{*,1}az^{-1}\\
&=\; (1-cz^{-1}a\psi_{*,1}+O(z^{-2}))z^{-1}\int \frac{xdH(x)}{xz^{-1}(1-cz^{-1}a\psi_{*,1}+O(z^{-2}))+1} - \psi_{*,1}az^{-1}\\
&\sim\; (1-cz^{-1}a\psi_{*,1}+O(z^{-2}))z^{-1}\int \frac{xdH(x)}{xz^{-1}+1} - \psi_{*,1}az^{-1} \qquad (69)\\
&\sim\; (1-cz^{-1}a\psi_{*,1}+O(z^{-2}))z^{-1}\int (x-x^2z^{-1})dH(x) - \psi_{*,1}az^{-1}\\
&\sim\; z^{-1}\psi_{*,1}a-\psi_{*,2}a^2z^{-2}-cz^{-2}a^2\psi_{*,1}^2-\psi_{*,1}az^{-1} \;+\; O(z^{-3})\\
&=\; -z^{-2}(\psi_{*,2}+c\psi_{*,1}^2)a^2 \;+\; O(z^{-3})
\end{aligned}$$

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

$$
\begin{aligned}
&1 - zm(-z) - \psi_{*,1} a z^{-1} + z^{-2}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \\
&= \ (1 - cz^{-1}(\psi_{*,1}a - z^{-1}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \ + \ O(z^{-2}))) \\
&\times \int \frac{x dH(x)}{x(1 - cz^{-1}(\psi_{*,1}a - z^{-1}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \ + \ O(z^{-2}))) + z} - \psi_{*,1} a z^{-1} + z^{-2}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \\
&= \ (1 - cz^{-1}(\psi_{*,1}a - z^{-1}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \ + \ O(z^{-2}))) \\
&\times z^{-1} \int \frac{x dH(x)}{xz^{-1}(1 - cz^{-1}\psi_{*,1}a) + 1 + O(z^{-3})} - \psi_{*,1} a z^{-1} + z^{-2}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \\
&\sim \ (1 - cz^{-1}(\psi_{*,1}a - z^{-1}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \ + \ O(z^{-2})))z^{-1} \int x(1 - xz^{-1}(1 - cz^{-1}\psi_{*,1}a) + x^2 z^{-2}) \\
&- \psi_{*,1} a z^{-1} + z^{-2}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \\
&\sim \ (1 - cz^{-1}(\psi_{*,1}a - z^{-1}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \ + \ O(z^{-2})))z^{-1} \left(\psi_{*,1}a - z^{-1}\psi_{*,2}a^2 + z^{-2}a^3(\psi_{*,3} + c\psi_{*,2}\psi_{*,1}) \right) \\
&- \psi_{*,1} a z^{-1} + z^{-2}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \\
&= \ \psi_{*,1} a z^{-1} - z^{-2}\psi_{*,2}a^2 + z^{-3}a^3(\psi_{*,3} + c\psi_{*,2}\psi_{*,1}) \\
&- cz^{-2}\psi_{*,1}a(\psi_{*,1}a - z^{-1}\psi_{*,2}a^2) + cz^{-3}(\psi_{*,2} + c\psi_{*,1}^2)a^2\psi_{*,1}a \ + \ O(z^{-4}) - \psi_{*,1} a z^{-1} + z^{-2}(\psi_{*,2} + c\psi_{*,1}^2)a^2 \\
&= \ z^{-3}a^3(\psi_{*,3} + c\psi_{*,2}\psi_{*,1}) \\
&- cz^{-2}\psi_{*,1}a(-z^{-1}\psi_{*,2}a^2) + cz^{-3}(\psi_{*,2} + c\psi_{*,1}^2)a^2\psi_{*,1}a \ + \ O(z^{-4}) \\
&= \ z^{-3}a^3(\psi_{*,3} + 3c\psi_{*,2}\psi_{*,1} + c^2\psi_{*,1}^3) \ + \ O(z^{-4}) \, .
\end{aligned}
\tag{70}
$$

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,

$$\begin{aligned} MSE &= E[\|R_{t+1} - S_{t,1}\hat{\beta}\|^2] \\ &= \text{tr}\, E[R_{t+1}R_{t+1}'] \;-\; 2E[\beta' S_t' S_{t,1}\hat{\beta}_1] \;+\; \text{tr}\, E[S_{t,1}\hat{\beta}_1\hat{\beta}_1' S_{t,1}'] \\ &= \text{tr}\, E[R_{t+1}R_{t+1}'] \;-\; 2E[\beta' S_t' S_{t,1}\hat{\beta}_1] \;+\; \text{tr}\, E[\Psi_{1,1}\hat{\beta}_1\hat{\beta}_1'] \end{aligned} \tag{71}$$

where $\hat{\beta}_1$ is the estimate of the first component of the whole β 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 $\xi_{1,1}(z) = \lim T^{-1}\,\text{tr}\, E[(zI + \hat{\Psi})^{-1}\Psi]$ to denote $\xi(z; cq; q)$.

The following is true.

Lemma 8 *We have*

$$\begin{aligned} \mathcal{E}(z; c_1) &= \lim E[\beta' S_t' S_{t,1}\hat{\beta}_1] \\ &= b_* \frac{c_1}{c}(\psi_{*,1} - c_1^{-1} z\xi(z)) \;+\; b_* \frac{c^{-1}\xi_{2,1}(z)}{1 + \xi_{1,1}(z)} \end{aligned} \tag{72}$$

where

$$\xi_{2,1}(z) = \lim_{T\to\infty} \frac{1}{T}\,\text{tr}\, E[\Psi_{1,2}\Psi_{2,1}(zI + \hat{\Psi}_{T,1,t})^{-1}] \tag{73}$$

Proof of Lemma 8. We have

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

and

$$\frac{1}{T}\sum_t S_{t,1}' R_{t+1} = \frac{1}{T}\sum_{t=1}^{T} S_{t,1}'(S_t\beta + \varepsilon_{t+1}) = \;\hat{\Psi}_T\beta \;+\; q_T\,, \tag{74}$$

where

$$q_T = \frac{1}{T}\sum_{t=1}^{T} S'_{t,1}\varepsilon_{t+1} \tag{75}$$

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,

$$\hat{\beta} = (zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_{T,1}\beta_1 + \hat{\Psi}_{T,2}\beta_2\ +\ q_T)\,. \tag{76}$$

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

$$\begin{aligned}
&E[\beta' S'_t S_{t,1}\hat{\beta}] \\
&=\ E[(\beta'_1\Psi_{1,1} + \beta'_2\Psi_{2,1})(zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_{T,1}\beta_1 + \hat{\Psi}_{T,2}\beta_2)] \\
&=\ \mathrm{tr}\, E[\Psi_{1,1}\,(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_T\beta\beta'] \\
&+\ E[\beta'_1\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\beta_2] \\
&+\ E[\beta'_2\Psi_{2,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}\beta_1] \\
&+\ E[\beta'_2\Psi_{2,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}\beta_2] \\
&= \{by\ Lemma\ 1\} \\
&\overset{prob}{\to}\ b_*\, P^{-1}\,\mathrm{tr}\, E[\Psi_{1,1}\,(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}]\ +\ P^{-1}b_*\,\mathrm{tr}\, E[\Psi_{2,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}]\,.
\end{aligned} \tag{77}$$

The first term is

$$
\begin{aligned}
&P^{-1}\operatorname{tr} E[\Psi_{1,1}\,(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}] \\
&= P^{-1}\operatorname{tr} E[\Psi_{1,1}\,(zI+\hat{\Psi}_{T,1})^{-1}(zI+\hat{\Psi}_{T,1}-zI)] \;\;\rightarrow\;\; \frac{c_1}{c}(\psi_{*,1}-c_1^{-1}z\xi_{1,1}(z))\,.
\end{aligned}
\tag{78}
$$

To compute the second term in (77), we will need the following lemma.

Lemma 9

$$
\frac{1}{P}\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}] \;\rightarrow\; c^{-1}\xi_{2,1}(z)/(1+\xi_{1,1}(z)) \tag{79}
$$

Proof of Lemma 9. We have that, by symmetry over time, and using the Sherman-Morrison formula (27), we get

$$
\begin{aligned}
&\frac{1}{P}\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2}] \\
&= \frac{1}{P}\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1})^{-1}\frac{1}{T}\sum_{t=1}^{T}S_{t,1}S_{t,2}'] \\
&= \frac{1}{P}\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1})^{-1}S_{t,1}S_{t,2}'] \\
&= \frac{1}{P}\operatorname{tr} E[\Psi_{2,1}\Big((zI+\hat{\Psi}_{T,1,t})^{-1} \\
&- \frac{1}{T}(zI+\hat{\Psi}_{T,1,t})^{-1}S_{t,1}(I+\frac{1}{T}S_{t,1}'(zI+\hat{\Psi}_{T,1,t})^{-1}S_{t,1})^{-1}S_{t,1}'(zI+\hat{\Psi}_{T,1,t})^{-1}\Big)S_{t,1}S_{t,2}'] \\
&= \frac{1}{P}\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S_{t,1}S_{t,2}'] \\
&- \frac{1}{P}\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S_{t,1}(I+C_T)^{-1}C_TS_{t,2}'] \\
&= \frac{1}{P}\operatorname{tr} E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t})^{-1}\Psi_{2,1}'] \\
&- \frac{1}{P}E[S_{t,2}'\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S_{t,1}(1+C_T)^{-1}C_T]
\end{aligned}
\tag{80}
$$

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 (34),

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

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

By a similar argument,

$$\frac{1}{T} S'_{t,2} \Psi_{2,1}(zI + \hat{\Psi}_{T,1,t})^{-1} S_{t,1} \to \xi_{2,1}(z) \tag{81}$$

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,

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

The proof of Lemma 9 is complete. □

Lemma 8 follows now from (77) □

Lemma 11 *We have*

$$\begin{aligned}
&\mathcal{L}(z) = \lim \operatorname{tr}(\Psi_{1,1} E[\hat{\beta}\hat{\beta}']) \\
&= \frac{c_1}{c} b_*(\psi_{*,1}(q) - 2zc_1^{-1}\xi_{1,1}(z) - z^2 c_1^{-1}\xi'_{1,1}(z)) \;+\; (1 + b_* P^{-1} \operatorname{tr} \Psi_{2,2})(\xi_{1,1}(z) + z\xi'_{1,1}(z)) \\
&+ b_*(1+\xi(z))^{-2} c_1^{-1} \widehat{\xi}_{2,1} \\
&- 2b_*(\xi_{1,1}(z) + z\xi'_{1,1}(z))(1+\xi_{1,1}(z))^{-1} c_1^{-1} \xi_{2,1}(z)
\end{aligned} \tag{83}$$

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

$$\begin{aligned}
&\operatorname{tr}(\Psi_{1,1} E[\hat{\beta}\hat{\beta}']) \\
&= \operatorname{tr}(\Psi_{1,1} E[(zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T(1,:)\beta \;+\; q_T)(\hat{\Psi}_T(1,:)\beta \;+\; q_T)'(zI + \hat{\Psi}_{T,1})^{-1}]) \\
&= \operatorname{tr}(\Psi_{1,1} E[(zI + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T(1,:)\beta \;+\; q_T)(\beta'\hat{\Psi}_T(1,:)' \;+\; q'_T)(zI + \hat{\Psi}_{T,1})^{-1}]) \\
&= \operatorname{tr}(\Psi_{1,1} E[(zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T(1,:)\beta\beta'\hat{\Psi}_T(1,:)' \;+\; q_T q'_T)(zI + \hat{\Psi}_{T,1})^{-1}]) \\
&= \operatorname{tr}(\Psi_{1,1} E[(zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T(1,:)\beta\beta'\hat{\Psi}_T(1,:)' \;+\; q_T q'_T)(zI + \hat{\Psi}_{T,1})^{-1}])
\end{aligned} \tag{84}$$

Formula (45) still holds with $\hat{\Psi}$ replaced by $\hat{\Psi}_{1,1}$ and calculations in (46) imply

$$\operatorname{tr}(\Psi_{1,1} E[(zI + \hat{\Psi}_{T,1})^{-1} q_T q'_T (zI + \hat{\Psi}_{T,1})^{-1}]) \;\to\; \xi_{1,1}(z) + z\xi'_{1,1}(z)\,. \tag{85}$$

It remains to deal with

$$
\begin{aligned}
&\operatorname{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T(1,:)\beta\beta\hat{\Psi}_T(1,:))(zI+\hat{\Psi}_{T,1})^{-1}]) \\
&= \operatorname{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_{T,1}\beta_1\beta_1\hat{\Psi}_{T,1})(zI+\hat{\Psi}_{T,1})^{-1}]) \\
&+ \operatorname{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_{T,1,2}\beta_2\beta_2\hat{\Psi}_{T,1,2})(zI+\hat{\Psi}_{T,1})^{-1}]) \\
&\stackrel{prob}{\to} P^{-1}b_*\operatorname{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}^2(zI+\hat{\Psi}_{T,1})^{-1}]) \\
&+ P^{-1}b_*\operatorname{tr}E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1,2}\hat{\Psi}_{T,1,2}(zI+\hat{\Psi}_{T,1})^{-1}]
\end{aligned}
\tag{86}
$$

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

$$
P^{-1}b_*\operatorname{tr}(\Psi_{1,1}E[(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}^2(zI+\hat{\Psi}_{T,1})^{-1}]) \;\to\; \frac{c_1}{c}b_*(\psi_{*,1}(q)-2zc_1^{-1}\xi_{1,1}(z)-z^2c_1^{-1}\xi_{1,1}(z))\,. \tag{87}
$$

Thus, it remains to deal with the second term in (86). We have

$$
\begin{aligned}
&P^{-1}b_*\operatorname{tr}E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1,2}\hat{\Psi}_{T,1,2}(zI+\hat{\Psi}_{T,1})^{-1}] \\
&= P^{-1}b_*\operatorname{tr}E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1,2}\hat{\Psi}_{T,1,2}(zI+\hat{\Psi}_{T,1})^{-1}] \\
&= P^{-1}b_*\frac{1}{T^2}\operatorname{tr}E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\sum_{t_1,t_2}S_{t_1,1}S'_{t_1,2}S_{t_2,2}S'_{t_2,1}(zI+\hat{\Psi}_{T,1})^{-1}] \\
&= P^{-1}b_*\frac{1}{T^2}\operatorname{tr}E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\Big(TS_{t_1,1}S'_{t_1,2}S_{t_1,2}S'_{t_1,1} \\
&+ T(T-1)\,S_{t_1,1}S'_{t_1,2}S_{t_2,2}S'_{t_2,1}\Big)(zI+\hat{\Psi}_{T,1})^{-1}] \\
&= Term1 \;+\; Term2\,.
\end{aligned}
\tag{88}
$$

Here,

$$
Term1 = P^{-1}b_*\frac{1}{T}\operatorname{tr}E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}S_{t_1,1}S'_{t_1,2}S_{t_1,2}S'_{t_1,1}(zI+\hat{\Psi}_{T,1})^{-1}] \tag{89}
$$

and

$$Term2 = (1 - T^{-1})P^{-1}b_* \operatorname{tr} E[\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1}S_{t_1,1}S'_{t_1,2}S_{t_2,2}S'_{t_2,1}(zI + \hat{\Psi}_{T,1})^{-1}]\,. \tag{90}$$

Using the Sherman-Morrison formula (27) and defining $C_T = S'_{t,1}(zI + \hat{\Psi}_{T,1,t})^{-1}S_{t,1}$, we get

$$(zI + \hat{\Psi}_{T,1})^{-1}S_{t_1,1} = (zI + \hat{\Psi}_{T,1,t})^{-1}S_{t_1,1}(1 + C_T)^{-1}\,, \tag{91}$$

and therefore

$$\begin{aligned} &Term1 \\ &= P^{-1}b_*\frac{1}{T}\operatorname{tr} E[\Psi_{1,1}(zI + \hat{\Psi}_{T,1,t})^{-1}S_{t_1,1}(1 + C_T)^{-1} \\ &\times S'_{t_1,2}S_{t_1,2}(1 + C_T)^{-1}S'_{t_1,1}(zI + \hat{\Psi}_{T,1,t})^{-1}] \\ &= P^{-1}b_*\frac{1}{T}\operatorname{tr} E[S'_{t_1,1}(zI + \hat{\Psi}_{T,1,t})^{-1}\Psi_{1,1}(zI + \hat{\Psi}_{T,1,t})^{-1}S_{t_1,1}S'_{t_1,2}S_{t_1,2}(1 + C_T)^{-2}] \end{aligned} \tag{92}$$

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

$$\frac{1}{T}S'_{t,1}(zI+\hat{\Psi}_{T,1,t})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S_{t,1} \;\to\; \frac{1}{T}E[\operatorname{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1})] \tag{93}$$

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

$$Term1 \to b_*\widehat{\xi}_{1,1}(z)P^{-1}\operatorname{tr}\Psi_{2,2}/(1 + \xi_{1,1}(z))^2 \tag{94}$$

where we have defined

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

We will now need the following lemma.

Lemma 12 *We have*

$$\frac{1}{T} E[\mathrm{tr}(\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1})] \;\to\; \widehat{\xi}_{1,1}(z) = (\xi_{1,1}(z)+z\xi'_{1,1}(z))(1+\xi_{1,1}(z))^2 \quad (95)$$

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 (9) 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

$$\begin{aligned} &\frac{1}{T} S'_{t_1,1}(zI + \hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI + \hat{\Psi}_{T,1,t})^{-1}S'_{t_1,1} \\ &- \frac{1}{T}\,\mathrm{tr}\, E[\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1}] \;\to\; 0 \end{aligned} \quad (96)$$

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 (91) and (96),

$$
\begin{aligned}
\xi_{1,1}(z) \;&\approx\; \frac{1}{T}\,\mathrm{tr}\,E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}] \\
&=\; \frac{1}{T}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_{T,1})(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}] \\
&\approx\; -z\xi_{1,1}(z) \;+\; \frac{1}{T}\,\mathrm{tr}\,E[\hat{\Psi}_{T,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}] \\
&= \{\hat{\Psi}_{T,1} = T^{-1}\sum_t S_{t,1}S'_{t,1}\} \\
&=\; -z\xi'_{1,1}(z) \;+\; \frac{1}{T^2}\sum_t \mathrm{tr}\,E[S_{t,1}S'_{t,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}] \\
&=\; -z\xi'_{1,1}(z) \;+\; \frac{1}{T}\,\mathrm{tr}\,E[S_{t,1}S'_{t,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1})^{-1}] \\
&=\; -z\xi'_{1,1}(z) \;+\; \frac{1}{T}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_{T,1})^{-1}S_{t,1}S'_{t,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}] \\
&=\; -z\xi'_{1,1}(z) \\
&+\; \frac{1}{T}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_{T,1,t})^{-1}S_{t_1,1}(1+C_T)^{-1}S'_{t_1,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}] \\
&=\; -z\xi'_{1,1}(z) \\
&+\; \frac{1}{T}\,\mathrm{tr}\,E[(1+C_T)^{-2}S'_{t_1,1}(zI+\hat{\Psi}_{T,1})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t})^{-1}S_{t_1,1}] \\
&\approx\; -z\xi'_{1,1}(z) \;+\; (1+\xi_{1,1}(z))^{-2}\widehat{\xi}_{1,1}(z)
\end{aligned}
\tag{97}
$$

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

Thus, it remains to deal with $Term2$ in (88). By (91),

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

(98)

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,

$$E[S'_{t_2,1} Q S_{t_2,2} | S_{t_2}] = \operatorname{tr}(Q\Psi_{2,1}) \tag{99}$$

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

$$\begin{aligned} Term21 &= P^{-1} b_* \operatorname{tr} E[(1+C_T)^{-1} S'_{t_2,1} \\ &\times \left((zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \Psi_{1,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \right) S_{t_1,1} (1+C_T)^{-1} S'_{t_1,2} S_{t_2,2}] \\ &= P^{-1} b_* \operatorname{tr} E[(1+C_T)^{-2} S'_{t_2,1} Q S_{t_2,2}] \\ &= b_* E[(1+C_T)^{-2} P^{-1} \operatorname{tr}(Q\Psi_{2,1})] \,, \end{aligned} \tag{100}$$

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

$$\begin{aligned} P^{-1} \operatorname{tr}(Q\Psi_{2,1}) &= P^{-1} \operatorname{tr} \left((zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \Psi_{1,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_1,1} S'_{t_1,2} \Psi_{2,1} \right) \\ &= P^{-1} \operatorname{tr} \left(S'_{t_1,2} \Psi_{2,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \Psi_{1,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_1,1} \right) \\ &\overset{prob}{\to} P^{-1} \operatorname{tr} E[\Psi_{1,2} \Psi_{2,1} \left((zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \Psi_{1,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \right)] \,, \end{aligned} \tag{101}$$

where, as we explain in the main text, we pass to a subsequence if necessary to ensure the

limit exists. Thus, by (9),

$$Term21 \to b_*(1+\xi(z))^{-2}c_1^{-1}\widehat{\xi}_{2,1}\,. \tag{102}$$

Proceeding to the next term in (98), we get

$$\begin{aligned}
&Term22\\
&= P^{-1}b_*\,\mathrm{tr}\,E[(1+C_T)^{-1}S'_{t_2,1}\\
&\times\left(-\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}(1+C_T)^{-1}S'_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)\\
&S_{t_1,1}(1+C_T)^{-1}S'_{t_1,2}S_{t_2,2}]
\end{aligned} \tag{103}$$

We have

$$\begin{aligned}
&\frac{1}{T}S'_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}\\
&\to \frac{1}{T}\,\mathrm{tr}\,E[\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}]\\
&= \widehat{\xi}_{1,1}(z)
\end{aligned} \tag{104}$$

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

$$\begin{aligned}
&Term22\\
&\to P^{-1}b_*\,\mathrm{tr}\,E[(1+C_T)^{-1}S'_{t_2,1}\\
&\times\left(-(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}(1+C_T)^{-1}\widehat{\xi}_{1,1}(z)\right)(1+C_T)^{-1}S_{t_1,2}S'_{t_2,2}]\\
&\to -b_*\widehat{\xi}_{1,1}(z)(1+\xi_{1,1}(z))^{-3}\,\mathrm{tr}\,E[\Psi_{2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}S'_{t_1,2}]\\
&\to -b_*\widehat{\xi}_{1,1}(z)(1+\xi_{1,1}(z))^{-3}c_1^{-1}\xi_{2,1}(z)\,,
\end{aligned} \tag{105}$$

where we have used Lemma 10 to pass to the limit.[27] Proceeding to the next term in (98), we get

$$
\begin{aligned}
Term23 \;\approx\; & P^{-1} b_* E[(1+C_T)^{-1} S'_{t_2,1} \\
& \left(- (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \Psi_{1,1} \frac{1}{T} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_2,1} (1+C_T)^{-1} S'_{t_2,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \right) \\
& S_{t_1,1} (1+C_T)^{-1} S'_{t_1,2} S_{t_2,2}] \\
& = -b_* E[X_T Y_T]
\end{aligned}
\tag{106}
$$

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

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

Integrating over S_{t_2} gives

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

[27] 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.

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 (98) satisfies.

$$Term23 \;\to\; -b_*\widehat{\xi}_{1,1}(z)(1+\xi_{1,1}(z))^{-3}c_1^{-1}\xi_{2,1}(z)\,. \tag{107}$$

Finally, the last term in (98) is given by

$$\begin{aligned}
&Term24 = P^{-1}b_*\,\mathrm{tr}\,E[(1+C_T)^{-1}S'_{t_2,1}\\
&\times\left(\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}(1+C_T)^{-1}S'_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\right.\\
&\left.\times\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_2,1}(1+C_T)^{-1}S'_{t_2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)\\
&\times S_{t_1,1}(1+C_T)^{-1}S'_{t_1,2}S_{t_2,2}]\\
&=\; P^{-1}b_*\,\mathrm{tr}\,E[(1+C_T)^{-4}S_{t_2,2}S'_{t_2,1}\\
&\times\left(\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}S'_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\right.\\
&\left.\times\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_2,1}S'_{t_2,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)\\
&S_{t_1,1}S'_{t_1,2}]
\end{aligned} \tag{108}$$

We will need the following

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,

$$\begin{aligned} & E[S_{t,2}S_{t,1}'ZS_{t,1}S_{t,1}'] \\ &= \Psi_{2,1}(Z+Z)\Psi_{1,1} + (Q_2\,\mathrm{diag}((E[X^4]-3)Q_1ZQ_1)Q_2 + \mathrm{tr}(Z\Psi_{1,1})\Psi_{2,1}) \end{aligned} \tag{109}$$

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 $S_{t,k}' = X_t'Q_k$. Then,

$$E[S_{t,2}S_{t,1}'\beta\gamma'S_{t,1}S_{t,1}'] = E[Q_2\,X_tX_t'Q_1\beta\gamma'Q_1'X_tX_t'Q_1] \tag{110}$$

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

$$\begin{aligned} E[X_tX_t'\tilde\beta\tilde\gamma'X_tX_t']_{k_1,k_2} &= E[\sum_{l_1,l_2} X_{k_1}X_{l_1}\tilde\beta_{l_1}\tilde\gamma_{l_2}X_{l_2}X_{k_2}] \\ &= E[X_{k_1}^2X_{k_2}^2](\tilde\beta_{k_1}\tilde\gamma_{k_2} + \tilde\beta_{k_2}\tilde\gamma_{k_1}) + \sum_\ell \tilde\beta_\ell\tilde\gamma_\ell E[X_{k_1}X_{k_2}X_\ell^2] \\ &= \tilde\beta_{k_1}\tilde\gamma_{k_2} + \tilde\beta_{k_2}\tilde\gamma_{k_1} \end{aligned} \tag{111}$$

At the same time,

$$\begin{aligned} E[X_tX_t'\tilde\beta\tilde\gamma'X_tX_t']_{k_1,k_1} &= E[\sum_{l_1,l_2} X_{k_1}^2X_{l_1}\tilde\beta_{l_1}\tilde\gamma_{l_2}X_{l_2}] \\ &= \sum_\ell \tilde\beta_\ell\tilde\gamma_\ell E[X_{k_1}^2X_\ell^2] \\ &= \tilde\beta_{k_1}\tilde\gamma_{k_1}(E[X_{k_1}^4]-1) \ + \ \tilde\beta'\tilde\gamma \end{aligned} \tag{112}$$

Summarizing,

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

Thus, by formula (110), we get

$$E[S_{t,2}' S_{t,1} \beta\gamma' S_{t,1}' S_{t,1}] = Q_2' Q_1 (\beta\gamma' + \gamma\beta') Q_1' Q_1 + (Q_2' \,\text{diag}((E[X^4] - 3)\tilde{\beta}_{k_1}\tilde{\gamma}_{k_1}) Q_2 + (\tilde{\beta}'\tilde{\gamma}) Q_2' Q_1)\,, \tag{113}$$

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

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

Thus,

$$\Psi = \begin{pmatrix} Q_1' Q_1 & Q_1' Q_2 \\ Q_2' Q_1 & Q_2' Q_2 \end{pmatrix} = \begin{pmatrix} \Psi_{1,1} & \Psi_{1,2} \\ \Psi_{2,1} & \Psi_{2,2} \end{pmatrix} \tag{114}$$

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}[28]. Define

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

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

Then, we can rewrite (108) as

$$\begin{aligned}
Term24 &= P^{-1} b_* \operatorname{tr} E[(1+C_T)^{-4} S_{t_2,2} S'_{t_2,1} \\
&\times \left(\frac{1}{T} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_1,1} S'_{t_1,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \Psi_{1,1} \right. \\
&\left. \times \frac{1}{T} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_2,1} S'_{t_2,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} \right) \\
&S_{t_1,1} S'_{t_1,2}] \\
&= P^{-1} b_* \operatorname{tr} E[(1+C_T)^{-4} E[S_{t_2,2} S'_{t_2,1} Z S_{t_2,1} S'_{t_2,1} | S_{t_1}] (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_1,1} S'_{t_1,2}]
\end{aligned} \tag{115}$$

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 (115), we get that everything reduces to computing two expectations:[29]

$$Expectation1 = P^{-1} \operatorname{tr} E[\Psi_{2,1} Z \Psi_{1,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_1,1} S'_{t_1,2}] \tag{116}$$

and

$$Expectation2 = P^{-1} \operatorname{tr} E[\Psi_{2,1} \operatorname{tr}(Z\Psi_{1,1}) (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_1,1} S'_{t_1,2}] \tag{117}$$

For $Expectation2$, we have

$$\begin{aligned}
Expectation2 &= P^{-1} \operatorname{tr} E[\Psi_{2,1} \operatorname{tr}(Z\Psi_{1,1}) (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_1,1} S'_{t_1,2}] \\
&= P^{-1} \operatorname{tr} E[S'_{t_1,2} \Psi_{2,1} (zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1} S_{t_1,1} \operatorname{tr}(Z\Psi_{1,1})] .
\end{aligned} \tag{118}$$

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

We know that the quantities

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

and

$$\begin{aligned}&\frac{1}{T}\operatorname{tr}\left((zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}S'_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)\\ &= \frac{1}{T}\operatorname{tr}\left(S'_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}\right)\end{aligned} \tag{119}$$

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

$$\begin{aligned}&Expectation1 = P^{-1}\operatorname{tr}E[\Psi_{2,1}Z\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}S'_{t_1,2}]\\ &= P^{-1}\operatorname{tr}E[S_{t_1,2}S'_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}Z'\Psi_{1,2}]\\ &= P^{-1}\operatorname{tr}E[S_{t_1,2}S'_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\\ &\times\left((zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S_{t_1,1}S'_{t_1,1}\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\right)\Psi_{1,2}]\end{aligned} \tag{120}$$

We can now once again apply Lemma 13 and get

$$\begin{aligned}&E[S'_{t_1,2}S_{t_1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\\ &\times(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}S'_{t_1,1}S_{t_1,1}]\\ &= \Psi_{2,1}(\hat{Z}+\hat{Z}')\Psi_{1,1} \;+\; \operatorname{tr}(\hat{Z}\Psi_{1,1})\Psi_{2,1}\end{aligned} \tag{121}$$

where

$$\hat{Z} = (zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,1}\frac{1}{T}(zI+\hat{\Psi}_{T,1,t_1,t_2})^{-1} \tag{122}$$

Therefore,

$$Expectation1 = P^{-1} \operatorname{tr} E[\left(\Psi_{2,1}(\hat{Z} + \hat{Z}')\Psi_{1,1} \ + \ \operatorname{tr}(\hat{Z}\Psi_{1,1})\Psi_{2,1} \right) \frac{1}{T}(zI + \hat{\Psi}_{T,1,t_1,t_2})^{-1}\Psi_{1,2}] \tag{123}$$

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

$$\begin{aligned}
&\operatorname{tr}(\Psi_{1,1}E[\hat{\beta}\hat{\beta}']) \\
&= \ \operatorname{tr}(\Psi_{1,1}E[(zI + \hat{\Psi}_{T,1})^{-1}(\hat{\Psi}_T\beta\beta'\hat{\Psi}_T \ + \ q_T q_T')(zI + \hat{\Psi}_{T,1})^{-1}]) \\
&= \operatorname{tr}(\Psi_{1,1}E[(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_T(1,:)\beta\beta'\hat{\Psi}_T(1,:)'(zI + \hat{\Psi}_{T,1})^{-1}]) \ + \ \xi_{1,1}(z) + z\xi_{1,1}'(z) \\
&\overset{prob}{\to} \ P^{-1}b_* \operatorname{tr}(\Psi_{1,1}E[(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,1}^2(zI + \hat{\Psi}_{T,1})^{-1}]) \\
&+ \ P^{-1}b_* \operatorname{tr} E[\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2,1}\hat{\Psi}_{T,2,1}'(zI + \hat{\Psi}_{T,1})^{-1}] + \ \xi_{1,1}(z) + z\xi_{1,1}'(z) \\
&= \ \frac{c_1}{c}b_*(\psi_{*,1}(q) - 2zc_1^{-1}\xi_{1,1}(z) - z^2c_1^{-1}\xi_{1,1}'(z)) \ + \ \xi_{1,1}(z) + z\xi_{1,1}'(z) \\
&+ \ P^{-1}b_* \operatorname{tr} E[\Psi_{1,1}(zI + \hat{\Psi}_{T,1})^{-1}\hat{\Psi}_{T,2,1}\hat{\Psi}_{T,2,1}'(zI + \hat{\Psi}_{T,1})^{-1}] \\
&= \ \frac{c_1}{c}b_*(\psi_{*,1}(q) - 2zc_1^{-1}\xi_{1,1}(z) - z^2c_1^{-1}\xi_{1,1}'(z)) \ + \ (1 + b_*P^{-1} \operatorname{tr} \Psi_{2,2})(\xi_{1,1}(z) + z\xi_{1,1}'(z)) \\
&+ Term21 + Term22 + Term23 + Term24 \\
&\to \ \frac{c_1}{c}b_*(\psi_{*,1}(q) - 2zc_1^{-1}\xi_{1,1}(z) - z^2c_1^{-1}\xi_{1,1}'(z)) \ + \ (1 + b_*P^{-1} \operatorname{tr} \Psi_{2,2})(\xi_{1,1}(z) + z\xi_{1,1}'(z)) \\
&+ b_*(1 + \xi(z))^{-2}c_1^{-1}\widehat{\xi_{2,1}} - 2b_*(\xi_{1,1}(z) + z\xi_{1,1}'(z))(1 + \xi_{1,1}(z))^{-1}c_1^{-1}\xi_{2,1}(z)
\end{aligned} \tag{124}$$

The proof of Lemma 11 is complete.

□

C Discussion of Related Literature

Formulas of Propositions 2 and 3 have been established in papers on random matrix theory, with Proposition 2 going back to Ledoit and Péché (2011). Hastie et al. (2019) prove an analog of Proposition 3 allowing for arbitrary β and expressing all quantities in terms of the distribution of projections of β onto the eigenvectors of Ψ (see also Wu and Xu, 2020). Furthermore, they establish non-asymptotic bounds on the rate of convegence. However, both Hastie et al. (2019) and Wu and Xu (2020) require that Ψ is strictly positive definite. By contrast, in our data analysis, we find that Ψ is close to degenerate. Richards et al. (2021) also allow for more general β structures and Ψ 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 random Fourier features used in our empirical analysis (or any other non-linear signal transformations). In contrast to these papers, we establish our results under much weaker conditions on the distribution of $X_{i,t}$ across i: The Lindenberg condition of Assumption 2. This is important for practical applications, where neither the independence of X_t nor equality (or boundedness) of their higher moments can be guaranteed.

D Additional Exhibits

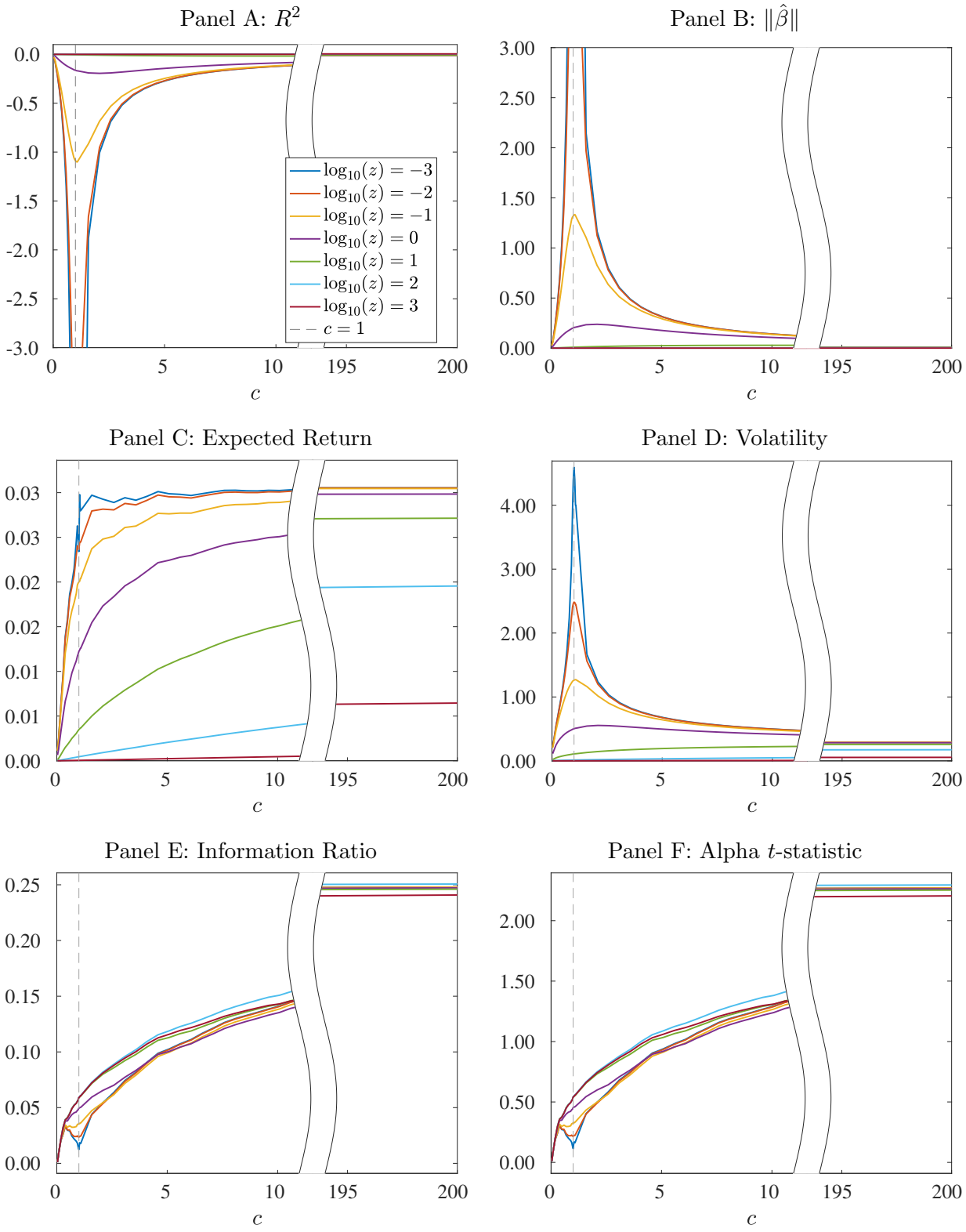

Figure 12: Out-of-sample Market Timing Performance With 60-month Training Window

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 using a range of P. Predictors are RFFs generated from 15 Welch and Goyal (2008) predictors with $\gamma = 2$.

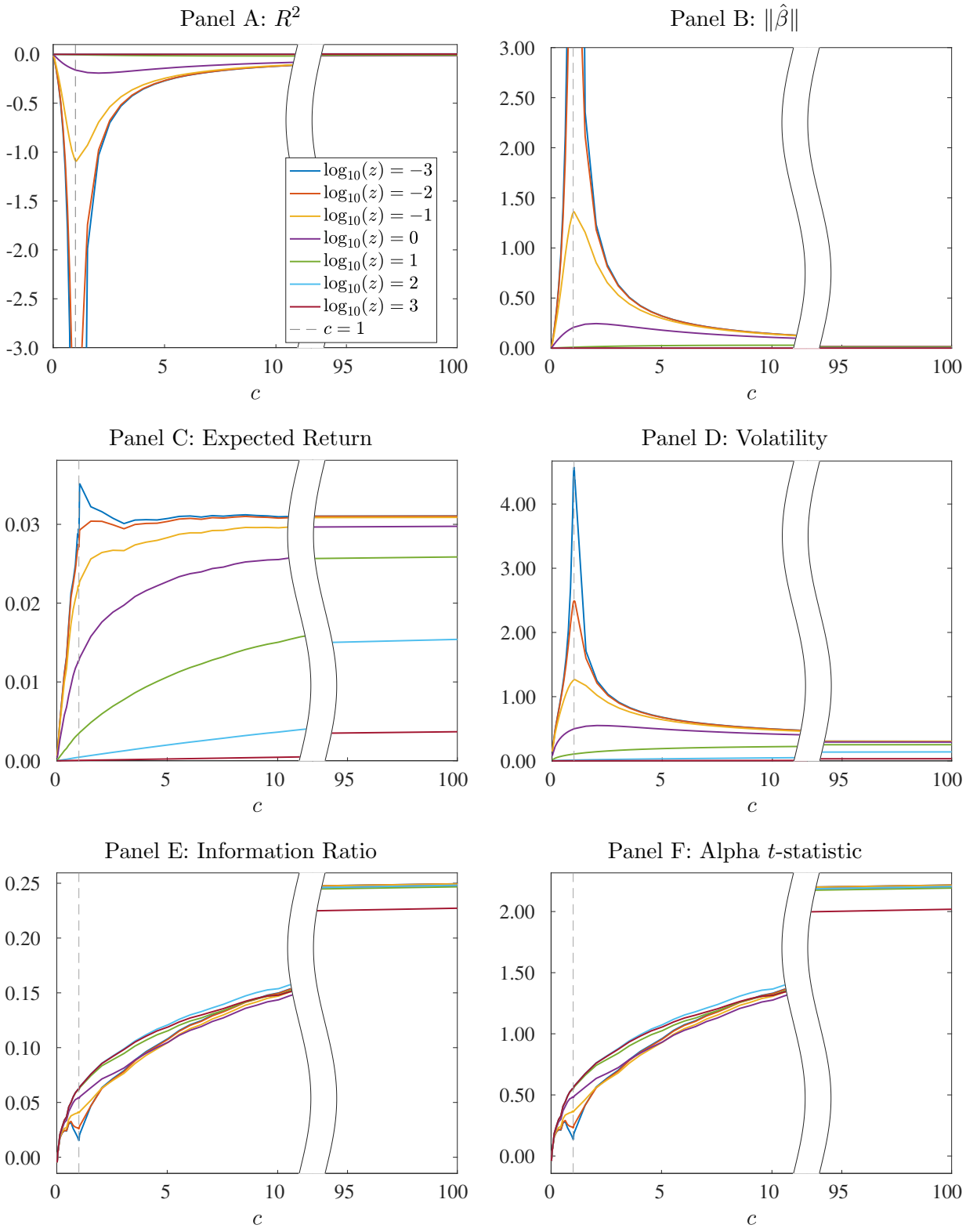

Figure 13: Out-of-sample Market Timing Performance With 120-month Training Window

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 using a range of P. Predictors are RFFs generated from 15 Welch and Goyal (2008) predictors with $\gamma = 2$.

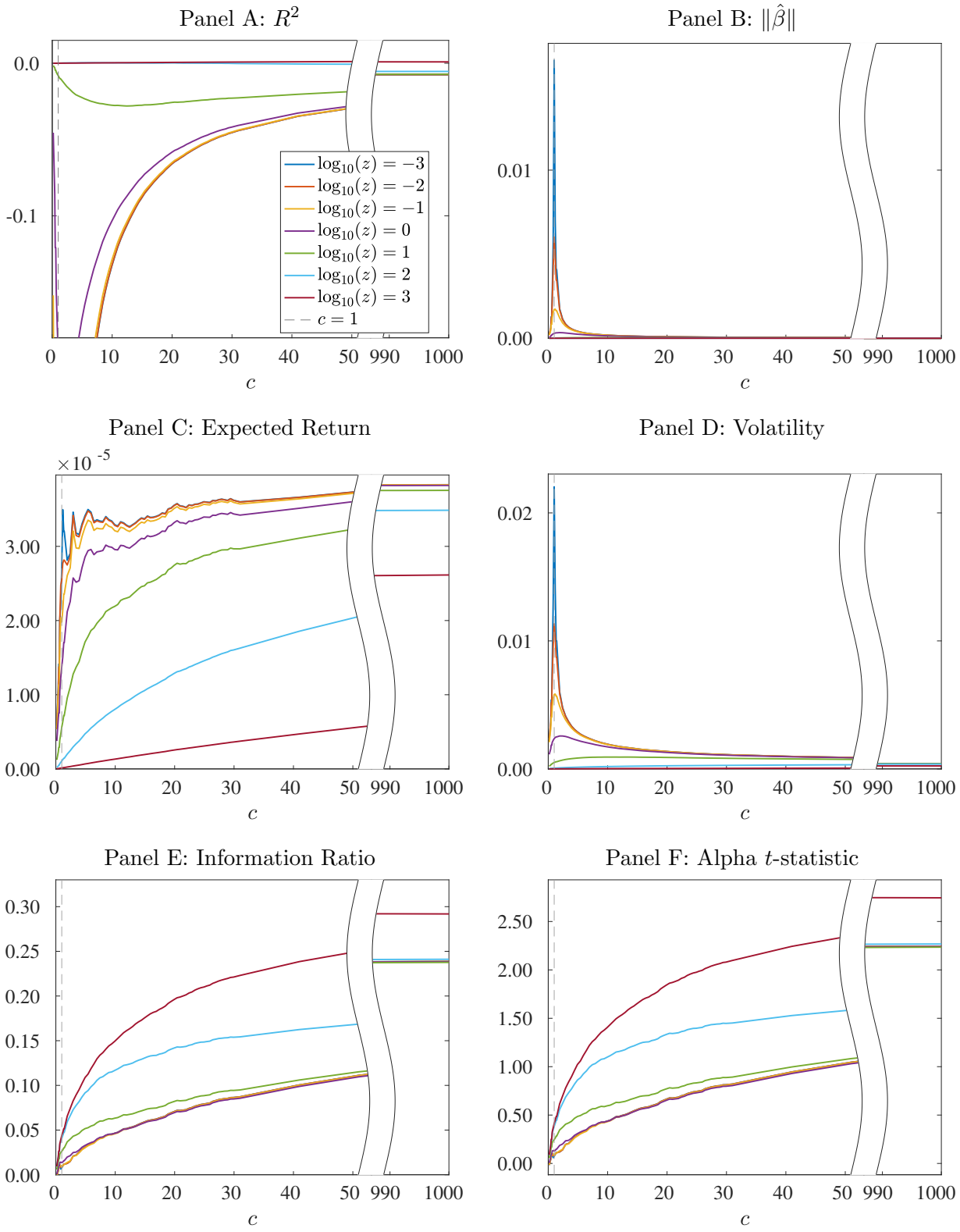

Figure 14: Out-of-sample Market Timing Performance With Un-standardized Returns

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 using a range of P. Predictors are RFFs generated from 15 Welch and Goyal (2008) predictors with $\gamma = 2$. In contrast to our main analysis, returns are not volatility-standardized in this figure.

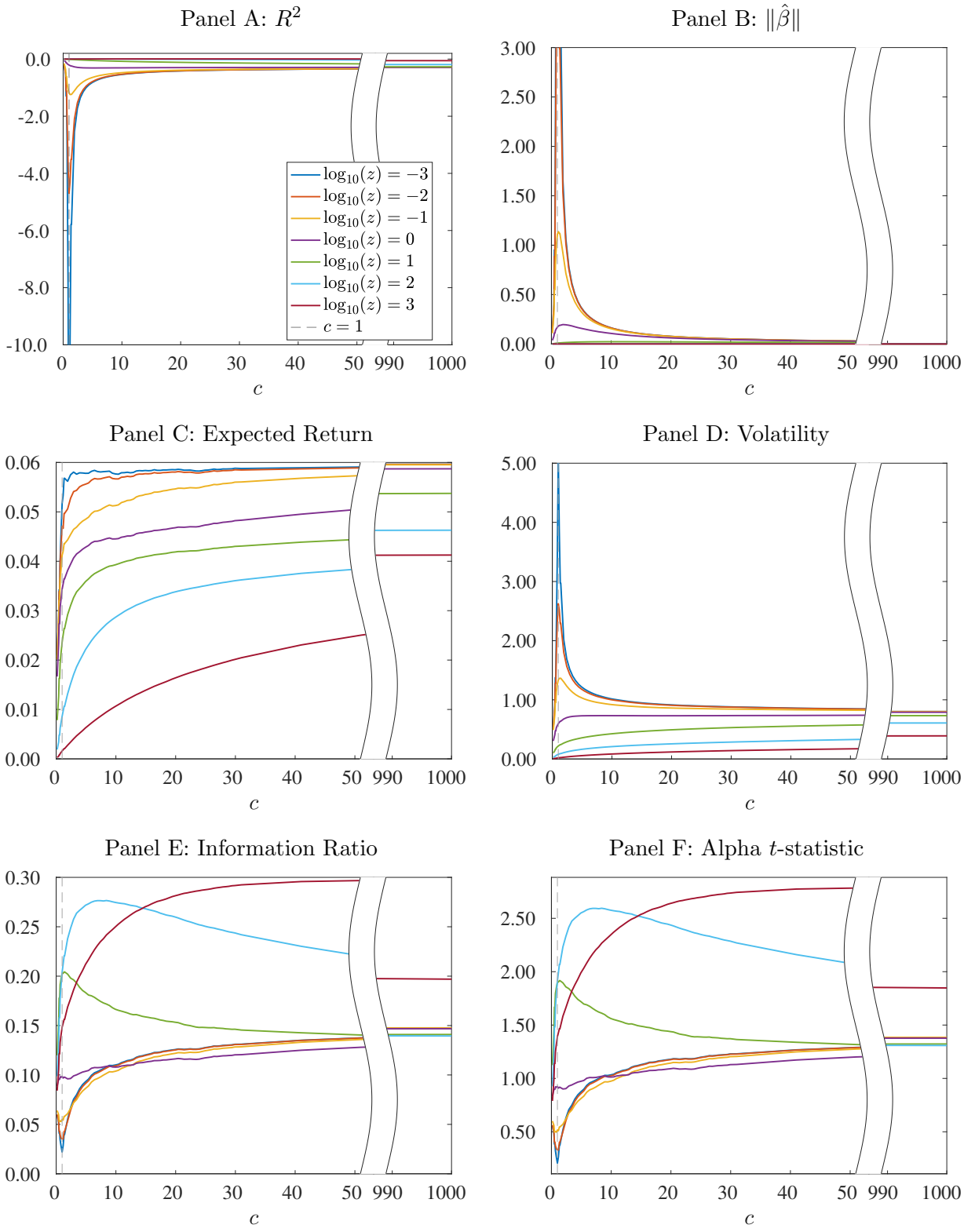

Figure 15: Out-of-sample Market Timing Performance With Bandwidth $\gamma = 1$

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 using a range of P. Predictors are RFFs generated from 15 Welch and Goyal (2008) predictors with $\gamma = 1$.

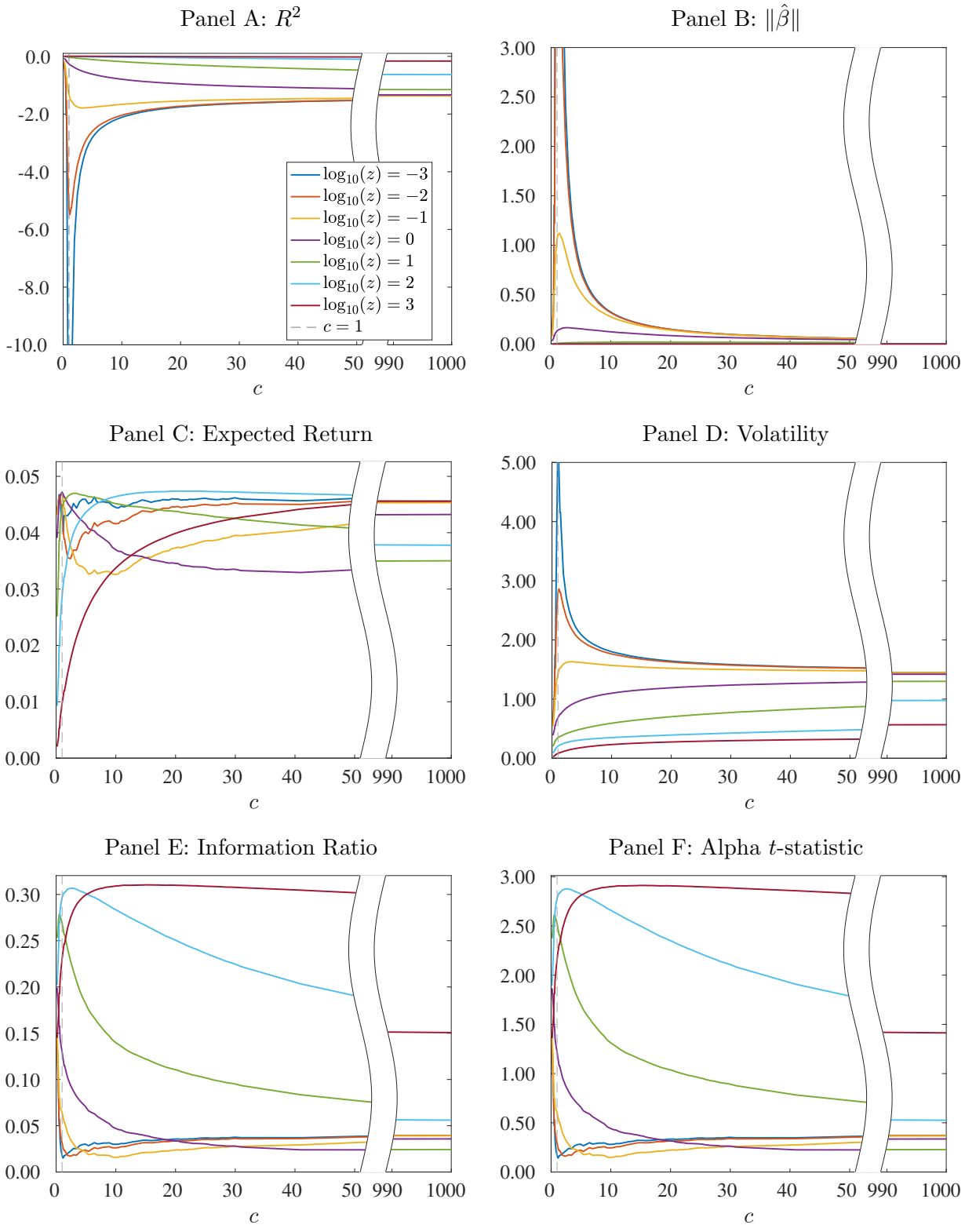

Figure 16: Out-of-sample Market Timing Performance With Bandwidth $\gamma = 0.5$

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 Welch and Goyal (2008) predictors with $\gamma = 0.5$.