The Journal of FINANCE
The Journal of THE AMERICAN FINANCE ASSOCIATION

THE JOURNAL OF FINANCE • VOL. LXXIX, NO. 1 • FEBRUARY 2024

The Virtue of Complexity in Return Prediction

BRYAN KELLY, SEMYON MALAMUD, and KANGYING ZHOU*

ABSTRACT

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

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 document significant empirical gains in portfolio performance through the use of machine learning, there is little theoretical

*Bryan Kelly is at Yale School of Management, AQR Capital Management, and NBER. 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; Daniel Buncic; James Choi; Frank Diebold; Egemen Eren; Paul Goldsmith-Pinkham; Amit Goyal; Ron Kaniel (discussant); Stefan Nagel (Editor); Andreas Neuhierl (discussant); Matthias Pelster (discussant); Olivier Scaillet (discussant); Christian Schlag (discussant); Akos Toeroek; Hui Wang (discussant); Guofu Zhou (discussant); seminar participants at AQR, Yale, Vienna University of Economics and Business, Philadelphia Fed, Bank for International Settlements, NYU Courant, and EPFL; and conference participants at the Macro Finance Society, Adam Smith Asset Pricing Conference, SFS Cavalcade North America Conference, Hong Kong Conference for Fintech, AI, and Big Data in Business, Wharton Jacobs-Levy Conference, Research Symposium on Finance and Economics, China International Risk Forum, Stanford SITE New Frontiers in Asset Pricing, and XXI Symposium. We are especially grateful to Mohammad Pourmohammadi for suggesting several essential improvements to our proofs and technical conditions. AQR Capital Management is a global investment management firm, which may or may not apply similar investment techniques or methods of analysis as described herein. The views expressed here are those of the authors and not necessarily those of AQR. Semyon Malamud gratefully acknowledges support from the Swiss Finance Institute and the Swiss National Science Foundation, Grant 100018_192692. We have read *The Journal of Finance*'s disclosure policy and have no conflicts of interest to disclose.

Correspondence: Bryan Kelly, Yale School of Management, AQR Capital Management, and NBER; 165 Whitney Ave., New Haven, CT 06511; e-mail: bryan.kelly@yale.edu.

DOI: 10.1111/jofi.13298

understanding of return forecasts and portfolios formed from heavily parameterized models.

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

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

where R is an asset return, G is a fixed set of predictive signals, and f is a smooth function. The predictors G may be known to the analyst, but the prediction function f is unknown. Rather than futile guess the functional form, the analyst relies on the universal approximation rationale (see, for example, Hornik, Stinchcombe, and White (1990)), that f can be approximated with a sufficiently wide neural network,

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

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

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

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

Answer: We prove that expected out-of-sample forecast accuracy and portfolio performance are *strictly increasing* in model complexity when appropriate shrinkage is applied (indeed, we derive the optimal degree of shrinkage to maximize expected out-of-sample model performance). The analyst should always use the largest approximating model that she can compute. In other words, when the true data-generating process (DGP) is unknown, the approximation gains achieved through model complexity dominate the statistical costs of heavy parameterization. The interpretation is not necessarily that asset returns are subject to a large number of fundamental driving forces. Rather, even when the driving variables (G_t) have low dimension, complex models

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

better leverage the information content of G_t by more accurately approximating the unknown and likely nonlinear prediction function.

To provide intuitive characterizations of forecast and portfolio behavior in complex models, our theoretical environment has two simplifying aspects. First, the machine learning models we study are restricted to high-dimensional linear models. As suggested by equation (2), this sacrifices little generality as a number of recent papers establish an equivalence between high-dimensional linear models and more sophisticated models such as deep neural networks (Jacot, Gabriel, and Hongler (2018), Allen-Zhu, Li, and Song (2019), Hastie et al. (2022)). In fact, equation (2) is a neural network with one hidden layer with P neurons and fixed input weights. Second, we focus on a single risky asset. Prediction is therefore isolated to the time-series dimension, and the portfolio optimization problem reduces to market timing.[2] These two simplifications make our key findings more accessible, yet neither is critical for our conclusions.

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

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

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

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

[4] This seemingly counterintuitive phenomenon is sometimes called "benign overfit" (Bartlett et al. (2020), Tsigler and Bartlett (2023)).

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

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

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

Third, while the results discussed so far refer primarily to the case of ridgeless regression, we show that machine learning portfolios tend to incrementally benefit from moving away from the ridgeless limit by introducing nontrivial shrinkage. The bias induced by heavier ridge shrinkage lowers the expected returns to market timing, but the associated variance reduction reins

[5] That is, R^2 is not just about predictive correlation. Consider a simple model with a single predictor and a coefficient estimate many times larger than the true value. This scale error will tend to drive the R^2 negative, but it will not 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. Related, Rapach, Strauss, and Zhou (2010) show that mean square forecast error (MSE) decomposes into a scale-free (correlation) component and a scale-dependent component. It is the scale-free component that is important for trading strategy performance. Leitch and Tanner (1991), Cenesizoglu and Timmermann (2012), and Rapach and Zhou (2013) also emphasize the importance of evaluating return prediction models based on their economic value in terms of trading strategy performance.

in the volatility of the strategy. The Sharpe ratio tends to benefit from higher shrinkage because the variance reduction overwhelms the deterioration in expected timing returns. This is especially true when $P \approx T$, where the behavior of ridgeless regression is most vulnerable.

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

We conduct an extensive empirical analysis that demonstrates the virtues of model complexity in a canonical asset pricing problem: predicting the aggregate U.S. equity market return.[6] In particular, we study market timing strategies based on predictions from very simple models with a single parameter to extremely complex models with over 10,000 parameters (applied to training samples with as few as 12 monthly observations). The data inputs to our models are 15 standard predictor variables from the finance literature compiled by Goyal and Welch (2008). To map our data analysis to the theory, we require a method that smoothly transitions from low- to high-complexity models while holding the underlying information set fixed. The random feature method of Rahimi and Recht (2007) is ideal for this. We use it to construct expanding neural network architectures that take the Goyal and Welch (2008) predictors as inputs and maintain the core ridge regression structure of our theory.

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

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

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

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

In this literature, the paper closest to ours is Hastie et al. (2022), who derive nearly optimal error bounds in finite samples for bias and risk in the ridge(less) regression under very general conditions.[9] They are also the first to introduce misspecified models in which some of the signals may be unobservable. In this paper, we focus on the (easier) asymptotic regime. We use a different method of proof and relax some of the technical conditions on the distributions of signals, using recent results of Yaskov (2016). In particular, we allow for nonuniformly 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. (2022), our key interest is in the misspecified model. While Hastie et al. (2022) focus on a specific form of misspecification and its ridgeless limit, we derive general expressions for asymptotic expected returns and volatility in terms of signal correlations.

Our paper also relates closely to a growing empirical literature that uses machine learning methods to analyze stock returns. The state-of-the-art market return prediction uses high-dimensional models with shrinkage and demonstrates robust out-of-sample predictive power. Rapach, Strauss, and Zhou (2010) use predictors from Goyal and Welch (2008) and forecast combination methods (which they show exert a strong shrinkage effect). Ludvigson and Ng (2007) and Kelly and Pruitt (2013) use principal components regression and partial least squares, respectively, to leverage large predictor sets for market return prediction and achieve shrinkage through dimension reduction. Dong et al. (2022) use 100 long-short "anomaly" portfolios to forecast the market return using a variety of forecasting strategies to implement shrinkage (more generally, see the recent survey by Rapach and Zhou (2022)). An emerging literature uses machine learning methods to forecast large panels of individual stock returns or portfolios, including Rapach and Zhou (2020), Kozak, Nagel, and Santosh (2020), Freyberger, Neuhierl, and Weber (2020), Gu, Kelly, and Xiu (2020), and Chen, Pelger, and Zhu (2023) (also see the survey by Kelly

[7] See, for example, Jacot, Gabriel, and Hongler (2018), Hastie et al. (2022), Du et al. (2018), Du et al. (2019), and Allen-Zhu, Li, and Song (2019).

[8] See, for example, Spigler et al. (2019), Belkin et al. (2019), Belkin, Rakhlin, and Tsybakov (2019), Belkin, Hsu, and Xu (2020), and Bartlett et al. (2020).

[9] See also Richards, Mourtada, and Rosasco (2021), who obtain less general results in an asymptotic setting (as in our paper).

and Xiu (2022)). Our paper offers theoretical justification for the successes of machine learning prediction documented in the asset pricing literature. Our theoretical results call for researchers to consider even larger information sets and higher dimensional approximations to further improve return forecasts (a rationale justified by our empirical analysis). Finally, our paper is related to Martin and Nagel (2022) and Da, Nagel, and Xiu (2022), who examine market efficiency implications of the high-dimensional prediction problem faced by investors, to Fan et al. (2022) who touch upon the "double descent" phenomenon in their analysis of structural machine learning models, and to financial econometrics applications of random matrix theory such as Fan, Fan, and Lv (2008), Ledoit and Wolf (2020), and Fan, Guo, and Zheng (2022).

The paper is organized as follows. In Section I, we lay out the theoretical environment. Section II presents the foundational results from random matrix theory from which we derive our main theoretical results. Section III 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 IV extends these results to the more practically relevant setting of misspecified models. We present our main empirical results in Section V. Section VI concludes. The Internet Appendix contains a variety of supplementary theoretical results and empirical robustness analyses.[10] We invite readers that are primarily interested in the qualitative theoretical points and the empirical analysis to skip the technical material of Sections I and II.

I. Environment

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

A. Asset Dynamics

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

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

with ε_{t+1} *independent and identically distributed (i.i.d.),* $E[\varepsilon_{t+1}] = E[\varepsilon_{t+1}^3] = 0, E[\varepsilon_{t+1}^2] = \sigma^2,\ E[\varepsilon_{t+1}^4] < \infty$,[11] *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 infor-

[10] The Internet Appendix is available in the online version of the article on *The Journal of Finance* website.

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

mation set embodied by the predictors, S. The interpretation of this assumption is not necessarily that asset returns are subject to many fundamental driving forces. Instead, it espouses the machine learning perspective discussed in the introduction: The DGP's functional form is unknown but may be approximated with richly parameterized models using a high-dimensional nonlinear expansion S of some underlying feature set.

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

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

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

Furthermore, $E[X_{i,t}] = E[X_{i,t}^3] = E[X_{i_1} X_{i_2} X_{i_3}] = 0$ and $E[X_{i,t}^2] = 1$, for all $i, i_1, i_2, i_3 = 1, \ldots, P$ and $E[X_{i_1} X_{i_2} X_{i_3} X_{i_4}] = 0$ whenever at least three indices among i_1, i_2, i_3, i_4 are different. Furthermore, the fourth moments $E[X_{i,t}^4]$ are uniformly bounded.

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, that is, 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 use $\lambda_k(\Psi)$, $k = 1, \ldots, P$, 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*

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

converges to a nonrandom probability distribution H supported on $[0, +\infty)$.[12] *Furthermore, Ψ is uniformly bounded as $P \to \infty$. We use*

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

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 that $\beta = \beta_P$ is random, $\beta = (\beta_i)_{i=1}^P \in \mathbb{R}^P$, with i.i.d. coordinates β_i that are independent*[13] *of S and R, and such that $E[\beta] = 0$,*

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

[13] The assumption of a random coefficient vector β is related to that in Gagliardini, Ossola, and Scaillet (2016).

$E[\beta\beta'] = P^{-1}b_{*,P}I$ for some constant $b_{*,P} = E[\|\beta\|^2]$,[14] *and satisfies $b_{*,P} \to b_*$ in probability for some $b_* > 0$. Furthermore, $E[\beta_i^4] \leq KP^{-2}$ for some $K > 0$.*

The randomness of β in Assumption 4 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 a nonzero mean and restate our analysis in terms of variances rather than second moments. The assumption $E[\beta\beta'] = P^{-1}b_{*,P}I$ imposes that the predictive content of signals is rotationally symmetric, that is, predictability is uniformly distributed across signals. This may seem restrictive, as commonly used return predictors would not satisfy Assumption 4. But it is closely aligned with the structure of feed-forward neural networks, in which raw features are mixed and nonlinearly propagated into final generated features whose ordering is essentially randomized by the initialization step of network training. Intuitively, we expect (and later confirm empirically) that the random-feature methodology that we use in our empirical analysis satisfies Assumption 4.[15]

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. It ensures that Sharpe ratios of timing strategies remain bounded as the number of predictors grows. In other words, our setting is one with many signals, each contributing a little bit of predictability.

A key aspect of our paper, and one rooted in Assumptions 2 and 4, is that realized out-of-sample expected 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.[16]

LEMMA 1: *As $P \to \infty$, we have*

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

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

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

[15] From a technical standpoint, it is possible to derive explicit expressions for portfolio performance without this assumption, but the expressions become more complex. In this case, the asymptotic behavior depends on the distribution of projections of β on the eigenvectors of Ψ (the signals' principal components). See Hastie et al. (2022). In particular, when β is concentrated on the top principal components, the phenomenon of benign overfit emerges (Bartlett et al. (2020), Tsigler and Bartlett (2023)) and the optimal ridge regularization is zero. We leave this generalization for future research.

[16] It is possible to use the results in Hastie et al. (2022) to extend our analysis to generic β distributions. We leave this important direction for future research.

B. Timing Strategies and Performance Evaluation

We study timing-strategy returns, defined as

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

where π_t is a timing weight that scales the position in the asset up and down to exploit time variation in the asset's expected returns.

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

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

While there are other possible performance criteria, we focus on this one 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.[17]

Our analysis centers on the following timing-strategy functional form:

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

This strategy takes positions equal to the asset's conditional expected return. Note that this timing strategy optimizes the *conditional* Sharpe ratio. It achieves the same Sharpe ratio as the conditional Markowitz solution, $\pi_t^{\text{Cond. MV}} = E_t[R_{t+1}]/\text{Var}_t[R_{t+1}] = S_t'\beta$, according to equation (3). While strategy π_t is conditionally mean-variance efficient, it is not the optimizer of the unconditional objective in (5), which takes the form $\pi_t^{\text{Uncond. MV}} = S_t'\beta/(1 + (S_t'\beta)^2)$.[18] In the proof of Proposition 1 in the Internet Appendix, we show that π_t in equation (6) and $\pi_t^{\text{Uncond. MV}}$ are equal up to third-order terms.[19] We study $\pi_t = S_t'\beta$ for the simplicity of its linearity in both β and S_t, but note that our conclusions are identical for $\pi_t^{\text{Uncond. MV}}$ because, in the limit as $P \to \infty$, the normalization factor $1 + (S_t'\beta)^2$ converges to a constant.[20]

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

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

[18] See Hansen and Richard (1987), Ferson and Siegel (2001), Abhyankar, Basu, and Stremme (2012).

[19] In particular, the Sharpe ratio in equation (5) 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.

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

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

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

The infeasible market-timing Sharpe ratio is

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

For comparison, under Assumptions 1 to 4, the unconditional first and second moments of the untimed 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 that the untimed asset has a Sharpe ratio of zero. This is just a normalization so that any positive market timing Sharpe ratio can be interpreted as pure excess performance arising from timing ability.

C. 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 MSE on an independent test sample. This is the fundamental notion of estimator "risk" from statistical theory, although we use the term "MSE" here to avoid confusion with portfolio riskiness. We can write MSE as

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

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

Note that a strategy $\pi_t = \beta'S_t$ based on the infeasible true β satisfies $E[\pi_t R_{t+1}] = E[\beta'\Psi\beta] = E[\pi_t^2]$.[21] 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

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

of the infeasible R^2 based on equation (3) and Lemma 1:

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

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

II. 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 increasingly complex, that is, as $P \to \infty$. Because this involves estimating infinite-dimensional parameters, traditional large-T asymptotics do not apply, and hence 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.

A. 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, $P/T \to c > 1$, although we will see that it also has important implications for the behavior of $\hat{\beta}(z)$ when $P/T < 1$.[22]

Consider first the OLS estimator, $\hat{\beta}(0)$. As P approaches T from below, the denominator of the least-squares estimator approaches the singularity. This produces explosive variance of $\hat{\beta}(0)$ and, in turn, explosive forecast error variance. As $P \to T$, the model begins to fit the data with zero error, so a common

[22] One could alternatively analyze "sparse" least-squares models that combine shrinkage with variable selection (e.g., based on LASSO). First, recent evidence of Giannone, Lenza, and Primiceri (2021) suggests that sparsity of predictive relationships in economics and finance is likely an illusion. Second, our empirical focus is on nonparametric models that seek to approximate a generic nonlinear function as a linear combination of generated features, and sparsity in the generated feature space is difficult to identify (see, for example, Ghorbani et al. (2020)). Third, analysis with ℓ_1 shrinkage is significantly more taxing from a theoretical standpoint. We thus leave sparse least-squares models to future research.

interpretation of the explosive variance of $\hat{\beta}(0)$ is an insidious overfit that does not generalize out-of-sample.

When P moves beyond T, there are more parameters than observations and the least-squares problem has multiple solutions. A particularly interesting solution invokes the Moore-Penrose pseudo-inverse, $(T^{-1}\sum_t S_t S_t')^+ \frac{1}{T}\sum_t S_t R_{t+1}$.[23] 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, but 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. (2022)). The machine learning literature has recently devoted substantial attention to understanding ridgeless regression in the high-complexity regime. The counterintuitive 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.

B. The Role of Random Matrix Theory

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

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

[23] Recall that the Moore-Penrose pseudo-inverse A^+ of a matrix A is defined as $A^+ = \lim_{z\to 0+}(zI + A'A)^{-1}A' = \lim_{z\to 0+}A'(zI + AA')^{-1}$.

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

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

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

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

From this identity, we immediately see the fundamental connection between ridge regularization and the distribution of eigenvalues for Ψ. The right-hand 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 nonrandom distribution H. We therefore have

$$m_{\Psi}(z) := \int \frac{1}{x - z} dH(x) = \lim_{P\to\infty} \frac{1}{P}\text{tr}\big((\Psi - zI)^{-1}\big). \tag{9}$$

The function $m_{\Psi}(z)$ is the *limiting* Stieltjes transform of the eigenvalue distribution of Ψ. Equation (9) is a powerful step toward 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 characterized entirely if we know Ψ's eigenvalue distribution.

The problem, of course, is that the true Ψ is unobservable. We only observe its sample counterpart, $\hat{\Psi}$, and 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

$$m(z; c) := \lim_{P\to\infty} \frac{1}{P}\text{tr}\big((\hat{\Psi} - zI)^{-1}\big).$$

In traditional finite P statistics, we would have a convergence between the sample covariance $\hat{\Psi}$ and the true covariance Ψ as $T \to \infty$. One might be tempted to think that $\lim_{P\to\infty} \frac{1}{P}\text{tr}((\hat{\Psi} - zI)^{-1})$ and $\lim_{P\to\infty} \frac{1}{P}\text{tr}((\Psi - zI)^{-1})$ 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 random matrix theory can help us understand it. In the Internet Appendix (see Theorem 2), 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 $c > 0$.[24] The next result shows that,

[24] Theorem 2 in the Internet 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 $m_{\Psi}(z) =$

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.[25]

PROPOSITION 2: *We have*

$$\lim_{T\to\infty} \frac{1}{T}\text{tr}((zI+\hat{\Psi})^{-1}\Psi) \;\to\; \xi(z;c) \tag{10}$$

almost surely, where

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

The quantity $\text{tr}\, E[(zI+\hat{\Psi})^{-1}\Psi]$ appears in virtually every expression we analyze to describe portfolio behavior. It depends on the interaction between the sample and true signal covariance matrix and arises in the computation of both the expected return and leverage of the timing strategy (see equation (8)). One might imagine, therefore, that we need to know the limiting eigenvalue distribution of both matrices (or their Stieltjes transforms, m and m_Ψ) to describe $\text{tr}\, E[(zI+\hat{\Psi})^{-1}\Psi]$. 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 that we know the eigenvalues of Ψ.[26]

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

$(1-z)^{-1}$, Marčenko and Pastur (1967) show that

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

By direct calculation, the expression above is indeed the unique positive solution to (IA4) when $m_\Psi(z) = (1-z)^{-1}$. While the eigenvalue distributions of the sample and true covariance matrices do not coincide, Theorem 2 describes the precise nonlinear way they relate to each other. In particular, when $P > T$, the matrix $\hat{\Psi}$ has $P - T$ zero eigenvalues and therefore, $P^{-1}\text{tr}((zI+\hat{\Psi})^{-1})$ contains a singular part, $P^{-1}(P-T)z^{-1} = (1-c^{-1})z^{-1}$.

[25] It is possible to develop *nonlinear* shrinkage estimators analogous to those developed by Ledoit and Wolf (2020) for covariance matrices. Such estimators would require knowledge of the true eigenvalue distribution of Ψ, which can be recovered from $m(z;c)$ using equation (IA4).

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

can continue to generate substantial trading gains. This is true even in the ridgeless case. Additional ridge shrinkage can boost performance even further. We precisely characterize these behaviors in the following sections.

III. 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.

A. 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

$$MSE(z;c) = \lim_{T,P\to\infty,\ P/T\to c} E\Big[\big(R_{t+1} - S_t'\hat{\beta}(z)\big)^2 | \hat{\beta}(z)\Big]. \tag{11}$$

Notably, while $\hat{\beta}(z)$ is random and depends on the sample realization, we show below that the limit in (11) is nonrandom. 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 I.C above, we discuss 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 ($z = 0$). The $R^2(0;0)$ is the benchmark for evaluating the loss of predictive accuracy due to high model complexity, even when data are abundant. Specifically, the R^2 of the least-squares estimator in the machine learning regime behaves as follows.

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

$$\begin{aligned} \mathcal{E}(z;c) &= \lim E[\hat{\pi}_t R_{t+1} | \hat{\beta}(z)] = b_* \nu(z;c), \\ \mathcal{L}(z;c) &= \lim E[\hat{\pi}_t^2 | \hat{\beta}(z)] = b_* \hat{\nu}(z;c) - c\nu'(z;c), \\ R^2(z;c) &= \frac{2\mathcal{E}(z;c) - \mathcal{L}(z;c)}{1 + b_*\psi_{*,1}}, \end{aligned} \tag{12}$$

where

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

As we show in the Internet 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 $z > z_*$. The $R^2(z;c)$ attains its maximum at $z_* = c/b_*$, where it is positive and given by

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

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

$$R^2(0;c) = R^2(0;0) - (1 + b_*\psi_{*,1})^{-1}\begin{cases}(c^{-1}-1)^{-1}, & c < 1\\ \mu(c), & c > 1\end{cases} \tag{13}$$

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

$$\lim_{c\to\infty} R^2(0;c) = 0 > \lim_{c\to 1} R^2(0;c) = -\infty\,. \tag{14}$$

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. The degradation in predictive accuracy due to complexity can be so severe that expected out-of-sample R^2 becomes extremely negative, particularly in the ridgeless case. Shrinkage can mitigate this and help preserve accuracy amid complexity. Shrinkage controls variance but introduces bias. Proposition 3 points out that the amount of shrinkage that optimizes the bias-variance trade-off is $z_* = c/b_*$.[27] More complex settings benefit from heavier shrinkage, while settings with a higher signal-to-noise ratio (higher b_*) benefit from lighter shrinkage (see, for example, Hastie et al. (2022)). The $\mathcal{E}$ and $\mathcal{L}$ are the limiting out-of-sample expected return and leverage of the timing strategy. Proposition 3 shows that these are the main determinants of out-of-sample R^2.

Figure 1 illustrates the theoretical behavior of the least-squares estimator derived in Proposition 3. The plots set Ψ to the identity matrix and fix $b_* = 0.2$

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

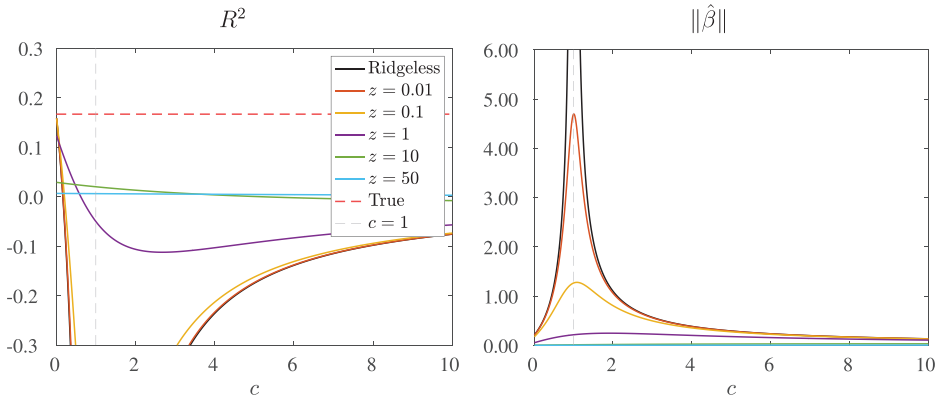

Figure 1. Expected out-of-sample R^2 and norm of least-squares coefficient. This figure shows the limiting out-of-sample R^2 and $\hat{\beta}$ norm as a function of c and z from Proposition 3 assuming Ψ is the identity matrix and $b_* = 0.2$. (Color figure can be viewed at wileyonlinelibrary.com)

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

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

To the right of $c = 1$, the number of predictors exceeds the sample size, and the "ridgeless" case is defined as the limit as $z \to 0$ (i.e., when the least-squares denominator is calculated via the pseudo-inverse of $\hat{\Psi}$). Counterintuitively, the R^2 begins to *rise* as model complexity increases.[28]

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

[28] This is an illustration of what the statistics literature refers to as benign overfitting.
[29] That is, $\beta' S_t = R_{t+1}$ for all $t \in [1, \ldots, T]$.

ridgeless least squares is a newly documented phenomenon in the statistics literature and an emerging topic of research.[30] It shows that even in very simple DGPs, 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 nontrivial ridge shrinkage. Allowing $z > 0$ improves R^2 except at very low levels of complexity. This is again a manifestation of the bias-variance trade-off. When $z > 0$, the norm of $\hat{\beta}$ is controlled, and the associated variance reduction outweighs the effects of bias when the model is complex.

It is useful to place our analysis thus far in the context of the literature. Some formulas of Propositions 2 and 3 have been established in papers on random matrix theory (e.g., Ledoit and Péché (2011)). Hastie et al. (2022) 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 nonasymptotic bounds on the rate of convergence. However, both Hastie et al. (2022) and Wu and Xu (2020) require that Ψ is strictly positive definite. By contrast, in our data analysis, we find that Ψ is nearly degenerate. Richards, Mourtada, and Rosasco (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 RFFs used in our empirical analysis (or any other nonlinear signal transformations). In contrast to these papers, we establish our results under much weaker conditions on the distribution of $X_{i,t}$ across i. This is important for practical applications, where neither the independence of X_t nor equality (or boundedness) of their higher moments can be guaranteed. Lastly, the novel techniques that we develop allow us to characterize the out-of-sample performance of misspecified models. While Hastie et al. (2022) study misspecification in certain stylized examples, we derive more general results allowing for generic covariance structures. To the best of our knowledge, our characterization is new in the literature (see Section IV).

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

B. Expected Out-of-Sample Market Timing Performance

We analyze the behavior of market timing based on the least-squares estimate,

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

Formula (12) derives the expected return of this strategy. The following proposition characterizes the expected out-of-sample risk-return trade-off of market timing in the high-complexity regime.

[30] See Spigler et al. (2019), Belkin et al. (2019), Belkin, Rakhlin, and Tsybakov (2019), Belkin, Hsu, and Xu (2020), and Hastie et al. (2022).

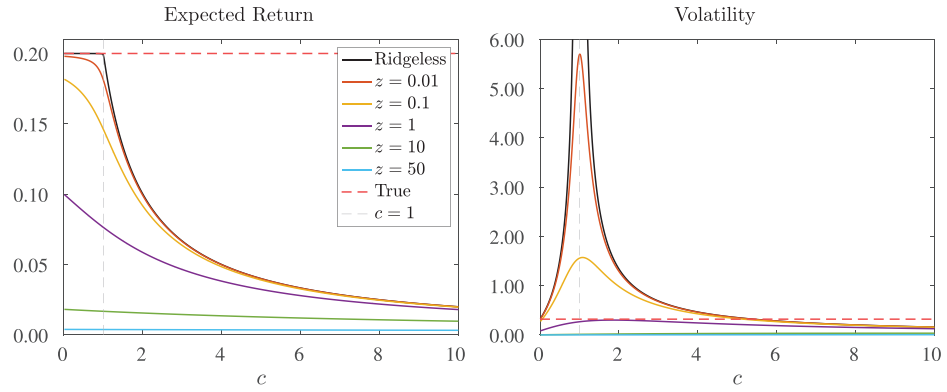

Figure 2. Expected out-of-sample risk and return of market timing. This figure shows the limiting out-of-sample expected return and volatility of the market timing strategy as a function of c and z from Proposition 3 assuming Ψ is the identity matrix and $b_* = 0.2$. (Color figure can be viewed at wileyonlinelibrary.com)

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

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

in probability, with $\mathcal{E}$ and $\mathcal{L}$ given in (12). As a result, the Sharpe ratio satisfies

$$SR(z;c) = \frac{\mathcal{E}(z;c)}{\sqrt{\mathcal{V}(z;c)}} = \frac{1}{\sqrt{2 + \left(1 + b_*\psi_{*,1}\right)\frac{\mathcal{L}(z;c)}{(\mathcal{E}(z;c))^2}}}. \tag{15}$$

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

$$SR(z_*;c) = \frac{1}{\sqrt{2 + \left(1 + b_*\psi_{*,1}\right)\frac{1}{b_*\nu(z_*;c)}}} < SR(0;0), \tag{16}$$

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 black. The expected returns of least-squares timing strategies are always positive because they are quadratic in beta. When $c < 1$ (i.e., in the OLS case), the ridgeless timing strategy achieves the true expected return even though the

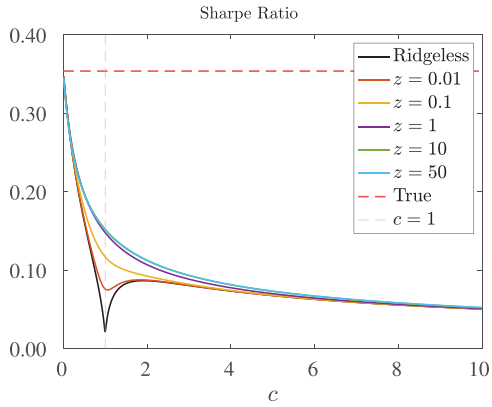

Figure 3. Expected out-of-sample Sharpe ratio of market timing. This figure shows the limiting out-of-sample Sharpe ratio of the market timing strategy as a function of c and z from Proposition 3 assuming Ψ is the identity matrix and $b_* = 0.2$. (Color figure can be viewed at wileyonlinelibrary.com)

corresponding R^2 is significantly negative in much of this range. The fact that the out-of-sample expected return is unimpaired reflects the unbiasedness of OLS, while the declining R^2 reflects the increasing forecast variance as c rises toward one. The return volatility of the timing strategy is likewise increasing in c for $c \in [0, 1]$ due to the rising forecast variance and maxes out at $c = 1$.

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

The other expected return and volatility curves show that the bias induced by a nontrivial 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 relatively high shrinkage levels like $z = 1$, the volatility of the timing strategy drops even below that of the infeasible best strategy while maintaining a meaningfully positive expected return.

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

Another interesting aspect of Figure 3 is that the Sharpe ratio benefits from nontrivial ridge shrinkage regardless of model complexity. Shrinkage is most

valuable near $c = 1$, where it reins in volatility substantially more than it reduces expected return. At both low levels of complexity ($c \approx 0$) and high levels of complexity ($c >> 1$), the Sharpe ratio is relatively insensitive to z.

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

C. A Note on R^2

At this point, we already see that a timing strategy with negative R^2 can have high average out-of-sample returns and thus positive out-of-sample Sharpe ratios.[31] More plainly, the positivity of out-of-sample R^2 is *not* a necessary condition for an economically valuable timing strategy. The least-squares timing strategies in our framework all have strictly positive out-of-sample expected return and Sharpe ratio regardless of shrinkage or model complexity (despite having enormously negative R^2 in many cases).

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

Out-of-sample R^2 and Sharpe ratio measurements serve different purposes. The R^2 helps evaluate forecast accuracy, while the Sharpe ratio is appropriate for evaluating the economic value of forecasts in asset allocation contexts. Much of the empirical literature on return prediction and market timing focuses its evaluations on out-of-sample predictive R^2 (see, for example, Goyal and Welch (2008)). Proposition 4 ensures that we can worry less about the

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

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.

IV. Machine Learning and Model Misspecification

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, and thus, these theoretical comparative statics cannot be taken to the data. Nevertheless, theory grounded on correct model specification is powerful for developing a conceptual understanding of machine learning portfolios.

A more empirically relevant theoretical setting would consider a single true DGP. It would then 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 $P_1 = 1, \ldots, P$ 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, misspecified estimator behavior depends on the correlation structure of observed and unobserved signals 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 nonrandom probability distribution $H(x; q)$. We say that signals are sufficiently mixed if $H(x; q)$ is independent of q. We also use*

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

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

In a misspecified 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:

$$\xi_{2,1}(z;cq;q) = \lim_{T\to\infty} T^{-1}\mathrm{tr}\, E[(zI+\hat{\Psi}_{1,1})^{-1}\Psi_{1,2}\Psi_{1,2}'] \geq 0, \tag{17}$$

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

The quantities in (17) account for covariances between observed and unobserved signals. While the existence of the limits in (17) cannot be guaranteed in general, the expectations are uniformly bounded for $z > 0$ (as the Ψ matrices are uniformly bounded for $z > 0$). Hence, by passing to a subsequence of T, P, we can always assume that the limits in (17) exist. In the Internet Appendix, we show that these limits actually exist for a class of correlation structures.

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

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

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

in probability, where

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

and

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

Furthermore,

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

In addition, we have

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

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

(ii) Expected leverage converges in probability to

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

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

$$R^2(z;cq;q) = \frac{2\mathcal{E}(z;cq;q)-\mathcal{L}(z;cq;q)}{1+b_*\psi_{*,1}(1)}. \tag{18}$$

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

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

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

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

In general, the behavior of the quantities in Proposition 5 depends in complex fashion on the correlations between observable and unobservable signals, as captured by the quantities (17). When both quantities (17) are zero, the expressions simplify significantly. It is straightforward to show that both quantities in (17) are zero if the matrices $\Psi_{1,2}$, $\Psi_{2,1}$ have uniformly bounded traces. For example, this is when $\Psi_{1,2}$ has a finite, uniformly bounded rank when $P,\ P_1 \to \infty$ (due to, say, a finite-dimensional factor structure in the signals). We thus obtain the following result.

PROPOSITION 6: *Suppose that* $\mathrm{tr}(\Psi_{1,2}\Psi_{2,1}) = o(P)$.[32] *Then,* $\xi_{2,1} = \widehat{\xi}_{2,1} = 0$. *Furthermore,*

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

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

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

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

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

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

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

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

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

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

is constant for $q > 1/c$.

The comparative statics of Section III.B 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 that when models are correctly specified, the best performance (in terms of R^2 and Sharpe ratio) comes from simple models. Naturally, a small correctly specified model will converge on the truth faster than a large correctly specified model. But this is not a very helpful comparison.

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

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

Figures 4, 5, and 6 illustrate the behavior of misspecified machine learning predictions and portfolios derived in Proposition 5. In this calibration, the true unknown DGP is assumed to have a complexity of $c = 10$. We continue to calibrate 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

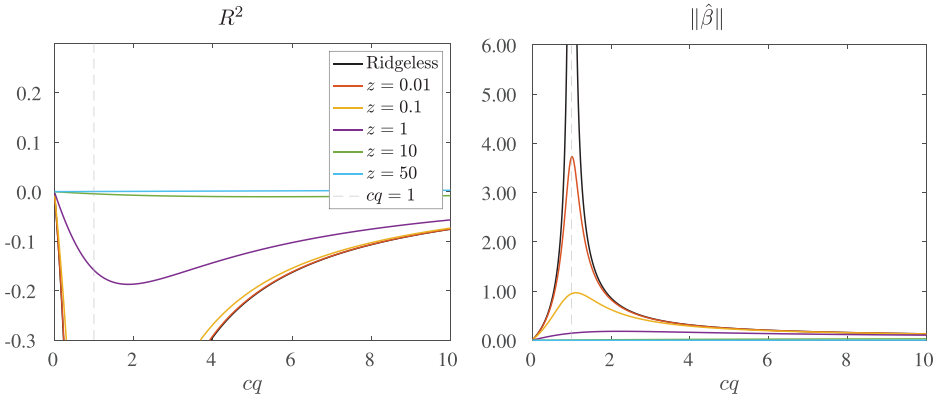

Figure 4. Expected out-of-sample prediction accuracy from misspecified models. This figure shows the limiting out-of-sample R^2 and $\hat{\beta}$ norm as a function of c and z from Proposition 6 assuming Ψ is the identity matrix, $b_* = 0.2$, and the complexity of the true model is $c = 10$. (Color figure can be viewed at wileyonlinelibrary.com)

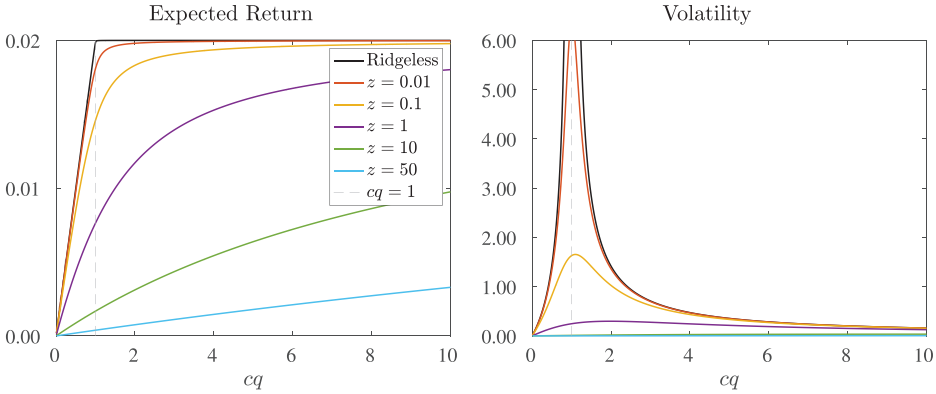

Figure 5. Expected out-of-sample risk and return from misspecified models. This figure shows the limiting out-of-sample expected return and volatility of the market timing strategy as a function of c and z from Proposition 6 assuming Ψ is the identity matrix, $b_* = 0.2$, and the complexity of the true model is $c = 10$. (Color figure can be viewed at wileyonlinelibrary.com)

thus severely misspecified) to highly complex ($q = 1$, $cq = 10$ and thus correctly specified). The left panel of Figure 4 shows the expected out-of-sample R^2. The cost of misspecification 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 ($cq \approx 1$) 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 $cq > 1$ due to the implicit regularization of high-complexity ridgeless regression, just as in the earlier analysis. More generally,

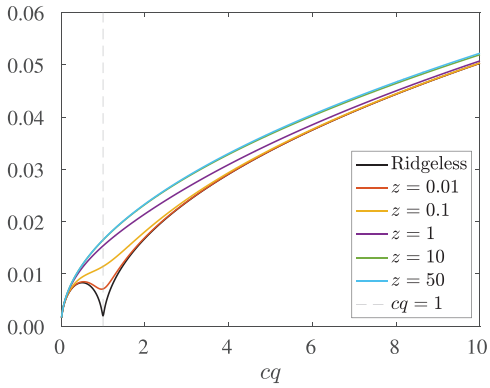

Figure 6. Expected out-of-sample Sharpe ratio from misspecified models. This figure shows the limiting out-of-sample Sharpe ratio of the market timing strategy as a function of c and z from Proposition 6 assuming Ψ is the identity matrix, $b_* = 0.2$, and the complexity of the true model is $c = 10$. (Color figure can be viewed at wileyonlinelibrary.com)

the patterns for R^2, $\hat{\beta}$ norm, and portfolio volatility share similar qualitative patterns as those in Figure 1.

The most important difference compared to Figure 1 is the pattern for the out-of-sample expected return of the market timing strategy (Figure 5, left panel). Expected returns are now low for simple strategies due to their poor approximation of the DGP. Increasing model complexity monotonically increases expected timing returns. In the ridgeless case, the benefit of added complexity reaches its maximum of $\mathcal{E}(0; 1; c^{-1}) = b_* \psi_{*,1} c^{-1}$ when $cq = 1$. A surprising fact is that the ridgeless expected return is exactly flat as complexity rises beyond $cq = 1$, in which case the benefits of incremental improvements in DGP approximation are exactly offset by the gradually rising bias of ridgeless shrinkage; see formula (19).

This new fact that the expected return rises monotonically with model complexity in the misspecified setting induces a similar pattern in the out-of-sample Sharpe ratio, shown in Figure 6. Rather than decreasing in complexity as we saw in the correctly specified setting, the expected return improvement from additional complexity leads the Sharpe ratio to also increase with complexity. Consistent with Theorem 1, this is particularly true with nontrivial 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 misspecified empirical models, complexity is a virtue. It improves the expected out-of-sample market timing performance in terms of both expected return and Sharpe ratio.

It is instructive to compare our findings with the phenomenon of double descent, where by absent regularization, out-of-sample MSE has a nonmonotonic pattern in model complexity (Belkin et al. (2019), Hastie et al. (2022)). The mirror image of double descent in *MSE* is the "double ascent" behavior of the ridgeless Sharpe ratio (Figure 6). As Theorem 1 shows, Sharpe ratio double ascent is an artifact of insufficient shrinkage. With the right amount of shrinkage,

complexity becomes a virtue even in the low-complexity regime (when $cq < 1$): The hump disappears, and "double ascent" turns into "permanent ascent."

V. Virtue of Complexity: Empirical Evidence From Market Timing

In this section, we present empirical analyses that are direct empirical analogs to the theoretical comparative statics for misspecified models in Section IV.

A. 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 excess return of the CRSP value-weighted index. The information set we use for prediction consists of the 15 predictor variables from Goyal and Welch (2008) available monthly over the sample from 1926 to 2020.[33]

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).[34] In contrast, predictors are standardized using an expanding window historical standard deviation (given the much higher persistence of most predictors). We require 36 months of data to ensure enough stability in our initial predictor standardization, so the final sample we bring to our analysis began in 1930. We perform this standardization to align the empirical analysis with our homoskedastic theoretical setting. Our results are insensitive to this step—none of our findings are sensitive to variations in how standardizations are implemented.

B. Random Fourier Features

We seek models that take the form of equation (3).[35] To evaluate our theory, we also seek a framework that will allow us to smoothly transition from low-

[33] 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. Most of these variables are based on market prices and are available at month end. Our date convention for inflation is that used by Goyal, Welch, and Zafirov (2023) and the data set graciously provided by Amit Goyal. Note that while inflation for month t is typically reported two weeks into month $t + 1$, the Goyal, Welch, and Zafirov (2023) date convention views the price data upon which the official inflation statistic is based as part of the time t information set. We show in Internet Appendix Figure IA12 and Table IA2 that our results are essentially unchanged if we exclude inflation from our analysis.

[34] For returns, we calculate standard deviation from the uncentered second moment due to the noisiness of estimating mean monthly returns in short windows.

[35] As in equation (3), we exclude the intercept from our regressions. If we include a constant as an additional regressor in our high complexity regressions, the associated intercept is shrunken so heavily that it has no effect on the results reported in Table I.

complexity models to high-complexity models. To do so, we adopt an influential methodology from the machine learning literature known as RFF (Rahimi and Recht (2007), Rahimi and Recht (2008)).[36] Let G_t denote our 15×1 vector of predictors. The RFF methodology converts G_t into a pair of new signals,

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

where $S_{i,t}$ uses the vector ω_i to form a random linear combination of G_t, which is then fed through the trigonometric functions.[37] 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 nonlinear transformation in (20). If one desires a very low-dimensional model in (3), say $P = 2$, one can generate a single pair of RFFs. For a very high-dimensional model, say $P = 10{,}000$, one can instead draw many random weight vectors ω_i, $i = 1, \ldots, 5{,}000$. The larger the number of random features, the richer the approximation that (3) provides to the general functional form $E[R_{t+1}|G_t] = f(G_t)$, where f is some smooth nonlinear function. Indeed, the RFF approach is a wide two-layer neural network with fixed weights in the first layer (in the form of ω_i) and optimized weights in the second layer (in the form of the regression estimates for β).

C. Out-of-Sample Performance

To conduct the empirical analogue of the theoretical analysis in Figures 4, 5, and 6, we consider one-year, five-year, and 10-year rolling training windows ($T = 12$, 60, or 120) and a large set of RFFs (as high as $P = 12{,}000$). These choices are guided by our desire to investigate the role of model complexity, defined in the empirical analysis as $c = P/T$. The advantages of short training samples like $T = 12$ are (i) we can reach extreme levels of model complexity with smaller P and thus less computing burden and (ii) it shows that the virtue of complexity can be enjoyed in small samples. But none of our conclusions are sensitive to this choice as we document all of the same patterns for training windows of $T = 60$ and 120.

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

[37] The parameter γ controls the Gaussian kernel bandwidth in the generation of RFFs. Random features can be generated in several ways (for a survey, see Liu et al. (2021)). 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. We find, however, that our results are insensitive to using other random feature schemes.

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

(i) Generate 12,000 RFFs according to (20) with bandwidth parameter γ.[38]
(ii) Fix a model defined by the number of features $P \in \{2, \ldots, 12{,}000\}$ and ridge shrinkage parameter $\log_{10}(z) \in \{-3, \ldots, 3\}$. The set of predictors S_t for regression (3) corresponds to the first P RFFs from (i).
(iii) Given the model in (ii), and fixing a training window $T \in \{12, 60, 120\}$, conduct a recursive out-of-sample prediction and market timing strategy. For each $t \in \{T, \ldots, 1{,}091\}$, estimate (3) using training observations $\{(R_t, S_{t-1}), \ldots, (R_{t-T+1}, S_{t-T})\}$.[39] Then, from the estimated regression coefficient, construct out-of-sample return forecast $\hat{\beta}' S_t$ and timing strategy return $\hat{\beta}' S_t R_{t+1}$.
(iv) From the sequence of out-of-sample predictions and strategy returns in (iii), calculate the average $\|\hat{\beta}\|^2$ across training samples, the out-of-sample R^2, and the out-of-sample average return, volatility, and Sharpe ratio of the timing strategy.[40]

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

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

The first conclusion from these figures is that the out-of-sample empirical behavior of machine learning predictions is a strikingly close match to the VoC curves predicted by our theory. In particular, compare the empirical results of Figure 7 to the theoretical results under model misspecification from Figure 4. The beta estimates and out-of-sample R^2 demonstrate explosiveness at the interpolation boundary and recovery in the high-complexity regime. Figures IA1 and IA2 (reported in the Internet Appendix in the interest of space) document identical patterns for training windows of 60 and 120 months.

[38] We set $\gamma = 2$. Our results are generally insensitive to γ, as discussed in Section V.F.

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

[40] Our empirical R^2 calculation is one minus the ratio of out-of-sample forecast error variance to out-of-sample realized return variance. Our empirical Sharpe ratio calculation uses the centered standard deviation in the denominator.

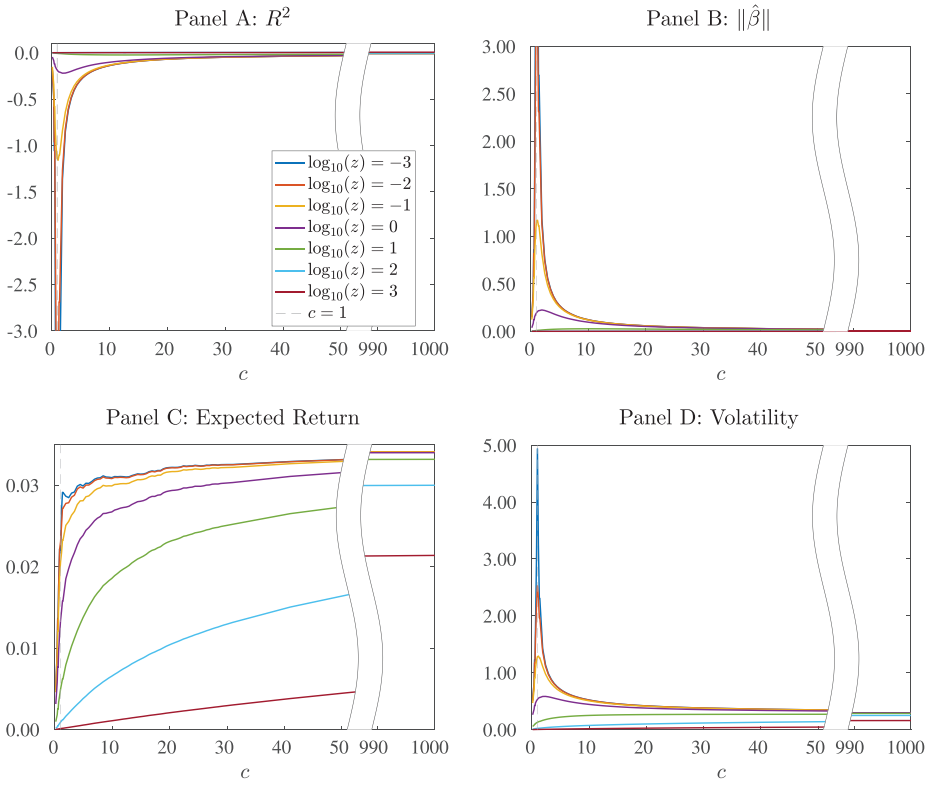

Figure 7. Out-of-sample market timing performance ($T = 12$). This figure shows the out-of-sample prediction accuracy and portfolio performance estimates for the empirical analysis described in Section V.C. The training window is $T = 12$ months and RFF count P (or cT) ranges from 2 to 12,000 with $\gamma = 2$. (Color figure can be viewed at wileyonlinelibrary.com)

Extreme behavior at the interpolation boundary makes it difficult to fully appreciate the patterns in R^2. Figure IA3 in the Internet Appendix provides more detail by plotting the out-of-sample R^2 zooming-in on the range $[-10\%, 1\%]$. Here, we see more clearly that high complexity and regularization together produce a positive out-of-sample R^2. In this plot, regularization comes in two forms, both directly through higher z and more subtly through higher c (which allows ridgeless regression to find solutions with small $\hat{\beta}$ norm). For large z, the R^2 is almost everywhere positive for all training windows.

The most intriguing aspect of Figure 7 is the clear increasing pattern in out-of-sample expected returns as model complexity rises. For $z = 10^{-3}$, which roughly approximates the ridgeless case, we see a nearly linear upward trend in average returns as c rises from zero to one. Beyond $c = 1$, the ridgeless expected return is nearly 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.

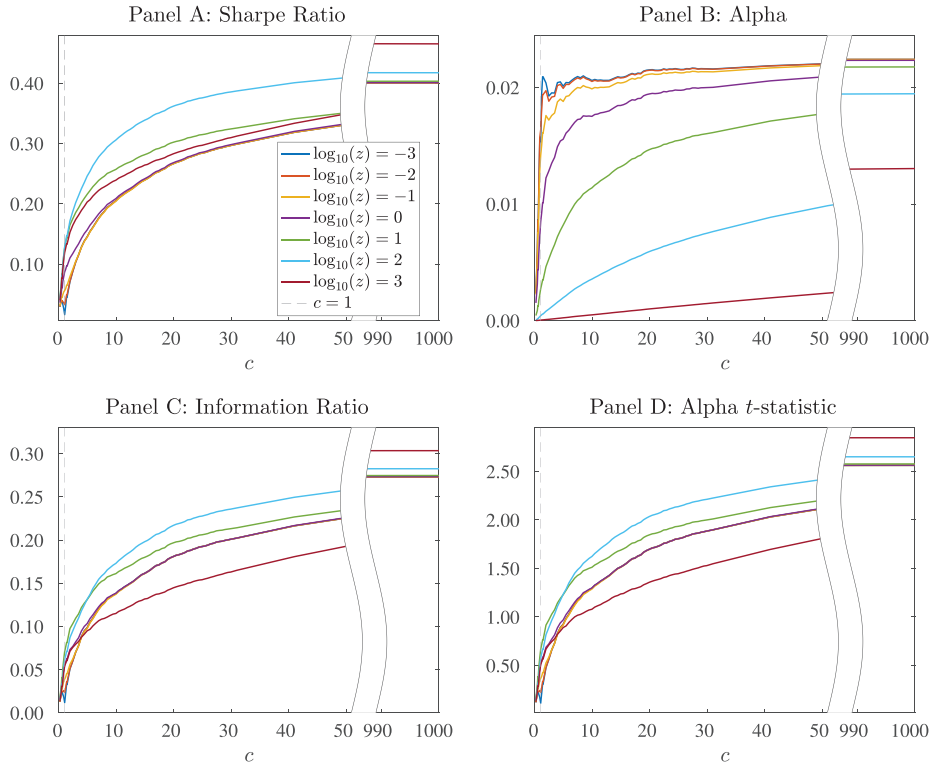

Figure 8. Out-of-sample market timing performance ($T = 12$). This figure shows the out-of-sample prediction accuracy and portfolio performance estimates for the empirical analysis described in Section V.C. The training window is $T = 12$ months and RFF count P (or cT) ranges from 2 to 12,000 with $\gamma = 2$. Alphas are versus a static position in the volatility-standardized market portfolio. (Color figure can be viewed at wileyonlinelibrary.com)

Internet Appendix Figures IA1 and IA2 again document an identical expected return pattern for longer training windows.

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

In our theoretical setting, we normalize the expected return of the untimed asset to zero. This is not the case of course for the U.S. market return. Therefore, to adjust for buy-and-hold market exposure, we calculate the out-of-sample alpha, alpha t-statistic, and information ratio (IR) of the timing strategy return via time-series regression on the untimed market. Figure 8 shows that the market timing alpha and IR inherit the same patterns as the average return and Sharpe ratio. In the high-complexity regime, we find IRs around

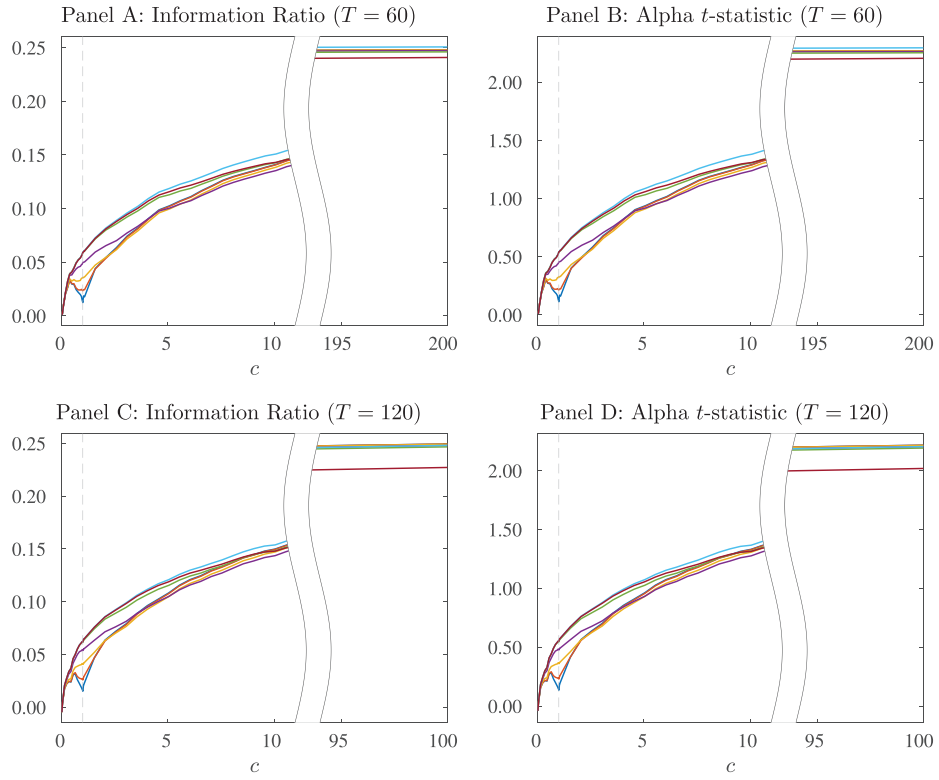

Figure 9. Out-of-sample market timing performance ($T = 60, 120$). This figure shows the out-of-sample prediction accuracy and portfolio performance estimates for the empirical analysis described in Section V.C. The training window is $T = 60$ or 120 months and RFF count P (or cT) ranges from 2 to 12,000 with $\gamma = 2$. Alphas are versus a static position in the volatility-standardized market portfolio. (Color figure can be viewed at wileyonlinelibrary.com)

0.3 and significant alpha t-statistics ranging from 2.6 to 2.9 depending on the amount of ridge shrinkage. Figure 9 repeats this analysis for training windows of 60 and 120 months, where we find similar IRs of roughly 0.25 with t-statistics above 2.0 for high-complexity models.

What do market timing strategies look like in the high-complexity regime? Figure 10 plots $\hat{\pi}(z, c)$ for the highest complexity and shrinkage configurations of our empirical model ($P = 12{,}000$ and $z = 10^3$, averaged across 1,000 sets of random feature weights). The three lines correspond to training windows of 12, 60, and 120 months. Positions show the same patterns for all training windows; their time-series correlations are 90% ($T = 12$ with $T = 60$), 87% ($T = 12$ with $T = 120$), and 97% ($T = 60$ with $T = 120$).[41] The plot shows six-month moving

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

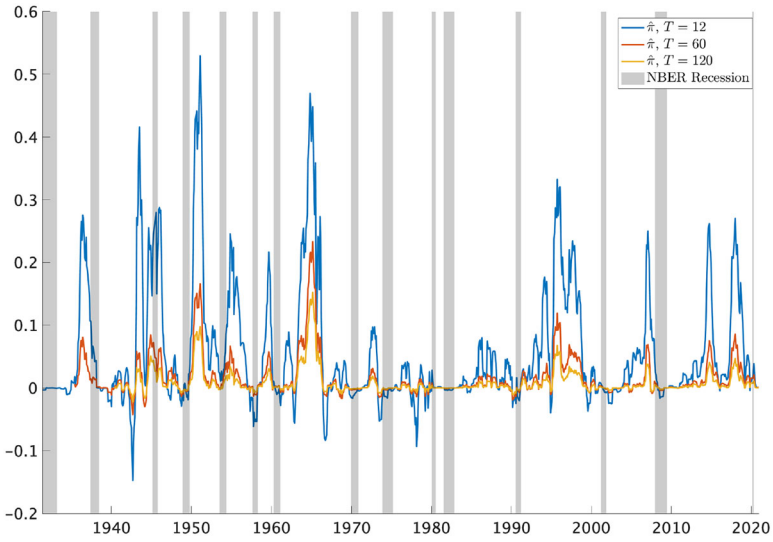

Figure 10. Market timing positions. This figure shows the out-of-sample prediction accuracy and portfolio performance estimates for the empirical analysis described in Section V.C. The training window is $T = 12$, 60, or 120 months with $P = 12{,}000$, $z = 10^3$, and $\gamma = 2$. Positions are averaged across 1,000 sets of random feature weights. Plots show the six-month moving average of positions to improve readability. (Color figure can be viewed at wileyonlinelibrary.com)

averages of raw positions for better readability (our trading results are based on the raw positions and not the moving averages).

The timing positions in Figure 10 are remarkable. First, they show that the high-complexity strategy is long-only at heart. Negative bets are infrequent and small relative to positive bets. The machine learning model thus heeds the guidance of Campbell and Thompson (2008) "that many predictive regressions beat the historical average return, once weak restrictions are imposed on the signs of coefficients and return forecasts." However, unlike Campbell and Thompson (2008), the machine seems to learn this rule without being given an explicit constraint.[42]

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 substantially reduces its position in the market before the recession (the exception is the eight-month recession of 1945). And it does this on a purely out-of-sample basis.

D. Comparison with Goyal and Welch (2008)

Our results seem at odds with the primary conclusion of Goyal and Welch (2008). These authors argue that the enterprise of market return prediction,

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

which has occupied a large amount of 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 as in that paper. What is the source of the discrepancy?

The conclusions of Goyal and Welch (2008) are based on their findings of consistently negative out-of-sample prediction R^2. They do not analyze the performance of timing strategies based on expected returns or Sharpe ratios.[43] We revisit their analysis with a focus on timing strategy performance using the same recursive out-of-sample prediction scheme as in the analysis of Figures 7 and 8. We use rolling 12-, 60-, and 120-month training windows (Panels A, B, and C, respectively), and we focus on a version of what Goyal and Welch (2008) call the "kitchen sink" regression. Our implementation uses 15 monthly predictors in a linear ridgeless regression.[44]

The first finding of Table I is that we confirm the conclusions of Goyal and Welch (2008). Note that with monthly data, a model with 15 regressors already has nontrivial complexity even for long training windows, and for the 12-month training window, its complexity even exceeds one. Monthly return forecasts using linear ridgeless regression behave egregiously. The monthly out-of-sample R^2 from ridgeless regression ($z = 0^+$) is large and negative at less than -100% (-9764% to be precise!). The timing strategy based on these predictions is also poor. The Sharpe ratio is -0.11 and is insignificantly different from zero. This seems perhaps not so terrible given the wildness of the forecasts, but it is due to the fact that the strategy's volatility is so high. Its maximum loss is 98 standard deviations. In light of our theoretical analysis, this agreement with the conclusions of Goyal and Welch (2008) is perhaps unsurprising. With $P = 15$ and $T = 12$, this analysis takes place near the interpolation boundary. Thus, forecasts and timing-strategy returns are expected to be highly volatile, as our estimates confirm. In Panels B and C, we repeat the analysis with longer training windows ($T = 60$ and 120). Longer training windows lead to less variable ridgeless regression estimates, producing higher (though still negative) R^2, and improving the Sharpe ratio.

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

[43] Updating the original Goyal and Welch (2008) analysis, Goyal, Welch, and Zafirov (2023) provide some evidence of timing-strategy performance for market return predictors.

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

Table I
Comparison with Goyal and Welch (2008)

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

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

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

The "Nonlinear" model in Table I refers to the machine learning timing strategy with $c = 1{,}000$ and $z = 10^3$ (averaged across 1,000 sets of random weight draws). In Panel A, the out-of-sample R^2 is 1% per month, with a Sharpe ratio of 0.46 and an IR of 0.31 versus the market. It also has a significant IR of 0.26 ($t = 2.5$) versus the best linear strategy ($z = 10^3$). One of the most attractive aspects of the machine learning strategy is its low downside risk. Its worst month

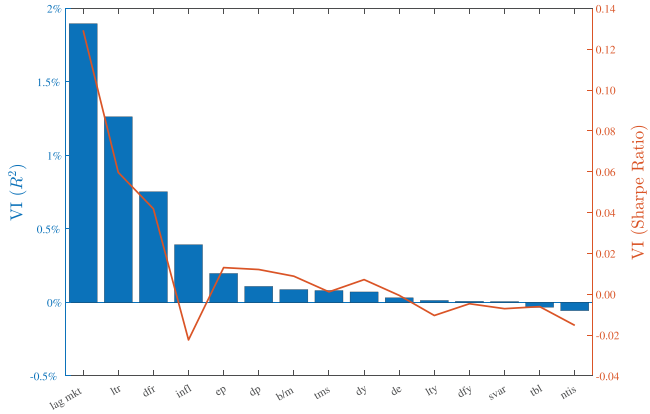

Figure 11. Variable importance. This figure shows the variable importance (VI) for the i^{th} predictor that is the change in performance, defined as out-of-sample R^2 or Sharpe ratio, moving from the full model with 15 variables to the reestimated model using 14 variables (excluding variable i). (Color figure can be viewed at wileyonlinelibrary.com)

was a loss of 1.23 standard deviations, and its skewness is positive, 2.48. These attractive tail risk properties of the machine learning model are not reflected in the Sharpe ratio. Still, they would be an important utility boost for investors who care about non-Gaussian risks. Note that the machine learning strategy accomplishes this using the identical information set as the linear strategy; it exploits this information in a high-dimensional, nonlinear way. Using longer training windows (Panels B and C) leads to the same conclusions.

E. Variable Importance

These results above beg the question: how can such large models learn predictive patterns in training windows as short as 12 months, particularly when several raw predictors are highly persistent (e.g., dividend yield and T-bill rate)? The short answer is that a number of the 15 raw predictors are, in fact, highly variable over short horizons, and these variables are the most important contributors to the performance of the high-complexity model. To shed more detailed light on this answer, we analyze the contribution of each variable to overall model performance. We reestimate the machine learning model omitting each of the 15 predictor variables one by one. We calculate "variable importance" (VI) for the i^{th} predictor as the change in performance (defined as out-of-sample R^2 or Sharpe ratio) moving from the full model with 15 variables to the reestimated model using 14 variables (excluding variable i).

Figure 11 plots the results for the 12-month training window (with $P = 12{,}000$, $z = 10^3$, and averaged across 1,000 sets of random feature weights). The three most important variables are also the three predictors with the highest average variation in 12-month windows (i.e., the least persistent

predictors).[45] Excluding the lagged market return ("lag mkt"), long-term bond return ("ltr"), or default return ("dfr") from the random features model reduces the out-of-sample monthly prediction R^2 by 1.9%, 1.3%, and 0.8%, respectively. In other words, the complex model is particularly adept at leveraging information in short-horizon fluctuations among predictors. The VI calculations tell the same story when we measure it in terms of R^2 (bars) or Sharpe ratio (line).

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

The VI results show that some linear predictors have impressive individual performance. To show that machine learning performance is not driven by these simple linear effects, Internet Appendix Table IA1 reports IRs of the machine learning strategy on the linear univariate timing strategy of each predictor (the univariate timing strategy is defined as the product of a predictor at time t with the market return at $t + 1$).

The machine learning model has a large and highly significant IR over every linear strategy. We also calculate its IR versus all 15 univariate strategies simultaneously ("All").[46] In this case, we find an IR of 0.32 ($t = 2.9$), providing further direct evidence for the nonlinear benefits of complexity.

Naturally, interpretation is a challenge for complex nonlinear models. Internet Appendix Figure IA5 makes progress in this direction by illustrating the nonlinear prediction patterns associated with each of the 15 predictors. To trace the impact of predictor i on expected returns, we fix the prediction model estimated from a given training sample and fix the values of all variables other than i at their values at the time of the forecast. Next, we vary the value of the i^{th} predictor from its full-sample min (corresponding to -1 in the plots) to its full-sample max (corresponding to $+1$) and record how the return prediction varies. We then average this prediction response function across all training windows and plot the result.

The figure illustrates a few interesting patterns. First, we see that when certain indicators of macroeconomic risk are at their lowest (in particular, stock market variance "svar" and credit spreads on risky corporate debt "dfy"), the machine learning model forecasts positive returns. However, once these variables reach even moderate levels, the return prediction drops to zero. This is

[45] Figure IA4 in the Internet Appendix reports the average variation of each predictor in 12-month training windows.

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

consistent with the time-series pattern in Figure 10, which shows that timing positions (i.e., expected returns) drop to zero heading into recessions. In fact, all predictors demonstrate a similar "risk on/risk off" predictive pattern in which certain values trigger positive market bets; otherwise, they advocate positions near zero.

F. The Extent of Nonlinearity and Other Robustness

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

$$\sin(\gamma\omega_i' G_t) = \gamma\omega_i' G_t + O(\gamma^2),\ \cos(\gamma\omega_i' G_t) = 1 - \gamma\omega_i' G_t + O(\gamma^2). \quad (21)$$

Suppose for simplicity that we only have the sin features. Then, defining $\Omega = \frac{1}{P^{1/2}}(\omega_i)_{i=1}^P \in \mathbb{R}^{15\times P}$, we have that the model is equivalent to a model with random *linear* features, $S_t = \Omega' G_t$.[47]

This begs the question: is there an optimal degree of nonlinearity? In general, the answer is no. In the high-complexity regime, different choices of γ deliver different approximations of the true DGP, with none strictly dominating the others. Mei, Misiakiewicz, and Montanari (2022) show that high model complexity poses an insurmountable obstacle for any random feature regression—it is impossible to learn the "true" dependency $R_{t+1} = f(G_t) + \varepsilon_{t+1}$ when the model is complex. In this case, different random feature generators recover different aspects (projections) of the truth on different subspaces. As a result, we would expect linear and nonlinear random features to contain complementary information. This is clearly reflected in the results of Table I.[48]

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

Next, we analyze the robustness of our main findings in subsamples. We report model performance splitting the test sample into halves, as shown in Internet Appendix Figures IA9, IA10, and IA11 for training windows $T = 12$, 60, and 120, respectively. The left side of each figure reports machine learning timing-strategy out-of-sample performance from 1930 to 1974, and the right

[47] See Proposition IA1 in Section V of the Internet Appendix.

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

side from 1975 to 2020. The figures show that the patterns of out-of-sample timing- strategy performance with respect to complexity and shrinkage do not depend on the subsample. Average out-of-sample returns rise monotonically with complexity and decrease with ridge shrinkage; volatility abates when we move past the interpolation boundary and is further dampened by shrinkage. IRs 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 in the first sample. But volatilities are roughly the same, so IRs are about half as large in the second sample. This is consistent with the machine's trading patterns plotted in Figure 10. Starting around 1968, the machine finds notably fewer buying opportunities and, when it does, takes smaller positions than in the earlier sample.

Finally, we compare the performance of the machine learning model with a 12-month training window to a 12-month time-series momentum strategy (Moskowitz, Ooi, and Pedersen (2012)). If regressors are highly persistent, they will appear roughly static in a typical 12-month window. In this case, forecasts from a high-complexity regression will behave very similarly to time-series momentum.[49] In Section VII of the Internet Appendix, we explain this issue in more detail. We also show that our results are not driven by this "short window and persistent regressor" mechanism. Instead, as emphasized in Section V.E, our machine learning model performance is driven by relatively high-frequency fluctuations among the predictors. We also show that the machine learning timing strategy has economically large and statistically significant alpha over time-series momentum.

VI. Conclusion

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

In this paper, we offer new theoretical insights 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

[49] We are grateful to the editor for pointing this out.

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.

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

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

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

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

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

Our theoretical analysis (along with that of Belkin et al. (2019), Hastie et al. (2022), and 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.

Initial submission: June 21, 2022; Accepted: December 16, 2022
Editors: Stefan Nagel, Philip Bond, Amit Seru, and Wei Xiong

REFERENCES

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

Ali, Alnur, J. Zico Kolter, and Ryan J. Tibshirani, 2019, A continuous-time view of early stopping for least squares regression, in Kamalika Chaudhuri and Masashi Sugiyama, eds., *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, 1370–1378 (Naha, Okinawa, Japan), PMLR.

Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song, 2019, A convergence theory for deep learning via over-parameterization, in Kamalika Chaudhuri and Ruslan Salakhutdinov, eds., *Proceedings of the 36th International Conference on Machine Learning*, 242–252 (Long Beach, California), PMLR 97.

Bai, Zhidong, and Wang Zhou, 2008, Large sample covariance matrices without independence structures in columns, *Statistica Sinica* 18, 425–442.

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

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

Belkin, Mikhail, Daniel Hsu, and Ji Xu, 2020, Two models of double descent for weak features, *SIAM Journal on Mathematics of Data Science* 2, 1167–1180.

Belkin, Mikhail, Alexander Rakhlin, and Alexandre B. Tsybakov, 2019, Does data interpolation contradict statistical optimality? In Kamalika Chaudhuri and Masashi Sugiyama, eds., *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, 1611–1619 (Naha, Okinawa, Japan), PMLR.

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

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

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

Chen, Luyang, Markus Pelger, and Jason Zhu, 2023, Deep learning in asset pricing, *Management Science*, Articles in Advance, 1–37.

Cochrane, John H., 2011, Presidential address: Discount rates, *Journal of Finance* 66, 1047–1108.

Da, Rui, Stefan Nagel, and Dacheng Xiu, 2022, The statistical limit of arbitrage, Working paper, Chicago Booth.

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

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

Du, Simon, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai, 2019, Gradient descent finds global minima of deep neural networks, in Kamalika Chaudhuri and Ruslan Salakhutdinov,

eds., *Proceedings of the 36th International Conference on Machine Learning*, 1675–1685 (Long Beach, California), PMLR.

Du, Simon S., Xiyu Zhai, Barnabas Poczos, and Aarti Singh, 2018, Gradient descent provably optimizes over-parameterized neural networks, Working paper, arXiv, Cornell University.

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

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

Fan, Jianqing, Zheng Tracy Ke, Yuan Liao, and Andreas Neuhierl, 2022, Structural deep learning in conditional asset pricing, Working paper, SSRN.

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

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

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

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

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

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

Goyal, Amit, Ivo Welch, and Athanasse Zafirov, 2023, A comprehensive 2021 look at the empirical performance of equity premium prediction II, Working paper, Swiss Finance Institute.

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

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

Hastie, Trevor, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani, 2022, Surprises in high-dimensional ridgeless least squares interpolation, *The Annals of Statistics* 50, 949–986.

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

Jacot, Arthur, Franck Gabriel, and Clément Hongler, 2018, Neural tangent kernel: Convergence and generalization in neural networks, *Advances in Neural Information Processing Systems* 31.

Kelly, Bryan, and Seth Pruitt, 2013, Market expectations in the cross-section of present values, *Journal of Finance* 68, 1721–1756.

Kelly, Bryan, and Dacheng Xiu, 2022, Financial machine learning, Working paper, Yale.

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

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

Ledoit, Olivier, and Sandrine Péché, 2011, Eigenvectors of some large sample covariance matrix ensembles, *Probability Theory and Related Fields* 151, 233–264.

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

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

Liu, Fanghui, Xiaolin Huang, Yudong Chen, and Johan A. K. Suykens, 2021, Random features for kernel approximation: A survey on algorithms, theory, and beyond, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 7128–7148.

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

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

Martin, Ian W. R., and Stefan Nagel, 2022, Market efficiency in the age of big data, *Journal of Financial Economics* 145, 154–177.

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

Mei, Song, and Andrea Montanari, 2022, The generalization error of random features regression: Precise asymptotics and the double descent curve, *Communications on Pure and Applied Mathematics* 75, 667–766.

Moskowitz, Tobias J., Yao Hua Ooi, and Lasse Heje Pedersen, 2012, Time series momentum, *Journal of Financial Economics* 104, 228–250.

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

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

Rapach, David, and Guofu Zhou, 2013, Forecasting stock returns, in Graham Elliott and Allan Timmermann, eds., Handbook of Economic Forecasting, volume 2, 328–383 (Elsevier).

Rapach, David, and Guofu Zhou, 2022, Asset pricing: Time-series predictability, *Oxford Research Encyclopedia of Economics and Finance*.

Rapach, David E., Jack K. Strauss, and Guofu Zhou, 2010, Out-of-sample equity premium prediction: Combination forecasts and links to the real economy, *Review of Financial Studies* 23, 821–862.

Rapach, David E., and Guofu Zhou, 2020, Time-series and cross-sectional stock return forecasting: New machine learning methods, *Machine Learning for Asset Management: New Developments and Financial Applications* 1–33.

Richards, Dominic, Jaouad Mourtada, and Lorenzo Rosasco, 2021, Asymptotics of ridge (less) regression under general source condition, in Arindam Banerjee and Kenji Fukumizu, eds., *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 3889–3897 (SanDiego, California, USA), PMLR.

Rudi, Alessandro, and Lorenzo Rosasco, 2017, Generalization properties of learning with random features, *Advances in Neural Information Processing Systems* 30.

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

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

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

Tsigler, Alexander, and Peter L. Bartlett, 2023, Benign overfitting in ridge regression, *Journal of Machine Learning Research* 24, 123–131.

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

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

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1: Internet Appendix.
Replication Code.