Seemingly Virtuous Complexity in Return Prediction [*]

Stefan Nagel[†]
University of Chicago,
NBER, CEPR, and CESifo

June 20, 2025

Return prediction with Random Fourier Features (RFF)—a very large number, P, of nonlinear transformations of a small number, K, of predictor variables—has become popular recently. Surprisingly, this approach appears to yield successful out-of-sample stock market index return predictions even when trained in rolling windows as small as $T = 12$ months with P in the thousands, and without shrinkage. However, this apparent virtue of complexity is an illusion. When $P \gg T$, the RFF-based predictive regression closely approximates a kernel ridgeless regression that takes the original K predictors as input. The resulting forecast is effectively a weighted average of past returns, with weights highest on periods whose predictor vectors are most similar to the current one. In short training windows, similarity simply means recency, so the forecast reduces to a weighted average of recent returns—essentially a momentum strategy. Moreover, similarity of current and past predictor vectors is decreasing in predictor volatility, which makes it a volatility-timed momentum strategy. Crucially, the volatility-timed momentum nature of the strategy arises mechanically from predictor similarity. The strong historical performance of the RFF-based strategy thus reflects the coincidental historical success of volatility-timed momentum, not predictive information extracted from training data.

[*] I am grateful for comments to Daniel Buncic, Alvaro Cartea, Darrell Duffie, Damir Filipovic, Amit Goyal, Leonid Kogan, Ralph Koijen, Serhiy Kozak, Markus Pelger, Ivo Welch, and Dacheng Xiu. Many thanks to Ruilin Wang for excellent research assistance.

[†] University of Chicago, Booth School of Business, 5807 South Woodlawn Avenue, Chicago, IL 60637, e-mail: stefan.nagel@chicagobooth.edu

I. Introduction

Research in machine learning has found that heavily overparametrizing models such that they fit the training data perfectly can result in good out-of-sample predictions. In a linear regression setting, this means that the number of covariates may far exceed the number of observations in the training data [Belkin, Hsu, Ma, and Mandal (2019), Bartlett, Long, Lugosi, and Tsigler (2020), Hastie, Montanari, Rosset, and Tibshirani (2022)]. These findings in machine learning have inspired a fast-growing literature in empirical asset pricing that uses Random Fourier Features (RFF)—a very large number, P, of randomized nonlinear transformations of a small number, K, of predictor variables—for return prediction and for modeling of the stochastic discount factor (SDF).[1]

Kelly, Malamud, and Zhou (2024) (KMZ), the pioneering paper using this approach, presents a stunning result. In a time-series setting for predicting excess returns on the CRSP value-weighed index, regressions with $P = 12,000$ RFF derived from $K = 15$ variables—predictors from Welch and Goyal (2008) augmented with lagged index returns—produce a market timing strategy with strong out-of-sample performance even with rolling training data windows as short as $T = 12$ months and in ridgeless regression, i.e., without explicit shrinkage. This is a stunning result because most of the Goyal-Welch predictor variables are highly persistent and the forecasting target is extremely noisy. Conventional wisdom holds that extracting useful predictive signals from such variables requires sample sizes spanning decades—not just a single year.

In line with conventional wisdom, I find that the high-complexity ridgeless regression using rolling training windows of $T = 12$ months and $P = 12,000$ RFF fails to extract meaningful predictive information from the training data. Instead, the resulting market-timing strategy effectively reduces to a certain volatility-timed momentum strategy. Crucially, the ridgeless regression does not learn from the training data that a volatility-timed momentum strategy is profitable. Rather, the resemblance to a volatility-timed momentum strategy arises mechanically from the structure of the RFF representation and the persistence of the underlying predictors. That this mechanically induced strategy happens to perform well in historical data is merely coincidental.

The reason the strategy based on thousands of RFF predictors collapses to a simple volatility-timed momentum strategy stems from the fundamental properties of RFF. When $P \gg K$, dot

[1] See Kelly, Malamud, and Zhou (2024), Jensen, Kelly, Malamud, and Pedersen (2024), Didisheim, Ke, Kelly, and Malamud (2023), Didisheim, Ke, Kelly, and Malamud (2024), Kelly, Malamud, and Zhou (2022),

products of RFF vectors converge to Gaussian kernels in the space of the original K predictors, as shown in Rahimi and Recht (2007) and Sutherland and Schneider (2015). Consequently, forecasts from ridgeless regression using RFF essentially equal those from kernel ridgeless regression with Gaussian kernels. In this setting, return forecasts are constructed as weighted averages of the $T = 12$ lagged returns in the training window, where the weights depend on the distance between the current K-dimensional predictor vector and each lagged predictor vector. For example, if the one-month lagged predictor vector is more similar to the current one than the two-month lagged predictor vector, then the corresponding one-month lagged return receives a higher weight in the construction of the return forecast than the two-month lagged return. Because most predictor variables are persistent, recent predictor vector observations tend to be more similar to the current one, leading to higher positive weights on recent returns—producing a momentum-like forecast. Moreover, when the predictor variables are less volatile, distances between predictor vectors are smaller, resulting in higher weights. Thus, the strategy embeds a form of volatility timing, where the overall strength of the momentum signal varies inversely with predictor volatility in the training window. The combination of these effects yields a volatility-timed momentum strategy.

This makes clear that the high-complexity ridgeless regression with RFF does not learn complex relationships between the K predictor variables and future returns from the short training data. The only information used in the construction of the market-timing strategy is the information encoded in the distances between lagged predictor vectors and the current predictor vector. In a short training sample, this information reflects two mechanical features. First, closer lags of predictor vectors have smaller distance to the current predictor vector. This is a property of any vector-autoregressive process with persistence and does not reflect predictive content that the predictor variables may have for future returns. Second, there are times when the distances between predictors are smaller than in other times. This reflects time-varying volatility of shocks to the predictors. This volatility does not embody predictive content of the predictors for future returns.

Conceptually, the method performs as intended. With a very long training sample, it would search across the historical record—potentially spanning decades—for instances in which the predictor vector resembled the current one, and use the returns following these instances to forecast future returns. However, when the training window is very short, as in KMZ's rolling approach with $T = 12$ months, the method cannot look for instances of predictor similarity in the distant past. As a result, it simply averages the most recent few returns in the training window, which correspond to the predictor vectors most similar to the current one.

Empirically, a simple volatility-timed momentum strategy that assigns linearly declining weights to the past 12 months returns, scaled by the inverse of predictor volatility over the same window, generates market-timing positions that closely resemble those of the RFF-based strategy and achieves comparable out-of-sample performance. When included as an explanatory factor, this simple strategy captures most of the abnormal returns attributed to the RFF-based approach. Like the RFF-based strategy, it tends to reduce exposure ahead of recessions—not because of any foresight, but because it systematically de-risks when predictor volatility increases.

To demonstrate that the ridgeless regression with RFF does not learn from the training data that a volatility-timed momentum strategy is profitable, I generate artificial return data by adding a simulated MA(2) component with strong negative autocorrelation to actual market index returns. These artificial returns now exhibit return reversals rather than momentum. Nevertheless, the RFF-based ridgeless regression continues to produce forecasts that place positive weights on recent returns. As a result, the corresponding market-timing strategy delivers negative abnormal returns out-of-sample on the artificial data. This illustrates that the RFF-based regression does not learn from the training data whether momentum or reversal dynamics are present; it mechanically imposes a momentum-like structure regardless of the underlying return process.

While most of my analysis focuses on the time-series predictive regression setting of KMZ, I also show that similar logic applies in a cross-sectional asset pricing setting when P RFF of K firm characteristics are used to construct RFF factors from a panel of stock returns. When $P \gg K$, a high-complexity SDF with thousands of RFF factors as in Didisheim, Ke, Kelly, and Malamud (2024) (DKKM) involves dot products that converge to Gaussian kernels of the K firm characteristics. The resulting mean-variance efficient portfolio weights implied by the estimated SDF are kernel-smoothed past stock returns, where the smoothing is based on similarity in firm characteristics across stock-time observations in the training panel. For example, the portfolio weight of stock n at time t then reflects an average from training-period stocks that exhibited similar characteristics as stock n at t. When the training sample is as short as $T = 12$ months and $P \gg T$, the closest matches in characteristic space are likely to be recent observations of stock n itself. This imparts a momentum-like nature to the strategy. As in the time-series setting, a volatility-timing component also emerges: lower volatility in the firm characteristics leads to tighter clustering in characteristic space, resulting in higher weights on past returns.

Beyond the specific settings of KMZ and DKKM, this analysis highlights a broader limitation that small training sample sizes impose on complex return prediction. When the true expected return function is high-dimensional, but sample sizes are small, estimators are constrained to explore only a very small subspace of the predictor space, leaving most of it unexamined. This limits variation in the signal, while noise is large. In a time-series setting, with training sample sizes as short as $T = 12$ months, little can be learned from the training data about the expected return function. But even with sample sizes spanning several decades, the available data may still be small relative to the dimensionality of the predictor space, especially when accounting for nonlinear transformations. This reflects a fundamental challenge in empirical asset pricing: the ratio of sample size to model complexity is likely far lower than in many other machine learning applications.

This paper connects to several others in the literature. Most directly related, several recent papers examine the analysis in KMZ. Berk (2023) raises the concern that the reported Sharpe ratios reported in KMZ are not actually achievable since they are actually averages across 1000 draws of the random weights in the construction of RFF. Relatedly, Buncic (2025) shows that the increasing relation between out-of-sample performance and complexity disappears when RFF-based return forecasts are first aggregated across different draws of RFF weights before computing the OOS performance measures rather than aggregating the OOS performance measures after constructing forecasts. In my analysis I focus on the ridgeless regression limit where P is very large and the choice of aggregation method has little effect. Cartea, Jin, and Shi (2025) shows that the measurement errors in return predictors can limit the virtue of complexity. Fallahgoul (2025) develops information-theoretic bounds on learning in KMZ's setting, showing that learning of complex predictive relationships is impossible with the training sample sizes employed by KMZ. These results are consistent with my analysis of the limitations imposed by sample size, but obtained with a different, complementary approach.

The kernel ridgeless regression representation used in my analysis connects to recent papers that explore kernel-based approaches in asset pricing. Kozak (2023) uses the kernel trick to represent dot products of high-dimensional nonlinear transformations of firm characteristics in terms of kernels.

Filipović and Pasricha (2022) use Gaussian process regression in asset pricing. Filipović, Pelger, and Ye (2022) employ kernel ridge regression in bond yield curve estimation.

The fact that sample size limits what can be learned about complex functions from training data relates to theoretical research that studies what economic decision makers can learn from data in high-dimensional settings, and how this affects equilibrium pricing. Martin and Nagel (2022) consider investors learning about asset fundamentals. Molavi, Tahbaz-Salehi, and Vedolin (2024) study the effects of limits on the complexity of models that investors can entertain. In Da, Nagel, and Xiu (2024), statistical arbitrageurs are learning about investment opportunities

II. Properties of Predictive Regressions with Random Fourier Features in the High-Complexity Case

I focus on the case of ridgeless regression, which yields the most striking results in KMZ. In this setting, the predictive regression appears to achieve substantial out-of-sample performance despite the absence of explicit ridge regularization and with the number of predictor variables, P, vastly outnumbering the number of observations used to estimate the regression, T. The analysis below centers on this high-complexity, small-sample ridgeless case, which generates the most surprising and counterintuitive findings in KMZ.

II.A. Preliminaries: Predictability Induced by Standardization of the Dependent Variable

KMZ standardize the dependent variable in the predictive regressions, and the return to be forecasted, with the standard deviation of returns over the previous 12 months. This can generate predictability that does not exist in the returns before standardization. Appendix A discusses the issue in more detail. The resulting bias in predictability is small, but to completely avoid it, I deviate from KMZ and work with returns that are not standardized.

II.B. Properties of Ridgeless Regression when $P > T$

Let T denote the number of observations used in rolling predictive regressions, and hence the length of the training window. The dependent variable observations are returns from $t-T$ to t, collected in the vector $\boldsymbol{r}_t$. The realizations of the P predetermined predictor variables during the training window are collected in the P columns of the $T \times P$ matrix $\boldsymbol{Z}_{t-1} = (\boldsymbol{z}_{t-1} \quad \boldsymbol{z}_{t-2} \quad ... \quad \boldsymbol{z}_{t-T})'$. I focus on the case $P > T$.

In a regression of returns on the predictor variables, the ridgeless OLS estimator in this case is

$$\hat{\boldsymbol{b}}_t = (\boldsymbol{Z}'_{t-1}\boldsymbol{Z}_{t-1})^{+}\boldsymbol{Z}'_{t-1}\boldsymbol{r}_t, \tag{1}$$

where $^{+}$ denotes the Moore-Penrose pseudoinverse. As the number of predictors is higher than the number of return observations in $\boldsymbol{r}_t$, the regression perfectly fits the training data in the window of length T.

With the predictor variable observations in period t collected in the vector $\boldsymbol{z}_t$, the predicted value for r_{t+1} is

$$\hat{r}^{\text{rff}}_{t+1|t} = \boldsymbol{w}'_t\boldsymbol{r}_t, \qquad \text{with} \quad \boldsymbol{w}'_t = \boldsymbol{z}'_t(\boldsymbol{Z}'_{t-1}\boldsymbol{Z}_{t-1})^{+}\boldsymbol{Z}'_{t-1}. \tag{2}$$

The predicted return, and hence the market timing position taken by this strategy, is therefore a weighted average of the T returns in $\boldsymbol{r}_t$. Considering that $P > T$ and using the rules for the Moore-Penrose pseudoinverse, we can rewrite the weights as

$$\boldsymbol{w}'_t = \boldsymbol{z}'_t\boldsymbol{Z}'_{t-1}(\boldsymbol{Z}_{t-1}\boldsymbol{Z}'_{t-1})^{-1}. \tag{3}$$

This provides an interpretation of the T weights: they represent the coefficients in a regression of P predictor variable observations in the vector $\boldsymbol{z}_t$ on lagged observations of the predictors in periods $t-1$ to $t-T$. Roughly speaking, this regression evaluates which of the past vectors of predictors $\boldsymbol{z}_{t-1}, \boldsymbol{z}_{t-2}, ..., \boldsymbol{z}_{t-T}$ is most similar to $\boldsymbol{z}_t$. The regression coefficients, and hence the weights $\boldsymbol{w}_t$ then reflect this similarity. For example, if the predictors follow an autoregressive process with persistence, a $\boldsymbol{z}_{t-k}$ that is closer in time to $\boldsymbol{z}_t$ will tend to be more similar to $\boldsymbol{z}_t$ than predictor vectors that are more distant in time. As a consequence, the elements of the weight vector $\boldsymbol{w}$ corresponding to small k will tend to be higher than those for larger k. When these weights are then applied to lagged returns in the construction of $\hat{r}_t = \boldsymbol{w}'_t\boldsymbol{r}_t$, this results in a a version of a momentum strategy, with higher weights on the most recent returns.

That the predicted return is a weighted average of past returns is always true for any predicted regression estimated on past returns in the high-complexity case where the number of predictor variables exceeds the number of observations in the training window. The question is whether the weights placed on these past returns reflect any predictive information that predictor variables have about future returns. As I will show next, when the predictors are constructed as RFF of a small number of variables, more can be said about how the information from predictors is used in construction of the weights.

II.C. Forecasts from Ridgeless Regression with Random Fourier Features Approximate Forecasts from Kernel Ridgeless Regression

KMZ construct $\boldsymbol{z}$ as a very large number of nonlinear transformations of a small number of predictor variables from the Goyal-Welch data set as predictors. Specifically, the nonlinear transformations take the form of Random Fourier Features (RFF). Results on the convergence of dot products of RFF help shed light on what happens to KMZ's ridgeless regression estimator when $P \gg T$.

The $K = 15$ predictor variables used by KMZ include 14 predictor variables from the Goyal-Welch data set augmented with the one-month lagged market index return. KMZ then form RFF where consecutive elements i and $i+1$ of $\boldsymbol{z}_t$ are constructed as

$$\begin{pmatrix} z_{i,t} \\ z_{i+1,t} \end{pmatrix} = \sqrt{\frac{2}{P}} \begin{pmatrix} \cos(\gamma \boldsymbol{\omega}_i' \boldsymbol{x}_t) \\ \sin(\gamma \omega_{i+1}' \boldsymbol{x}_t) \end{pmatrix}, \quad \boldsymbol{\omega}_i \sim \text{ IID N}(0, \boldsymbol{I}), \quad i = 1, ..., P/2, \tag{4}$$

with $\gamma = 2$. KMZ form up to 6,000 of such pairs.[2] Focusing on the highest number they consider, we then have $P = 2 \times 6,000 = 12,000$. KMZ then standardize the RFF, dividing by the within-training-window standard deviation of each RFF. I first ignore this standardization and will incorporate it in the next step.

Rahimi and Recht (2007) show that RFF can approximate kernels. Sutherland and Schneider (2015) show similar results for the RFF specification used by KMZ. The analysis in these papers focuses on the case of large training data sets. Kernel methods require the evaluation of $k(\boldsymbol{u}, \boldsymbol{v}) = k(\boldsymbol{u} - \boldsymbol{v})$ for every pair of datapoints. With large data sets, this entails a huge computational cost, and approximation of kernels with RFF can dramatically reduce computational cost. This is the typical use-case for RFF. The idea is to reduce computational complexity relative to direct computation of kernels. In contrast, in KMZ's setting, training data sets are small and direct

[2] KMZ do not pre-multiply by $\sqrt{2/P}$, but this is inconsequential for the resulting portfolio weights of the market timing strategy as this scalar factor cancels out in the weights (3).

computation of kernels is not a challenge. As shown in Rahimi and Recht (2007), when P is large relative to K, dot products of RFF accurately approximate a kernel

$$\boldsymbol{z}_t' \boldsymbol{z}_{t-k} \approx k(\boldsymbol{x}_t, \boldsymbol{x}_{t-k}) \tag{5}$$

that only takes the K original characteristics as inputs. That dot products in a high-dimensional space of P nonlinearly transformed features can be computed as a kernel that takes inputs in the lower-dimensional space of the K original features is also known as the *kernel trick* in machine learning.[3] Given the standard normal distribution of the weights in KMZ's construction of the RFF, the kernel will be a Gaussian kernel

$$k(\boldsymbol{x}_t, \boldsymbol{x}_{t-k}) = \exp\left(-\frac{\gamma^2}{2}\|\boldsymbol{x}_t - \boldsymbol{x}_{t-k}\|_2^2\right), \tag{6}$$

as shown in Sutherland and Schneider (2015).

With the notation

$$k(\boldsymbol{x}_t, \boldsymbol{X}_{t-1}) = \begin{pmatrix} k(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}), & ..., & k(\boldsymbol{x}_t, \boldsymbol{x}_{t-T}) \end{pmatrix} \tag{7}$$

and

$$K(\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-1}) = \begin{pmatrix} k(\boldsymbol{x}_{t-1}, \boldsymbol{X}_{t-1}) \\ ..., \\ k(\boldsymbol{x}_T, \boldsymbol{X}_{t-1}) \end{pmatrix} \tag{8}$$

we then get

$$\boldsymbol{z}_t' \boldsymbol{Z}_{t-1}' \approx k(\boldsymbol{x}_t, \boldsymbol{X}_{t-1}), \qquad \boldsymbol{Z}_{t-1}\boldsymbol{Z}_{t-1}' \approx K(\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-1}), \tag{9}$$

which leads to the result that the predicted value in ridgeless regression with a very large number $P > T$ of RFF is approximately equal to the predicted value in the ridgeless limit case of a kernel ridge regression[4]

$$\hat{r}_{t+1|t} \approx k(\boldsymbol{x}_t, \boldsymbol{X}_{t-1}) K(\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-1})^{-1} \boldsymbol{r}_t. \tag{10}$$

[3] See Kozak (2023) for an application of the kernel trick in empirical asset pricing.

[4] I refer to this regression as a kernel ridgeless regression, which is distinct from a kernel regression—e.g. based on the Nadarya-Watson estimator—that simply weights observations of the dependent variable without involving the $K(.,.)^{-1}$ matrix.

The kernel form of the estimator in (10) is revealing about the nature of KMZ's market-timing strategy that results from using the predicted returns as time-varying portfolio weights. The estimator constructs a prediction of future returns by smoothing past returns. The predicted value in (10) is a weighted average of the T lagged returns in $\boldsymbol{r}_t$, where the weights depend simply on the similarity between $\boldsymbol{x}_t$ and the T columns of $\boldsymbol{X}$. The function $k(\boldsymbol{x}_t, \boldsymbol{X}_{t-1})$ evaluates the similarity of the current predictor vector $\boldsymbol{x}_t$ with the predictor vectors in the training data contained in the columns of $\boldsymbol{X}_{t-1}$. If among these columns there are $\boldsymbol{x}_{t-k}$ for some lags k that are close in distance to $\boldsymbol{x}_t$, while others are not, then the prediction $\hat{r}_{t+1|t}$ is close to an average of the r_{t-k+1} observations at those lags k.

With a very long sample of training data, this estimator would do something economically reasonable. Given a current predictor vector $\boldsymbol{x}_t$, it would look for instances in the training data when the predictor vector was similar. Then it would use the realized returns in the months following these instances as the prediction of future returns associated with $\boldsymbol{x}_t$. Essentially, this is nonparametric smoothing based on local averaging, where locality is determined by an evaluation of the kernel. For example, consider using the price-dividend ratio as a single predictor variable in this approach. With a long training sample, the approach would effectively ask at which points in time in the past the price-dividend ratio took values similar to the current price-dividend ratio.

However, when applied to very short training windows, e.g. $T = 12$ months as in KMZ, this approach is unlikely to learn predictive information from the training data. In fact, when the set of predictors includes ones with strong persistence—as is the case for the Goyal-Welch predictor variables—then the weights on past returns in the construction of the predicted value in (10) take a mechanical form. In this case, there are two effects that largely account for how the weights on past returns look like. First, the less distant $\boldsymbol{x}_{t-k}$ from $\boldsymbol{x}_t$, the higher the weight that r_{t-k+1} gets in the construction of $\hat{r}_{t+1|t}$. If elements of $\boldsymbol{x}$ have strong persistence, similarity will be higher for predictor vectors closer in time to t, and hence weights will be higher on training window returns closer in time to t. This makes the weights resemble a momentum strategy: $\hat{r}_{t+1|t}$ will be an average of lagged returns in recent months leading up to and including t. Second, in times when the $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-k}$ are subject to bigger noise shocks—which will tend to happen during periods of high volatility—they will differ more, and hence $k(\boldsymbol{x}_t, \boldsymbol{x}_{t-k})$ will be smaller, resulting in a smaller weights on past returns in the construction of $\hat{r}_{t+1|t}$. The combination of the two effects is basically a momentum strategy interacted with a volatility-timing strategy that makes the momentum strategy less aggressive when volatility is high.

To illustrate how the weights on past returns look like, I use the replication data provided by KMZ to construct their estimator and the return predictions that follow. Figure I shows the weights $\boldsymbol{w}_t$ averaged over 1,000 draws of random weights for 12,000 RFF and focusing on the ridgeless regression case with $T = 12$. The percentiles are based on the distribution over time of the averaged weights across these 1,000 draws of RFF.[5]

Panel A shows the time-series mean of the weights for the different return lags as well as the 10th and 90th percentiles. The results show that the ridgeless regression basically forms a momentum strategy with the highest weight on the most recent lagged return y_t and smaller weights on earlier return observations. As discussed, this can be anticipated from the kernel ridge regression representation of the return prediction in (10). Panel B shows the cross-sectional mean of the $T = 12$ weights every month. There is strong time-variation in these weights. As I will show, this time-variation is closely related to the reciprocal of a volatility measure, which means that KMZ's market-timing strategy essentially amounts to a volatility-timed momentum strategy.

II.D. Within-Window Standardization of Random Fourier Features Leads to a Scaled Version of the Forecasts from Kernel Ridgeless Regression

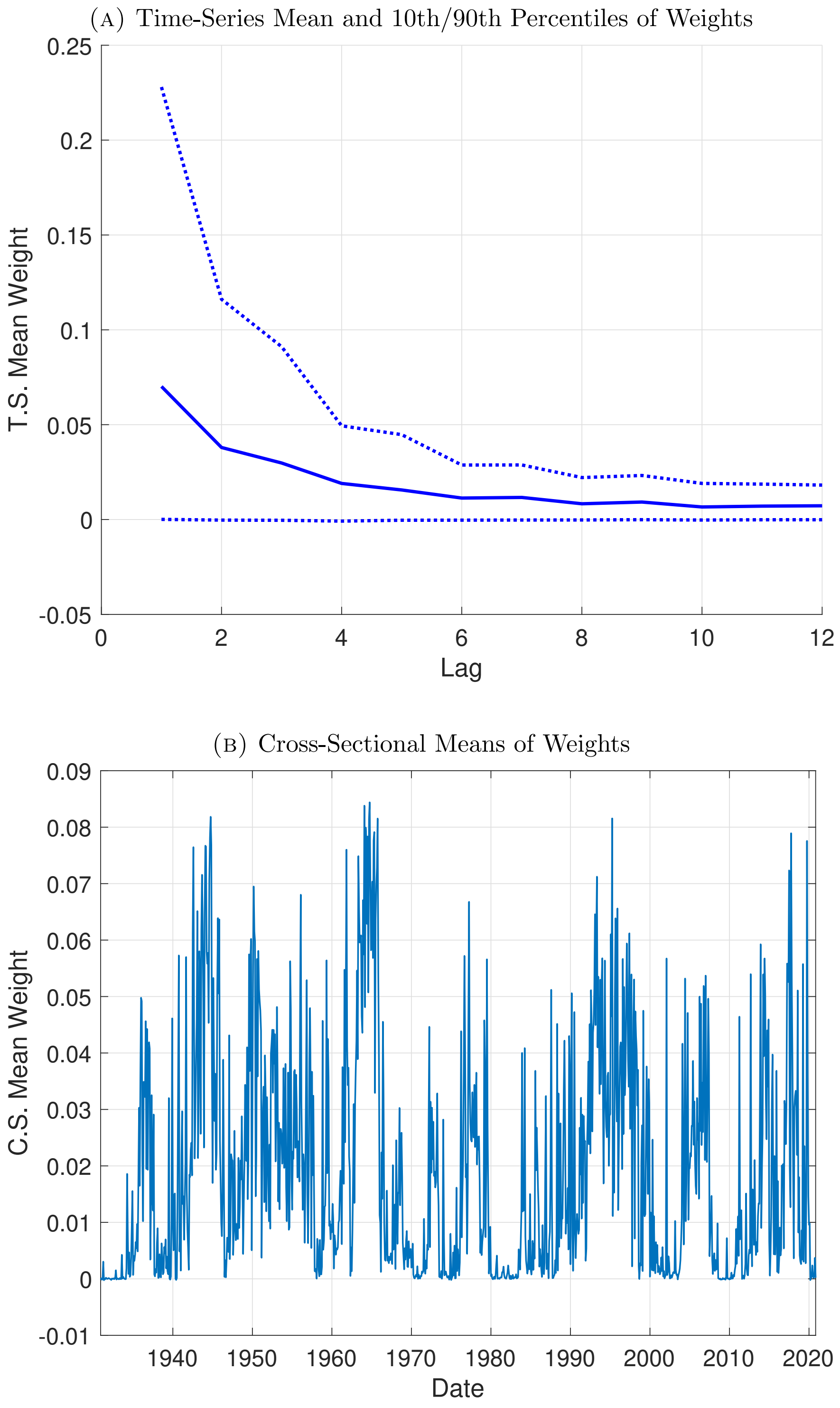
One issue that still remains to be addressed is that KMZ do not use the RFF directly as predictors. Instead, within each regression time window, they standardize the RFF by dividing with the within-window standard deviation of each RFF. Let $\tilde{\boldsymbol{z}}$ denote the standardized RFF. Their dot product is

$$\tilde{\boldsymbol{z}}_t'\tilde{\boldsymbol{z}}_{t-k}' = \boldsymbol{z}_t'\boldsymbol{\Omega}_{t-1}^{-1}\boldsymbol{z}_{t-k}, \tag{11}$$

where $\boldsymbol{\Omega}_{t-1} = \frac{1}{T-1}\,\mathrm{diag}(\boldsymbol{Z}_{t-1}'\boldsymbol{Z}_{t-1})$ is a diagonal matrix with the P within-window variances of the RFF on its diagonal. The results from Rahimi and Recht (2007) and Sutherland and Schneider (2015) on approximating kernels no longer apply to this weighted dot product of RFF.

While I do not have a closed-form result that relates the dot products of standardized RFF to kernels, empirically it turns out that in KMZ's setting the return prediction based on the

[5] The variation of in $\boldsymbol{w}_t$ for different draws of RFF random weights is miniscule. Hence, the mean weights and percentiles for a single draw of RFF random weights would look almost identical to Figure I.



(A) Time-Series Mean and 10th/90th Percentiles of Weights

(B) Cross-Sectional Means of Weights

Figure I

Weights on T Past Returns in Ridgeless Regression Return Prediction standardized RFF is extremely well approximated by a simple scaling up of the kernel in (10) to

$$\hat{r}^{\text{kernel}}_{t+1|t} \approx 1.69 \times k(\boldsymbol{x}_t, \boldsymbol{X}_{t-1}) K(\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-1})^{-1} \boldsymbol{r}_t. \tag{12}$$

I refer to the market timing strategy based on this scaled version of the kernel ridge regression as the kernel approach.

For the same RFF as those in Figure I, Figure II shows that the weights on past returns implied by KMZ's high-complexity regression with $P = 12,000$ RFF are almost exactly the same as those resulting from the kernel ridgeless regression approach in (12). Panel A shows the weights for the most recent lagged return r_t in the construction of $\hat{r}_{t+1|t}$. As the weight on the most recent lagged return is typically the biggest, this weight is the most important one. As Panel A shows, the weights on r_t in the RFF-based ridgeless regression (horizontal axis) are almost the same as those implied by the kernel ridgeless regression (vertical axis). The correlation is 0.99. Panel B shows similar results for the lag 2 return, i.e., the weight on r_{t-1}. For more distant lags not shown in the figure, the correlations between the different versions of weights are extremely high, too. This confirms that KMZ's ridgeless regression with thousands of RFF essentially boils down to a kernel ridgeless regression that takes a small number of K inputs.

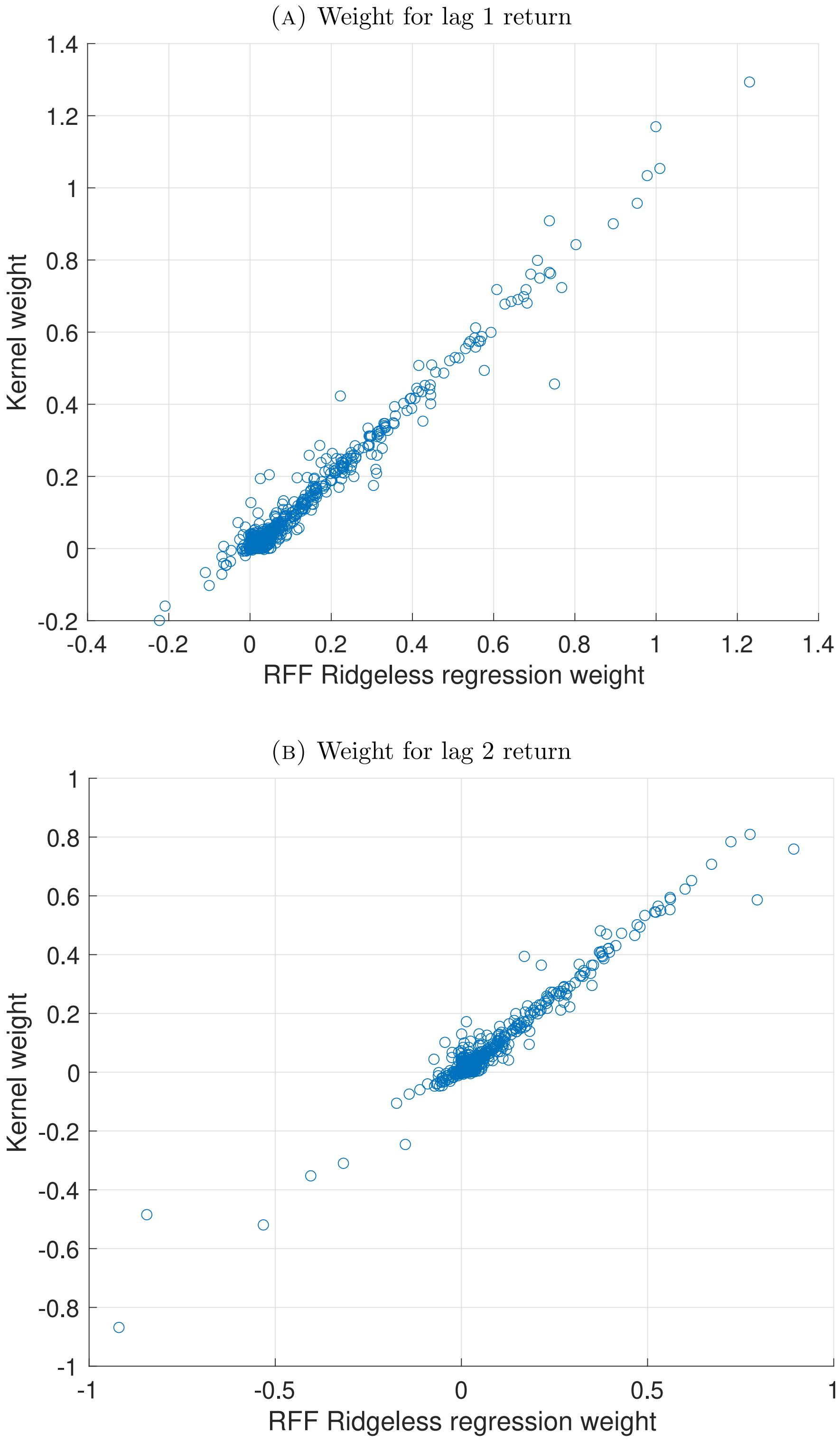
Figure III reinforces this conclusion. It shows the predicted returns, $\hat{r}^{\text{rff}}_{t+1|t}$ produced by the high-complexity regression and $\hat{r}^{\text{kernel}}_{t+1|t}$ produced by the kernel approach. They are very similar. The correlation is 0.93.

II.E. Nature of the Market-Timing Strategy when T is Small: Effectively a Volatility-Timed Momentum Strategy

The weight that the predicted return $r^{\text{kernel}}_{t+1|t}$ puts on return r_{t-k+1} depends, via the Gaussian kernel

$$k(\boldsymbol{x}_t, \boldsymbol{x}_{t-k}) = \exp\left(-\frac{\gamma^2}{2} \|\boldsymbol{x}_t - \boldsymbol{x}_{t-k}\|_2^2\right) \tag{13}$$

on the distance between $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-k}$. Broadly speaking, this distance reflects two effects. First, lags that are more distant in time (larger k) are less similar, which results in greater distance, and lower $k(\boldsymbol{x}_t, \boldsymbol{x}_{t-k})$. Second, the more volatile the predictors are in the training window, the greater the distance between $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-k}$. The two effects combined produce a momentum strategy (higher


(A) Weight for lag 1 return

(B) Weight for lag 2 return

Figure II
Weights on Past Returns implied by Ridgeless Regression and Scaled Kernel Ridgeless Regression

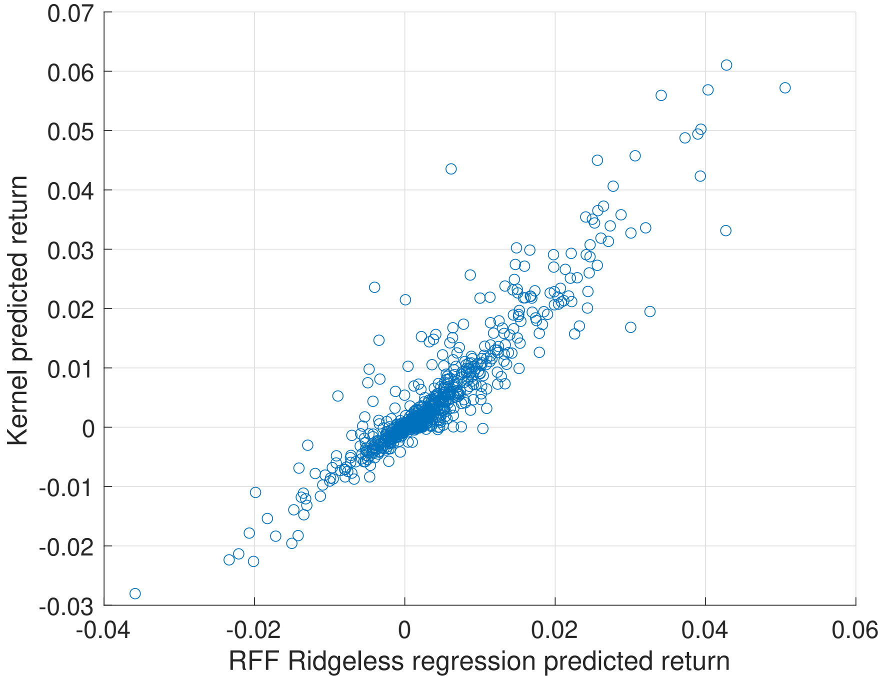


FIGURE III

Predicted Returns Implied by KMZ's Ridgeless Regression and Kernel Approach

positive weights for nearby lags of returns) that is volatility-timed (higher weights when predictor volatility is low).

Importantly, in a short training window with persistent predictors, the weights will, mechanically, always have this volatility-timed momentum form. The kernel ridgeless regression does not learn from the data that a volatility-timed momentum effect exists. This pattern of weights akin to a volatility-timed momentum strategy simply reflects the property of persistent predictor vectors that similarity is decreasing in time distance, and that the similarity is lower if predictor volatility is high. These properties of predictor vectors are always true irrespective of whether a volatility-timed momentum effect exists in the data or not. With much longer training windows the situation would be different. Then not only a few recent predictor vectors could have a high degree of similarity to the current one, but predictor vectors in the more distant past—years or decades ago—could be similar, too, as persistent predictor variables slowly cycle through their empirical range. By smoothing the return data based on predictor vector similarity, the regression could then learn nonlinear relationships between predictors and returns. But with windows as short as $T = 12$, this cannot happen.

Figure IV presents empirical evidence on the negative relation between predictor volatility and

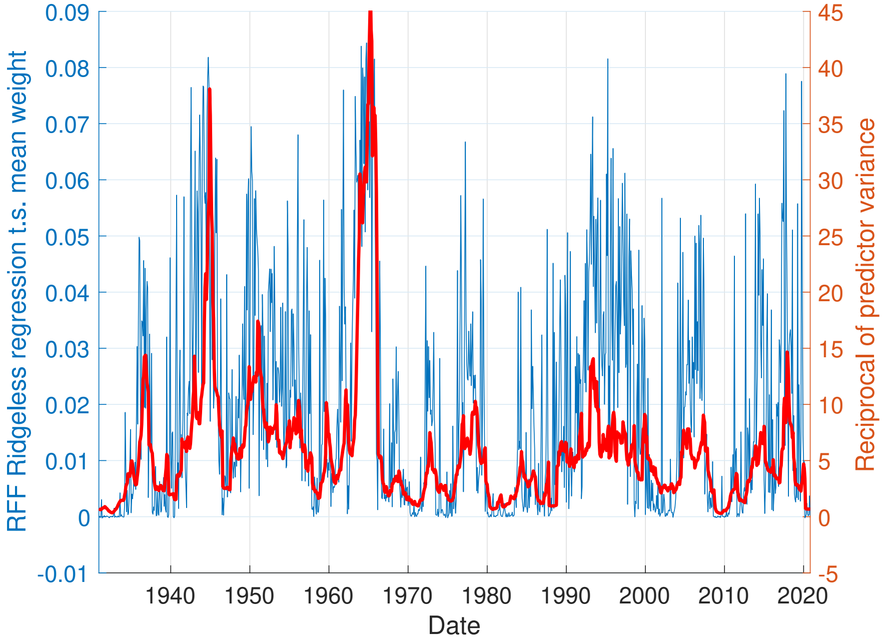


FIGURE IV
Ridgeless Regression Mean Weight on Past Returns and Reciprocal of Mean Predictor Variance

the magnitude of weights on past returns. The figure shows, for each month t, the mean of the elements of the weight vector $\boldsymbol{w}_t$ in (3) that multiplies the 12 training window returns in the construction of $\hat{r}_t^{\text{rff}}$. This is the same time series as in Panel B of Figure I. For comparison, the figure also shows the reciprocal of the mean of the variances of the 15 predictor variables within each training window. The figure shows that the mean weights on past returns produced by the ridgeless RFF strategy are highly correlated over time with the reciprocal mean predictor variance, consistent with the reasoning I outlined above: low predictor volatility implies higher weights assigned to recent past returns in the RFF-based strategy.

These results suggest that it should be possible to approximate KMZ's market timing strategy with a very simple one: a momentum strategy with weights on past returns that decay with lag length, similar to the decay of the time-series mean of weights shown in Panel A of Figure I, combined with a volatility-timing scaling factor that increases the magnitude of weights on past returns if the volatility of the predictor variables is low. The following market-timing strategy implements this idea:

$$\hat{r}_{t+1|t}^{\text{volmom}} = 0.05 \times \frac{1}{\hat{\sigma}_{x,t-1}^2} \times \sum_{k=0}^{11} \frac{12-k}{78} r_{t-k+1}. \tag{14}$$

Here $\hat{\sigma}_{x,t-1}^2$ is the average variance of the $K = 15$ predictors in the training window of length

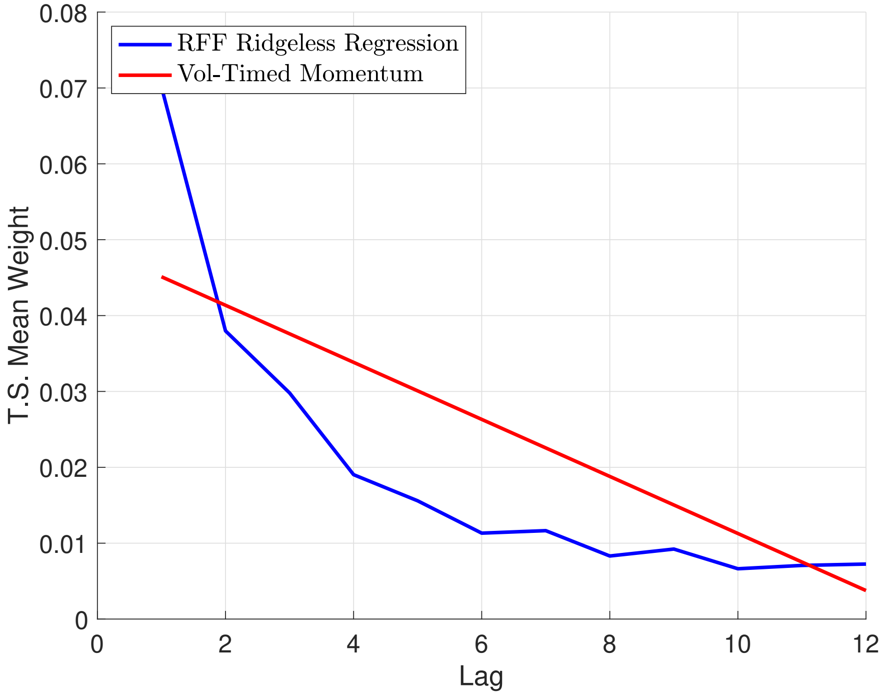


FIGURE V

Mean Weight on Past Returns in High-Complexity Ridgeless Regression with RFF and in the Volatility-Timed Momentum Strategy

$T = 12$. The weights on the past returns in the summation term are linearly declining with the lag and the division by 78 scales the weights inside the summation to have a sum of unity. The initial multiplication by 0.05 makes the average magnitude of the market-timing position $\hat{r}^{\text{volmom}}_{t+1|t}$ similar to the average magnitude of $\hat{r}^{\text{rff}}_{t+1|t}$, but this scaling has no effect on t-statistics and information ratios of the out-of-sample returns of the strategy.

Figure V compares the linearly declining weights of the volatility-timed momentum strategy with the weights implied by the KMZ's RFF-based strategy. They are not identical, but the broad pattern of decline with lag length is similar.

Figure VI shows the means of the past-return weights of the two strategies each month. Much of the time-series variation is shared between the two series.

II.F. Spanning Tests: Any Virtue in Complexity?

The analysis so far suggests that KMZ's RFF-based strategy should essentially be similar to one based on a kernel ridgeless regression that takes the original $K = 15$ predictor variables as inputs. In this section, I show that the kernel-based approach indeed produces very similar market-timing returns. The above analysis further suggests that KMZ's strategy is approximately similar to a

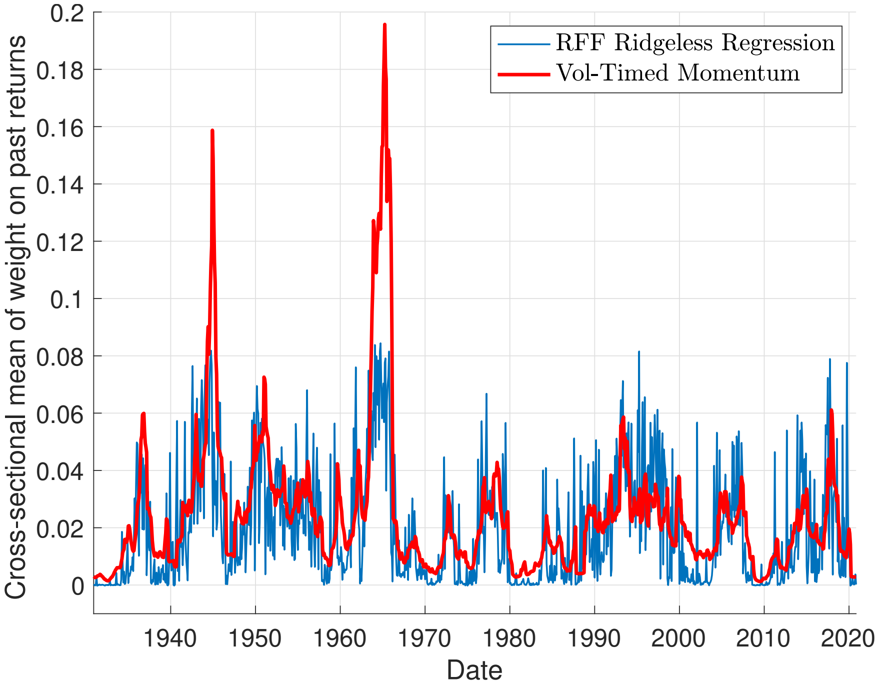


FIGURE VI
Mean Weight on Past Returns in High-Complexity Ridgeless Regression with RFF and Volatility-Timed Momentum Strategy

simple volatility-timed momentum strategy. In this section, I show that such a simple volatility-timed momentum strategy indeed gets close to explaining the returns earned by KMZ's strategy.

Panel A in Table I reports the alpha of each strategy relative to a one-factor model with the excess return of the CRSP value-weighted index as the single factor, as well as the corresponding t-statistics and information ratios.[6] The first column shows the result from KMZ that the high-complexity RFF strategy produces positive alpha, with high t-statistic (2.417) and information ratio (0.255).[7] As expected, the abnormal returns of the strategy based on kernel ridgeless regression are quite similar to the abnormal returns of the RFF-based strategy. Interestingly, the volatility-timed momentum strategy produces t-statistics and information ratios that are even slightly larger than those of the high-complexity RFF strategy. This means that the volatility-timed momentum strategy could potentially explain the high returns of the RFF-based strategy.

[6] For the RFF-based approach, I follow KMZ's method of calculating statistics for each draw of the random weights in the RFF construction, followed by averaging the statistics across the draws of random weights. Buncic (2025) criticizes this approach. I stick to it here to preserve comparability with KMZ's Figure 8.

[7] The t-statistic and information ratio are slightly lower than those that KMZ report in their Figure 8 for the ridgeless case ($\log_{10} z = -3$) and high complexity ($c = 1000$). The reason for this discrepancy is that I do not standardize the variable to be predicted to avoid the bias discussed in Appendix A. If I standardize returns in the same way as KMZ do, I obtain a t-statistic of 2.811 and an information ratio of 0.296, which exactly matches the results in KMZ's Figure 8.

TABLE I

Out-of-Sample Market Timing Performance

The RFF market timing strategy in the first column uses RFF as predictors and it is the same as in the ridgeless regression case in KMZ with $T = 12$ and $P = 12,000$ RFF, but without standardizing the predicted return, for the reasons discussed in Appendix A. The abnormal returns of the RFF-based market timing strategy are averaged over 1,000 draws of the random weights in the construction of RFF. The second column shows the market timing strategy based on the Gaussian kernel ridgeless regression in (12). The third column shows results for the volatility-timed momentum strategy in (14). Panel A shows alphas and information ratios relative to a one-factor model with the monthly excess return of the CRSP value-weighted index as the single factor. Panel B uses a two-factor model with the CRSP value-weighted index and the return on the kernel-based strategy as the two factors. Panel C uses the volatility-timed momentum strategy as the second factor along with the CRSP value-weighted index. Alphas are annualized in percent.

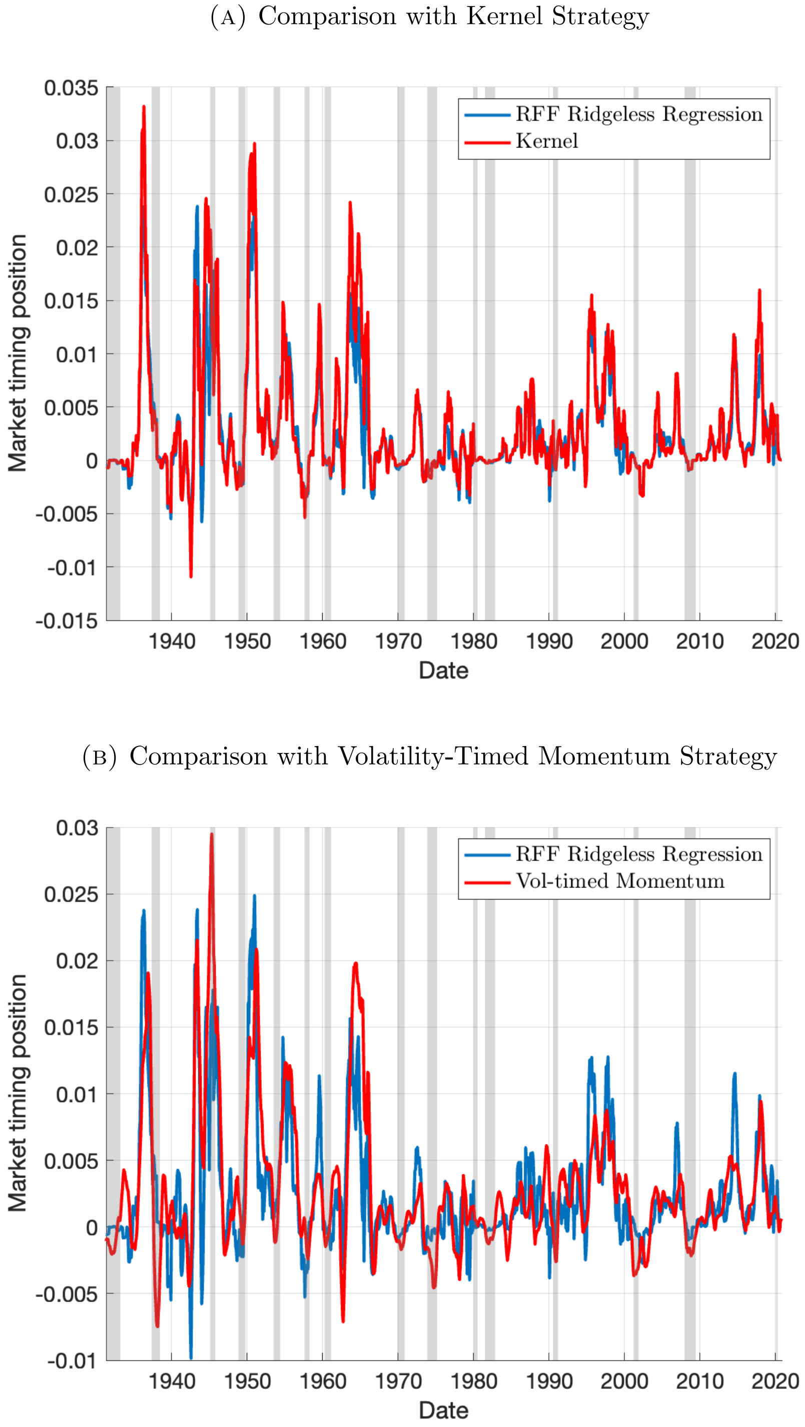
	High-Complexity RFF	Kernel	Vol-Timed Momentum
Panel A: One-Factor α (Market Factor)			
Alpha	0.034	0.040	0.034
(t-stat.)	(2.417)	(2.900)	(3.684)
Information Ratio	0.255	0.306	0.388
Panel B: Two-Factor α (Market and Kernel Factors)			
Alpha	-0.001		
(t-stat.)	(-0.122)		
Information Ratio	-0.013		
Panel C: Two-Factor α (Market and Vol-Timed Momentum Factors)			
Alpha	0.012		
(t-stat.)	(0.945)		
Information Ratio	0.100		

This is further reinforced by the spanning test in Panel B. Here I add the kernel-based strategy return as a second factor to the market factor. The alpha of the high-complexity RFF strategy drops to almost exactly zero, and the t-statistic and information ratios become very small. This is a natural outcome, given how similar the market timing positions of the two strategies are. Panel A in Figure VII shows that every month, the kernel-based strategy takes almost exactly the same position as the high-complexity RFF strategy.

Panel C uses the volatility-timed momentum strategy as second factor. Here the alpha doesn't drop as much, but it still falls by about 2/3 compared with Panel A and the t-statistic falls far below conventional levels of significance. Relatedly, Panel B of Figure VII shows that the market timing positions of the volatility-timed momentum strategy are very close to those of the high-complexity RFF strategy. This suggests that the volatility-timed momentum strategy, despite its extreme simplicity, explains much of the returns earned by the high-complexity RFF strategy.

Figure VII shows that both the kernel-based strategy and the volatility-timed momentum strategy tend to reduce market exposure ahead of recessions. KMZ interpret this pattern in the RFF-based strategy as evidence that "the machine learning strategy learns to divest leading up to recessions." However, the volatility-timed momentum structure of the weights on past returns in the RFF-based approach suggests a more mechanical explanation. The observed business-cycle pattern arises because the strategy places greater weight on past returns during periods of low predictor volatility—periods that typically fall outside recessions. This behavior is not the result of the machine learning that volatility-timing is beneficial, but rather a mechanical consequence of how the RFF-based strategy assigns weights based on predictor vector similarity in a short training window.

Figure VIII shows the cumulated abnormal returns of the three strategies with abnormal returns measured relative to the one-factor model with the market factor. The cumulated abnormal returns are divided by the full-sample standard deviation, which makes them interpretable as proportional to a cumulated information ratio. The volatility-timed momentum strategy overall has somewhat stronger performance than the others, but it is almost perfectly in sync with them in terms of the periods in which it gains and loses. Interestingly, almost all the gains for the three strategies accrued until 1970s. This could perhaps explain why the volatility-timed momentum for the aggregate market index is not a well-known anomaly.


(A) Comparison with Kernel Strategy

(B) Comparison with Volatility-Timed Momentum Strategy

Figure VII
Market Timing Positions of RFF-Based Ridgeless Regression, Kernel, and Volatility-Timed Momentum Strategy

Six-month moving averages of market timing positions $\hat{r}^{\text{rff}}_{t+1|t}$, $r^{\text{Kernel}}_{t+1|t}$, and $r^{\text{Volmom}}_{t+1|t}$ with $T = 12$, $P = 12,000$. The market timing positions $\hat{r}^{\text{rff}}_{t+1|t}$ are averaged over 1,000 draws of random weights in the construction of RFF.

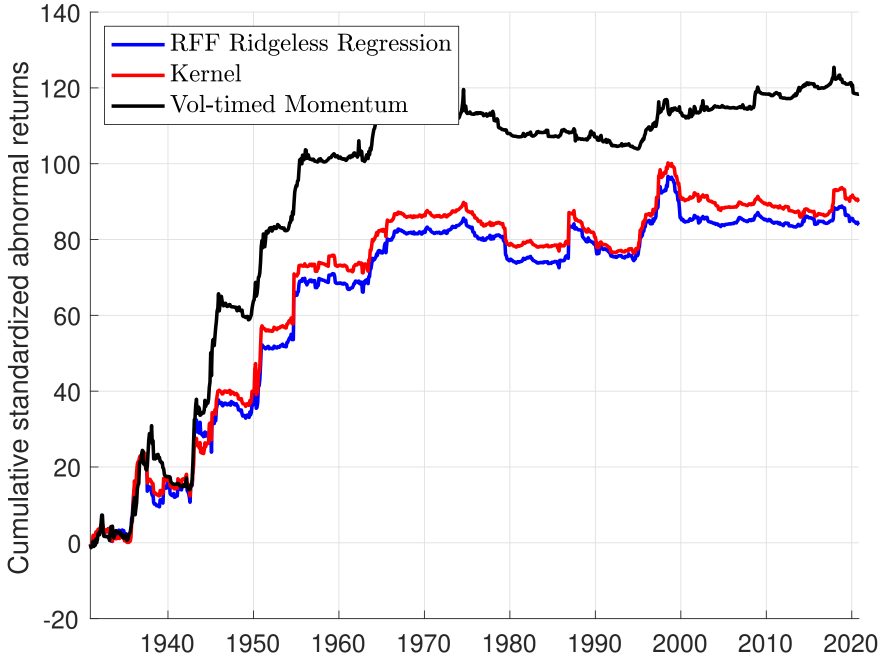


Figure VIII
Cumulative Standardized Out-of-Sample Abnormal Returns

Abnormal returns relative to one-factor model with the excess return of the CRSP value-weighted index as single factor, divided by the full-sample standard deviation of the abnormal return. The abnormal returns of the market timing strategy is based on high- ridgeless regression with $T = 12$ and $P = 12,000$ RFF, averaged over 1,000 draw of random weights in the construction of RFF.

II.G. Out-of-Sample Market Timing Performance in Artificial Data with Reversals

The ridgeless regression in KMZ's RFF-based approach does not learn from the data that a volatility-timed momentum effect exists; rather, it mechanically produces a volatility-timed momentum strategy regardless of the underlying properties of the return data. This structure emerges because, in short training windows with persistent predictors, the similarity of predictor vectors leads to systematically higher weights on recent returns during periods of low predictor volatility. The strong out-of-sample performance is therefore not the result of the model uncovering genuine predictive relationships, but a coincidence: a volatility-timed momentum effect happens to be present in the data, and the RFF-based strategy inadvertently exploits it.

While the mathematical argument is clear, it is still useful to demonstrate the point empirically. To that end, I modify the return data so that a volatility-timed momentum strategy no longer delivers positive returns. An algorithm that genuinely learns from the data—detecting whether or not volatility-timed momentum is present—should adapt and avoid pursuing such a strategy when applied to the modified data. In contrast, an algorithm that constructs a volatility-timed momentum

TABLE II
Out-of-Sample Market Timing Performance in Artificial Data with Reversals

The RFF market timing strategy in the first column uses RFF as predictors and it is the same as in the ridgeless regression case in KMZ with $T = 12$ and $P = 12,000$ RFF, but without standardizing the predicted return, for the reasons discussed in Appendix A, and here in applied to artificial data. The abnormal returns of the RFF-based market timing strategy are averaged over 1,000 draws of the random weights in the construction of RFF. The second column shows the market timing strategy based on the Gaussian kernel ridgeless regression in (12). The third column shows results for the volatility-timed momentum strategy in (14). Panel A shows alphas and information ratios relative to a one-factor model with the monthly excess return of the CRSP value-weighted index as the single factor. Panel B uses a two-factor model with the CRSP value-weighted index and the return on the kernel-based strategy as the two factors. Panel C uses the volatility-timed momentum strategy as the second factor along with the CRSP value-weighted index. Alphas are annualized in percent.

	High-Complexity RFF	Kernel	Vol-Timed Momentum
Panel A: One-Factor α (Market Factor)			
Alpha	-0.143	-0.137	-0.124
(t-stat.)	(-1.835)	(-1.757)	(-4.099)
Info Ratio	-0.193	-0.185	-0.432
Panel B: Two-Factor α (Market and Kernel Factors)			
Alpha	-0.013		
(t-stat.)	(-0.517)		
Information Ratio	-0.055		
Panel C: Two-Factor α (Market and Vol-Timed Momentum Factors)			
Alpha	-0.045		
(t-stat.)	(-0.628)		
Information Ratio	-0.066		

strategy for purely mechanical reasons, independent of predictive patterns in the training sample, will continue to do so even when it is no longer profitable.

Specifically, I modify the return data to such that it exhibits short-term reversals instead of momentum by adding an MA(2) process with negative autocorrelation to the original returns r_t. The artificial data of monthly market index returns then is

$$\tilde{r}_t = r_t + \xi_t - \theta_1 \xi_{t-1} - \theta_2 \xi_{t-2}, \tag{15}$$

with $\theta_1 = \theta_2 = 0.2$ and where $\xi_t \sim \mathrm{N}(0, 0.01)$. I chose the parameters of the MA(2) process such that the negative autocorrelation is big enough to overcome the momentum effect in the actual market index returns. I then replace the market index returns in KMZ's data with this artificial return series and redo the estimation.

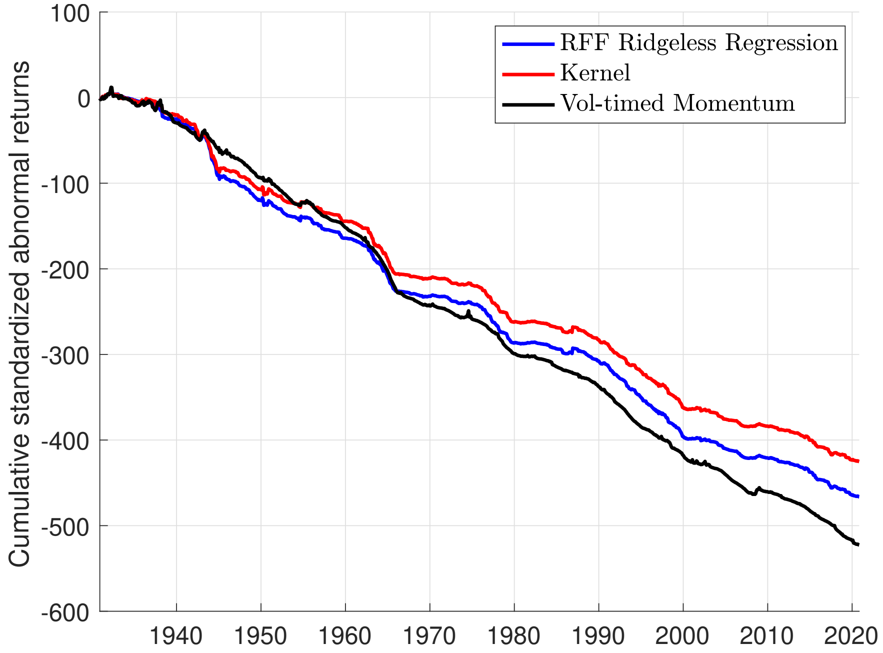


FIGURE IX

Cumulative Standardized Out-of-Sample Abnormal Returns in Artificial Data with Reversals

Abnormal returns relative to one-factor model with excess return of CRSP value-weighted index as single factor, divided by full-sample standard deviation of the abnormal return. The abnormal returns of the market timing strategy based on high-complexity ridgeless regression with $T = 12$ and $P = 12,000$ RFF are averaged over 1,000 draw of random weights in the construction of RFF.

Table II replicates the analysis from Table I, but using the artificial market index return data. As expected, the abnormal performance of the RFF-based strategy, the kernel-based strategy, and the volatility-timed momentum strategy is now uniformly negative. This confirms that the RFF-based approach continues to construct portfolio weights akin to a volatility-timed momentum strategy, despite the poor performance of such a strategy in the altered data. The RFF-based model does not learn from the data whether momentum (in the original data) or reversal (in the artificial data) dynamics are present. Instead, the structure of the weights is driven entirely by predictor-vector similarity. Even in the presence of return reversals, the most recent predictor vectors remain the most similar to the current one, leading the model to assign the highest positive weights to recent past returns—regardless of their predictive value.

Figure IX illustrates the out-of-sample performance in cumulative terms. In sharp contrast to Figure IX, the cumulative abnormal performance is negative. This reflects the inability of the RFF-based ridgeless regression, and the kernel approach that it approximates, to learn about the presence of reversals in the data in short training windows of only 12 months. These approaches still mechanically produce a volatility-timed momentum strategy, which performs badly in this artificial data.

III. Cross-Sectional Asset Pricing

In the analysis above, I focused on understanding the time-series predictability results in KMZ, which are especially puzzling given the conventional wisdom regarding the limitations of small-sample predictive regressions for stock market index returns. However, the findings also offer insight into the behavior of models that employ a large number of RFF-based factors in a cross-sectional asset pricing setting. While there are notable parallels with the time-series case, important differences also emerge, reflecting the distinct structure of cross-sectional prediction problems.

Consider now an unbalanced panel setting as in Didisheim, Ke, Kelly, and Malamud (2024) (DKKM), with $n = 1, ..., N_s$ stocks at time s, each with a vector of $j = 1, ..., K$ characteristics $\boldsymbol{x}_{n,s}$, stacked into a $N_s \times K$ matrix $\boldsymbol{X}_s = (\boldsymbol{x}_{1,s}, ..., \boldsymbol{x}_{N_s,s})'$. The researcher uses a training sample of length T, with $T < N_s$ for all s.

Construct RFF based on these characteristics for each stock n as

$$\begin{pmatrix} z_{n,i,s} \\ z_{n,i+1,s} \end{pmatrix} = \sqrt{\frac{2}{P}} \begin{pmatrix} \cos(\gamma \boldsymbol{\omega}_i' \boldsymbol{x}_{n,s}) \\ \sin(\gamma \omega_{i+1}' \boldsymbol{x}_{n,s}) \end{pmatrix} \quad \boldsymbol{\omega}_i \sim \text{ IID N}(0, \boldsymbol{I}), \quad i = 1, ..., P/2 \tag{16}$$

and, in each cross-section s, place the RFF in a $P \times N_s$ matrix $\boldsymbol{Z}_s$. DKKM randomly generate different values for γ for each i from a grid $[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$. Here I assume that γ is non-random.

Let $\boldsymbol{r}_{s+1}$ be an N_s-dimensional vector of stock returns. Forming cross-products of returns and lagged RFF delivers a P-dimensional vector of RFF factors

$$\boldsymbol{f}_{s+1} = \boldsymbol{Z}_s \boldsymbol{r}_{s+1}. \tag{17}$$

As in DKKM, I assume that the SDF has a representation as

$$M_s = 1 - \boldsymbol{\lambda}' \boldsymbol{f}_s. \tag{18}$$

Let $\hat{E}_t[.]$ denote a sample average in the training data sample from $t - T + 1$ to t. Solving for the minimum-norm solution of the sample moment conditions $\hat{E}_t[\boldsymbol{f}_s M_s] = 0$ delivers the ridgeless estimator of the prices of risk

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_t &= \left(\hat{E}_t[\boldsymbol{f}_s \boldsymbol{f}_s']\right)^+ \hat{E}_t[\boldsymbol{f}_s] \\ &= \left(\boldsymbol{F}_t' \boldsymbol{F}\right)^+ \boldsymbol{F}_t' \boldsymbol{\iota},\end{aligned} \tag{19}$$

where $\boldsymbol{\iota}$ is a conformable vector of ones and $\boldsymbol{F}_t = (\boldsymbol{f}_{t-T+1}, ..., \boldsymbol{f}_t)'$ has dimension $T \times P$.

Using the same approach as in (3), I can write $\hat{\boldsymbol{\lambda}}_t$ as

$$\hat{\boldsymbol{\lambda}}_t = \boldsymbol{F}_t' \left(\boldsymbol{F}_t \boldsymbol{F}_t'\right)^{-1} \boldsymbol{\iota}. \tag{20}$$

As usual, the prices of risk in the SDF that prices excess returns are the coefficients in a projection of 1 onto the space of factor excess returns (Hansen and Richard 1987). In the ridgeless case with $P > T$, the in-sample fit is perfect, so that $\boldsymbol{F}_t \hat{\boldsymbol{\lambda}}_t = \boldsymbol{\iota}$ and the fitted in-sample SDF is zero every period in the training sample.[8]

The prices of risk vector $\boldsymbol{\lambda}$ is proportional to the weights of the mean-variance efficient combination of the RFF factors. The estimated implied mean-variance efficient portfolio weights for the underlying N_t stocks, which can be applied out-of-sample to $\boldsymbol{r}_{t+1}$, are therefore proportional to

$$\boldsymbol{\omega}_t = \boldsymbol{Z}_t' \hat{\boldsymbol{\lambda}}_t = \boldsymbol{Z}_t' \boldsymbol{F}_t' \left(\boldsymbol{F}_t \boldsymbol{F}_t'\right)^{-1} \boldsymbol{\iota}. \tag{21}$$

As before, we can use the fact that dot products of large RFF feature vectors approximate kernels to write the estimator in terms of kernels. Specifically,

$$\begin{aligned}\boldsymbol{Z}_t' \boldsymbol{F}_t' &= \left[\boldsymbol{Z}_t' \boldsymbol{Z}_{t-T} \boldsymbol{r}_{t-T+1}, \quad ... \quad , \boldsymbol{Z}_t' \boldsymbol{Z}_{t-1} \boldsymbol{r}_t\right] \\ &\approx \left[K(\boldsymbol{X}_t, \boldsymbol{X}_{t-T}) \boldsymbol{r}_{t-T+1}, \quad ... \quad , K(\boldsymbol{X}_t, \boldsymbol{X}_{t-1}) \boldsymbol{r}_t\right],\end{aligned} \tag{22}$$

where $K(.,.)$ again denotes a Gaussian kernel matrix, as earlier, and the approximation follows for the same reasons I discussed earlier in Section II.C.

The $T \times T$ matrix $\boldsymbol{F}_t \boldsymbol{F}_t'$ has on its diagonal $\boldsymbol{f}_s' \boldsymbol{f}_s$, for $s = t - T + 1, ..., t$. These dot products

[8] In Section 4.5 in the theory part of their paper, DKKM discuss that the properties of the very large $P \times P$ matrix $\boldsymbol{F}_t' \boldsymbol{F}$ are difficult to characterize. However, in the ridgeless case, calculation of this matrix is not necessary—all we need is the much smaller $T \times T$ matrix $\boldsymbol{F}_t \boldsymbol{F}_t'$.

have the following approximation, again for the reasons discussed earlier in Section II.C:

$$\begin{aligned} \boldsymbol{f}_s'\boldsymbol{f}_s &= \boldsymbol{r}_s'\boldsymbol{Z}_{s-1}'\boldsymbol{Z}_{s-1}\boldsymbol{r}_s \\ &\approx \boldsymbol{r}_s' K(\boldsymbol{X}_{s-1}, \boldsymbol{X}_{s-1})\boldsymbol{r}_s. \end{aligned} \tag{23}$$

Off-diagonal elements are

$$\begin{aligned} \boldsymbol{f}_s'\boldsymbol{f}_{s-k} &= \boldsymbol{r}_s'\boldsymbol{Z}_{s-1}'\boldsymbol{Z}_{s-1-k}\boldsymbol{r}_{s-k} \\ &\approx \boldsymbol{r}_s' K(\boldsymbol{X}_{s-1}, \boldsymbol{X}_{s-1-k})\boldsymbol{r}_{s-k}, \end{aligned} \tag{24}$$

i.e., a kernel-weighted sum of serial and cross-serial comoments of returns at lag k. As serial correlation of returns is very small, and squared expected returns are small relative to the second moments of returns, the matrix $\boldsymbol{F}_t\boldsymbol{F}_t'$ can be approximated by setting the off-diagonal elements to zero. Writing it in terms of kernels makes transparent that the RFF-based estimator has a lot of similarity with the kernel trick approach in Kozak (2023).

With this approximation and (22), the weight vector in (21) becomes

$$\boldsymbol{\omega}_t \approx \sum_{s=t-T+1}^{t} \frac{K(\boldsymbol{X}_t, \boldsymbol{X}_{s-1})\boldsymbol{r}_s}{\boldsymbol{r}_{s+1}' K(\boldsymbol{X}_s, \boldsymbol{X}_s)\boldsymbol{r}_{s+1}}. \tag{25}$$

To discuss what the estimator is effectively doing, it is useful to focus on the element of the weight vector that applies to stock n,

$$\omega_{n,t} \approx \sum_{s=t-T+1}^{t} \frac{K(\boldsymbol{x}_{n,t}, \boldsymbol{X}_{s-1})\boldsymbol{r}_s}{\boldsymbol{r}_s' K(\boldsymbol{X}_{s-1}, \boldsymbol{X}_{s-1})\boldsymbol{r}_s}. \tag{26}$$

To determine the weight for stock n at the end of period t, this approach looks back at the panel of stock returns in the training data and constructs a weighted average of past returns of stocks. Consider one term in the summation. In the weighted average in the numerator, the kernel assigns higher weights to returns of stock observations with characteristics that are similar to $\boldsymbol{x}_{n,t}$, applying kernel smoothing to these returns. The denominator introduces a (co-)variance timing effect on the weights. If return volatility is high in period s, and especially when stocks with similar characteristics have similar returns, the denominator will be large, pulling towards zero the contribution of the summation term in (26) that corresponds to training sample period s.

DKKM find that their method produces high out-of-sample Sharpe ratios even with $T = 12$ months, although their baseline analysis finds higher Sharpe ratios with $T = 360$. As in the time-series setting, training windows as short as $T = 12$ raise the risk that the strategy behaves like a form of momentum strategy that performs well in the historical sample by coincidence but is not truly learned from the data. In this case, the portfolio weight assigned to stock n at time t depends a weighted average of returns across the entire return panel, where the weights favor stock-time observations with firm characteristics most similar to $\boldsymbol{x}_{n,t}$. Given the persistence of firm characteristics and the short training window, stock n's own past returns are likely to receive the most weight, imparting a strong momentum component. If other stocks have exhibited similar characteristics, their returns may also receive weight, introducing elements of group momentum or factor momentum. In addition, the denominator in (26) introduces a (co)variance-timing effect, increasing the influence of returns from periods with lower (co)variances. Taken together, the resulting strategy resembles a volatility-timed momentum strategy, shaped more by mechanical similarity and volatility patterns than by learned return predictability.

When the training window is substantially longer—as in DKKM's baseline analysis with $T = 360$—the strategy becomes more nuanced and is less likely to resemble a mechanical momentum-style approach. With $T = 360$, the estimator effectively scans the multi-decade return panel for stock-time observations whose firm characteristics resemble $\boldsymbol{x}_{n,t}$ and then smooths the associated returns to construct the numerator in (26). A longer training window increases the likelihood that relevant characteristics-similar observations are found in the more distant past, beyond the stock's own recent history, enabling the estimator to learn from a broader cross-section of the data about the typical returns linked to a given set of characteristics. The fact that DKKM report substantially higher out-of-sample Sharpe ratios with $T = 360$ than with $T = 12$ supports the interpretation that the estimator is capturing meaningful patterns in the data rather than simply reproducing the returns of a mechanical momentum-like strategy.

IV. Sample Sizes in Empirical Asset Pricing Limit Learning of Complex Functions

The evidence presented earlier in this paper shows that the out-of-sample performance of KMZ's market-timing strategy does not reflect predictive relationship learned from the data and therefore does not shed light on the virtue of complexity in return prediction. To be clear, this does not mean that there is no virtue of complexity in return prediction, but rather that KMZ's empirical analysis does not demonstrate the virtue of complexity. Conceptually, there remains a sound rationale for expecting models with a large number of predictors to outperform overly artificially sparse alternatives, provided appropriate regularization is applied, either explicitly through shrinkage or implicitly via minimum-norm estimators, such as in ridgeless regression. The key question for empirical asset pricing, however, is whether the benefits of model complexity can compensate for the severe limitations imposed by small sample sizes.

The training data sets available to asset pricing researchers are small by the standards of most machine learning applications. Much of the theoretical literature on the double-descent phenomenon, including the theoretical analysis in KMZ, focuses on asymptotic settings in which both T and P grow large. However, such asymptotic analysis does not address the central question facing empirical asset pricing: whether any meaningful predictability can be extracted from training samples of limited size. Conventional wisdom holds that windows as short as 60 or even 12 months—as employed in KMZ—are insufficient to uncover reliable return predictability. In this section, I examine this question focusing in finite-sample analysis.

Consider a time series of returns generated as

$$r_{t+1} = \boldsymbol{z}_t'\boldsymbol{b} + e_{t+1}, \qquad e_{t+1} \sim \mathrm{N}(0, \sigma^2), \tag{27}$$

where the P predictors are drawn as

$$\boldsymbol{z}_t \sim \mathrm{N}(0, \boldsymbol{I}) \tag{28}$$

every period t and the coefficients as

$$\boldsymbol{b} \sim \mathrm{N}\left(0, \frac{1}{P}\boldsymbol{I}\right), \tag{29}$$

so that $\mathbb{E}[\boldsymbol{b}'\boldsymbol{b}] = 1$. This setting uncorrelated predictors and normally distributed shocks is a special case of the DGP in the theoretical analysis of KMZ. The (infeasible) maximum squared Sharpe ratio achievable with perfect knowledge of $\boldsymbol{b}$ (see Appendix B.1) is

$$SR^2_{max} \approx \frac{1}{\sigma^2 + 2}. \tag{30}$$

This is an unconditional Sharpe ratio in the sense that it is not conditioned on $\boldsymbol{Z}$, consistent with an empirical approach as in KMZ that calculates the mean and variance of market-timing returns based on forecasts from rolling training windows with varying $\boldsymbol{Z}$. The Sharpe ratio in (30) does not depend on P. Therefore, by changing P, we can change the complexity of the true expected return function without changing the maximum achievable Sharpe ratio.

Let $\boldsymbol{Z}$ denote the matrix of $T \times P$ features and $\boldsymbol{r} = \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{e}$ the vector of returns observed in a training window of length $T < P - 1$. I focus first on the well-specified case where the econometrician uses all the P predictor variables present in the true model of expected returns. The ridgeless regression estimator of $\boldsymbol{b}$ is

$$\hat{\boldsymbol{b}} = \boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{r}. \tag{31}$$

With all predictor variables at the time of prediction, after the end of the training window, collected in $\boldsymbol{z}_t$, the estimated market-timing weight is

$$\hat{r}_t = \boldsymbol{z}_t'\hat{\boldsymbol{b}}. \tag{32}$$

A strategy based on these weights earns a squared Sharpe ratio (see Appendix B.2) of

$$SR^2 \approx \left(\frac{T}{P}\right) \frac{1}{\sigma^2 \left[1 + \frac{P+2+\sigma^2 P}{P-T-1}\right] + 3 - \frac{T}{P}}. \tag{33}$$

Several cases are instructive to consider here. First, as T approaches $P - 1$ from below, the variance term inside the brackets in the denominator of SR^2 explodes, causing the Sharpe ratio to collapse to zero. This gives rise to a double-ascent pattern in the Sharpe ratio, mirroring the more commonly discussed double-descent behavior of the squared forecast error. In this sense, there is a qualitative virtue of complexity: for fixed T, increasing P beyond the critical threshold where it is close to T improves the Sharpe ratio.

Second, even when $P \gg T$ and virtuous complexity is exploited to the fullest extent, the attainable Sharpe ratio remains tiny when T is small. In the empirically relevant case for studies like KMZ—where P is large and T is small—the Sharpe ratio remains far below its theoretical maximum SR_{max}. For example, with $T = 50$, $\sigma = 5$, $P = 1000$ we have $SR_{max} = 0.1925$, or about 0.67 on an annualized basis.[9] Yet in this case, equation (33) implies that Sharpe ratio of the the market-timing strategy based on ridgeless regression forecasts is only $SR = 0.0084$, roughly 1/25 of SR_{max}. Thus, while higher P helps lift the Sharpe ratio from zero when P is close to T to higher levels, these higher levels are still tiny. Ultimately, the size of the training sample imposes a binding constraint on the achievable Sharpe ratio.

To understand why small T limits the Sharpe ratio, it useful to see how the estimator processes the information in the signal component, $\boldsymbol{Zb}$, in the training data returns $\boldsymbol{r} = \boldsymbol{Zb} + \boldsymbol{e}$. We can write the ridgeless estimator as

$$\hat{\boldsymbol{b}} = \boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{b} + \text{noise}. \tag{34}$$

The first term on the right-hand side represents a projection of the P-dimensional vector $\boldsymbol{b}$ onto the T-dimensional random subspace spanned by the T rows of $\boldsymbol{Z}$. As a result, even abstracting from any distortions introduced by noise, the estimator can recover only a weighted average of the true coefficient vector $\boldsymbol{b}$, but not the full $\boldsymbol{b}$. In other words, much of the information in $\boldsymbol{b}$ remains unidentifiable from the limited sample.[10] As a consequence, non-zero forecast error will arise if $T < P$ even in a noiseless setting where $\boldsymbol{r} \approx \boldsymbol{Zb}$. This highlights a fundamental limitation imposed by training sample size. If it is indeed true that the true model of expected return is complex, then, with P in the thousands, it is impossible to learn much about the expected return function from samples as small as $T = 12$ or $T = 60$, even in the absence of noise. From the training data, the regression can only learn about return predictability in the T directions represented by the subspace spanned by the T rows of $\boldsymbol{Z}$. Other dimensions remain unexplored. As a consequence, the attainable squared Sharpe ratio based on the estimated strategy will be less than T/P times

[9] Scaling e_t and the elements of $\boldsymbol{z}_t$ by the same constant would not alter the Sharpe ratio.

[10] This is related to the notion that the effective complexity of a prediction model can be quantified with the effective degrees of freedom (EDF) of a fitted model [Efron (1986), Hastie and Tibshirani (1987)] calculated as the trace of the hat-matrix, which in this case is $\operatorname{tr}\left(\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\right) = T$. Fallahgoul (2025) uses learning theory to make related observations about the limitations induced by training sample size in KMZ's setting.

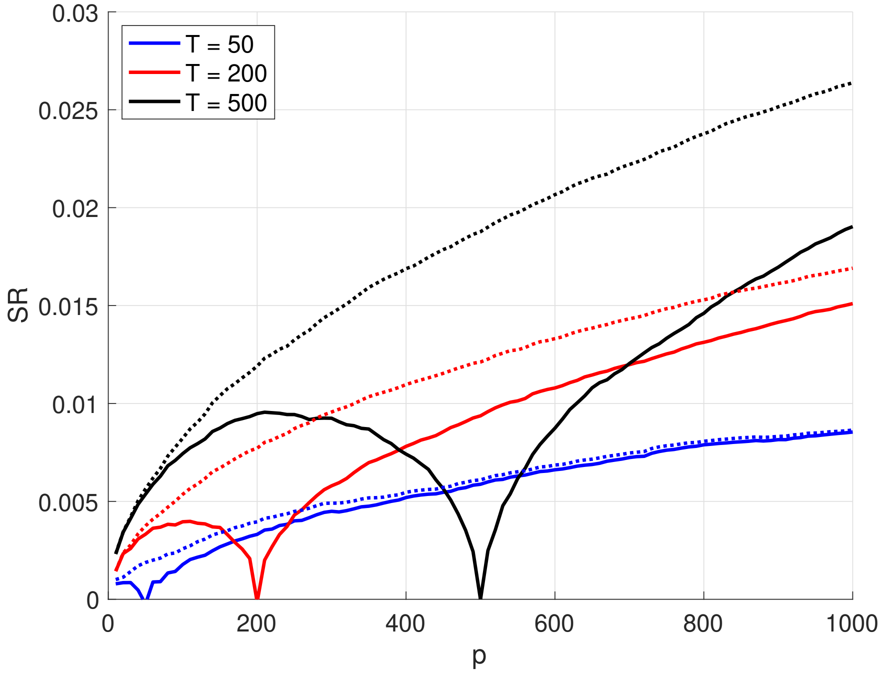


Figure X
Out-of-sample Sharpe ratios for different training sample sizes

Market-timing strategy based on ridgeless regression shown as solid lines, based on Bayesian posterior mean shown as dotted lines.

SR^2_{max}.

In the misspecified case—where the number of predictors used by the econometrician is smaller than the number of predictors in the true model of expected returns—there is an additional benefit to using a more complex model: incorporating more predictors reduces the omission of relevant predictive information. As KMZ note, this is a plausible characterization of the econometrician's problem in empirical asset pricing. Let P now denote the number of predictor variables in the true model, and $p \leq P$ is the number actually used by the econometrician. I simulate the Sharpe ratios achieved by the econometrician's market-timing strategy. Figure X displays the resulting Sharpe ratios for three value of T, holding fixed $P = 1,000$, $\sigma = 5$, implying $SR_{max} = 0.1925$, as before. In addition to the results for ridgeless regression, shown as solid lines, I also include the Sharpe ratio of a market-timing strategy that uses the Bayesian posterior mean as portfolio weight. This posterior mean is based on the prior belief that the elements of $\boldsymbol{b}$ are drawn from the same distribution that, in fact, generates them.

For all three values of T, higher p helps lift the Sharpe ratio, except in a region around $p = T$, due to the double-ascent property. However, as the figure shows, this dip in the Sharpe ratio can be avoided by employing optimal Bayesian shrinkage. For values of p close to P, the Sharpe ratio based on ridgeless regression and the Bayesian posterior mean are generally similar. That the Sharpe ratio is highest for high p shows that, qualitatively, there is a virtue of complexity. However, quantitatively, the benefits from higher complexity are minuscule when T is small. For $T = 50$ and $p = P$, the Sharpe ratio tops out at 0.0084 which is only around 1/25 of SR_{max}.

The key takeaway is that an empirical Sharpe ratio of approximately 0.30 with a training window of $T = 12$, as reported in KMZ cannot plausibly be attributed to the virtue of complexity in estimation. Even under conditions that fully exploit the benefits of model complexity, the Sharpe ratios achievable with such limited data are exceedingly small. As such, empirical findings of high Sharpe ratios in this setting cannot reflect predictive patterns learned from the training data—they must be driven by other factors. The insights from this analysis extend more broadly and generalize in several important directions:

Persistence. Predictor variables used in empirical asset pricing typically have some persistence. Persistence exacerbates the effects of small sample size because it reduces the in-sample variance of the predictor variable relative to the noise variance, which amplifies estimator error and forecast error. Put differently, with persistence, the estimator effectively only explores a very small part of the T-dimensional space spanned by the rows of $\boldsymbol{Z}$, and hence cannot learn about the regression slopes with much precision.

Cross-sectional asset pricing. In the cross-sectional setting of Section III, similar conclusions apply to some degree. The estimator of the P-dimensional vector of prices of risk in (20) represents the coefficients in a projection of 1 onto the training sample factor returns. We can decompose the vector of ones into "signal" and orthogonal noise as $\boldsymbol{\iota} = \boldsymbol{F}\boldsymbol{\lambda} + \boldsymbol{e}$. Then, inserting into (20) yields

$$\hat{\boldsymbol{\lambda}} = \boldsymbol{F}_t' \left(\boldsymbol{F}_t \boldsymbol{F}_t'\right)^{-1} \boldsymbol{F}\boldsymbol{\lambda} + \text{noise}. \tag{35}$$

The first term on the right-hand side represents the projection of the P-dimensional $\boldsymbol{\lambda}$ onto the T-dimensional random subspace spanned by the T rows of $\boldsymbol{F}_t$. As a result, the estimator can learn from the signal component only a weighted average of $\boldsymbol{\lambda}$, but not the full $\boldsymbol{\lambda}$. As in the time-series setting, when $T < P$, this limited view of the predictor space introduces out-of-sample forecast error beyond the estimation error caused by noise. Thus, the size of the training sample imposes a fundamental constraint on the dimensionality of factor risk pricing that the model can learn from the data. This limitation is not unique to DKKM; it also applies to other high-dimensional approaches that utilize large sets of factors or firm characteristics, e.g., as in Gu, Kelly, and Xiu (2020), Kozak, Nagel, and Santosh (2020), Chen, Pelger, and Zhu (2024).

There is, however, an important distinction from the time-series setting. In the cross-sectional asset-pricing context, the RFF-based factors in $\boldsymbol{F}_t$ are constructed as portfolio returns of the form $\boldsymbol{f}_{s+1} = \boldsymbol{Z}_s \boldsymbol{r}_{s+1}$. For a given T, a larger cross-sectional dimension N_s can help reduce the noise in these factor returns, thereby improving the estimator's ability to recover the true prices of risk. The extent to which a higher N_s reduces noise depends on the structure of the return covariance matrix. Specifically, it matters whether a given row of $\boldsymbol{Z}_s$ (which determines a factor's portfolio weights) is aligned with eigenvectors corresponding to large or small eigenvalues of the covariance matrix. If a factor portfolio loads primarily on low-variance directions (i.e., eigenvectors with small eigenvalues), the associated risk tends to diversify away, resulting in a cleaner signal. In contrast, if the factor loads on high-variance directions, diversification is less effective and noise remains substantial. When risk diversifies away but the corresponding price of risk remains large, an increase in N_s improves the signal-to-noise ratio, enhancing the estimator's ability to learn the true price of risk.

Return measurement frequency. Since return measurement frequency is a choice available to researchers, one might think that increasing frequency—and thereby raising the sample size T—could mitigate the limitations imposed by small training samples. Formally, this is correct: the first term on the right-hand side of (34) would represent a projection onto a higher-dimensional random subspace as measurement frequency increases. In a noiseless setting, this implies that the estimator could, in principle, learn about more directions in the predictor space. However, in practice, this benefit is minimal unless return-predictive signals also vary at high frequency. If predictors are instead persistent, increasing measurement frequency simply induces more smoothing. To see this, consider the ridgeless kernel regression representation (10). When a predictor vector $\boldsymbol{x}_s$ in a past period $s < t$ is similar to the current predictor vector $\boldsymbol{x}_{n,t}$ when measured at low frequency, then the predictor vectors measured at higher frequency around time s will also be similar. As a consequence, returns will be smoothed over all these adjacent higher-frequency periods, making it effectively a lower-frequency return. Ultimately, with persistent predictors, it is the total length of the training sample—rather than how finely the data is sampled within it—that determines how much of the predictor space is explored and, thus, how much signal can be extracted relative to the noise that obscures the true expected return function.

V. Conclusion

The empirical success of out-of-sample stock market index return prediction with a large number, P, of predictors constructed as Random Fourier Features (RFF)—randomly weighted and nonlinearly transformed versions of a small number, K, of original predictor variables—and small training sample size $T \ll P$ appears to suggest that a virtue of complexity allows the discovery of predictive relationships even when T is very small. However, the strong out-of-sample performance of the market-timing strategy in KMZ using training samples as small as $T = 12$ months is not evidence of virtuous complexity. Rather, it reflects fortunate coincidence: with small training sample sizes, the RFF approach in KMZ mechanically reduces approximately to a volatility-timed momentum strategy, which happened to perform well in the historical sample.

Crucially, the RFF-based regressions do not learn from the data that a volatility-timed momentum strategy is effective. Instead, the momentum-type weighting of past returns in the construction of predicted future returns arises mechanically. The predicted returns effectively take the form of a kernel-smoothed average of past returns, where the weights depend on the distance between past and current realizations of the K-dimensional original predictor vector. Because most of the K predictor variables underlying the RFF are persistent, recent observations tend to be more similar to the current predictor vector, leading to higher weights on recent returns—giving the strategy a momentum-like character. Moreover, when volatility is low, past and current predictor realizations are more tightly clustered, further concentrating the weights on recent months—imparting a volatility-timing component to the strategy.

Similar effects can arise in cross-sectional asset pricing applications where the SDF is expressed as a function of P factors, each constructed by weighting individual stock returns with RFF transformations of K original firm characteristics. When the training window is as short as 12 months, the predicted return for stock n effectively becomes a kernel-smoothed average of its own recent past returns, albeit possibly with some cross-stock smoothing as well if other stocks have recently exhibited firm characteristics similar to those of stock n.

These findings do not challenge the broader idea that complexity is virtuous—that is, a complex prediction model will often outperform an ad hoc, sparse, misspecified model that omits potentially relevant predictors. Rather, they challenge the notion that this virtue of complexity enables the discovery of substantial out-of-sample predictability in stock market index returns when models are trained on extremely small data sets. Conventional wisdom holds that identifying meaningful predictive relationships for stock returns requires training data spanning decades—not just a few years—if any such predictability exists at all. My empirical results are consistent with this conventional wisdom.

The fact that virtuous complexity cannot overcome the limitations of small training sample size is natural. When $T \ll P$, an overparametrized regression projects onto a very small T-dimensional random subspace within the P-dimensional space of predictors. Much of the predictive information embodied in predictors then necessarily remains hidden in unobserved dimensions, and the variance of the extractable signal is small relative to that of the noise. This also highlights a general limitation of what machine learning can achieve in empirical asset pricing: even when training samples are expanded to include all available data, sample sizes remain modest relative to the number of potentially useful informative variables and the complexity of predictive relationships.

References

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.

Berk, Jonathan, 2023, “Comment on ’The Virtue of Complexity in Return Prediction’,” Working paper, Stanford University.

Buncic, Daniel, 2025, “Simplified: A Closer Look at the Virtue of Complexity in Return Prediction,” Working paper, Stockholm University.

Cartea, Álvaro, Qi Jin, and Yuantao Shi, 2025, “The Limited Virtue of Complexity in a Noisy World,” Working paper, University of Oxford.

Chen, Luyang, Markus Pelger, and Jason Zhu, 2024, “Deep Learning in Asset Pricing,” *Management Science* 70, 714–750.

Da, Rui, Stefan Nagel, and Dacheng Xiu, 2024, “The Statistical Limit of Arbitrage,” Working paper, National Bureau of Economic Research.

Didisheim, Antoine, Shikun Barry Ke, Bryan T Kelly, and Semyon Malamud, 2023, “Complexity in Factor Pricing Models,” Working paper, National Bureau of Economic Research.

Didisheim, Antoine, Shikun Barry Ke, Bryan T Kelly, and Semyon Malamud, 2024, “APT or “AIPT”? The Surprising Dominance of Large Factor Models,” Working paper, National Bureau of Economic Research.

Efron, Bradley, 1986, “How Biased is the Apparent Error Rate of a Prediction Rule?,” *Journal of the American Statistical Association* 81, 461–470.

Fallahgoul, Hasan, 2025, “High-Dimensional Learning in Finance,” Working paper, arXiv.

Filipović, Damir, and Puneet Pasricha, 2022, “Empirical Asset Pricing via Ensemble Gaussian Process Regression,” Working paper, arXiv.

Filipović, Damir, Markus Pelger, and Ye Ye, 2022, “Stripping the Discount Curve-A Robust Machine Learning Approach”*Management Science*.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, “Empirical Asset Pricing via Machine Learning,” *Review of Financial Studies* 33, 2223–2273.

Hansen, Lars P., 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, T, A Montanari, S Rosset, and RJ Tibshirani, 2022, “Surprises in High-Dimensional Ridgeless Least Squares Interpolation,” *Annals of Statistics* 50, 949–986.

Hastie, Trevor, and Robert Tibshirani, 1987, “Generalized Additive Models: Some Applications,” *Journal of the American Statistical Association* 82, 371–386.

Jensen, Theis Ingerslev, Bryan T Kelly, Semyon Malamud, and Lasse Heje Pedersen, 2024, “Machine Learning and the Implementable Efficient Frontier,” Working paper 22-63, Swiss Finance Institute.

Kelly, Bryan, Semyon Malamud, and Kangying Zhou, 2024, “The Virtue of Complexity in Return Prediction,” *Journal of Finance* 79, 459–503.

Kelly, Bryan T, Semyon Malamud, and Kangying Zhou, 2022, “The Virtue of Complexity Everywhere,” Working paper, Swiss Finance Institute.

Kozak, Serhiy, 2023, “Kernel Trick for the Cross Section,” Working paper, University of Maryland.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2020, “Shrinking the Cross-Section,” *Journal of Financial Economics* 135, 271–292.

Martin, Ian, and Stefan Nagel, 2022, “Market Efficiency in the Age of Big Data,” *Journal of Financial Economics* 145, 154–177.

Molavi, Pooya, Alireza Tahbaz-Salehi, and Andrea Vedolin, 2024, “Model Complexity, Expectations, and Asset Prices,” *Review of Economic Studies* 91, 2462–2507.

Muirhead, Robb J. *Aspects of Multivariate Statistical Theory*, 1982).

Rahimi, Ali, and Benjamin Recht, 2007, “Random Features for Large-Scale Kernel Machines,” *Advances in neural information processing systems* 20.

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

Vershynin, Roman. *High-Dimensional Probability: An Introduction with Applications in Data Science* (2nd edition)., 2025).

Welch, Ivo, and Amit Goyal, 2008, “A Comprehensive Look at The Empirical Performance of Equity Premium Prediction,” *Review of Financial Studies* 21, 1455–1508.

Appendix

A. Spurious Predictability Induced by Standardizing Returns

The following analysis shows how standardizing can induce spurious predictability under the null of no true predictability.

Let excess returns on a stock market index return be $r_{t+1} = \mu + \sigma_t e_{t+1}$, where $\mu > 0$ and e_{t+1} is IID noise with $\mathbb{E}_t\, e_{t+1} = 0$ and $\mathbb{E}_t\, e^2_{t+1} = 1$. Assume that σ^2_t is time-varying, i.e., $\mathrm{var}(\sigma^2_t) > 0$. Assume that σ^2_t is observable to a researcher. The researcher examines market excess returns standardized by conditional volatility:

$$r_{t+1} = \frac{r_{t+1}}{\sigma_t} = \frac{\mu}{\sigma_t} + e_{t+1}. \tag{A.1}$$

The first term, $\frac{\mu}{\sigma_t}$, now has predictable variation and $\mathbb{E}_t\, r_{t+1} = \frac{\mu}{\sigma_t}$ is time-varying.

While a predictability test applied to r with predictor σ^{-1}_t would yield

$$\mathrm{cov}(r_{t+1}, \sigma^{-1}_t) = 0, \tag{A.2}$$

applied to y it yields evidence of predictability

$$\mathrm{cov}(r_{t+1}, \sigma^{-1}_t) = \mu\, \mathrm{var}(\sigma^{-1}_t) \tag{A.3}$$

and the regression slope coefficient in a regression of r_{t+1} on σ^{-1}_t is equal to μ and the intercept is zero. Therefore, the fitted prediction is $\mu\sigma^{-1}_t$, which is also equal to $\mathbb{E}_t\, r_{t+1}$. The predictable variation is

$$R^2 = \frac{\mu^2\, \mathrm{var}(\sigma^{-1}_t)}{\mathrm{var}(r_t)} \approx 0.43\% \tag{A.4}$$

which is not negligible in monthly data. This calculation uses $\mu = 0.0068$ (mean of the CRSP index return in the Goyal-Welch data set), $\mathrm{var}(\sigma^{-1}_t) = 108.80$ (using 12m lagged volatility of index returns as σ_t), and $\mathrm{var}(y_t) = 1.19$.

Now construct a timing strategy based on the predicted value

$$f_{t+1} = r_{t+1} x_t, \qquad \text{with} \quad x_t = \mu\sigma^{-1}_t \tag{A.5}$$

(The return of this timing strategy will be the same as the return of a volatility-timing strategy with weight $\mu\sigma^{-2}_t$ applied to the non-standardized return r_{t+1}, which exploits that μ does not vary with σ_t).

To get alpha, let's first get the covariance with the market excess return:

$$\begin{aligned} \mathrm{cov}\left(f_{t+1}, r_{t+1}\right) &= \mathrm{cov}\left(\frac{r_{t+1}}{\sigma_t}\left(\frac{\mu}{\sigma_t}\right), r_{t+1}\right) \\ &= \mathbb{E}\left[\frac{\mu\sigma^2_t e^2_{t+1}}{\sigma^2_t}\right] \\ &= \mu \end{aligned} \tag{A.6}$$

so

$$\begin{aligned}\beta &= \frac{\mu}{\mathrm{var}(r_{t+1})}\\ &= \frac{\mu}{\mathbb{E}[\sigma_t^2]}.\end{aligned} \tag{A.7}$$

Therefore,

$$\begin{aligned}\alpha &= \mathbb{E}\left[f_{t+1}\right] - \beta\mu\\ &= \mu^2\,\mathbb{E}[\sigma_t^{-2}] - \mu^2\,\mathbb{E}[\sigma_t^2]^{-1}\\ &= \mu^2\left(\mathbb{E}[\sigma_t^{-2}] - \mathbb{E}[\sigma_t^2]^{-1}\right)\\ &> 0,\end{aligned} \tag{A.8}$$

where the bound in the last line follows from Jensen's inequality. With the same moments as I used for the R^2 above, and after annualizing,

$$\alpha \approx 0.24. \tag{A.9}$$

Using the standardized market return as benchmark, as in KMZ, the alpha is smaller: Covariance with the standardized market return

$$\mathrm{cov}\left(f_{t+1}, r_{t+1}\right) = \mu\,\mathbb{E}[\sigma_t^{-1}] \tag{A.10}$$

and, since $\mathrm{var}(r_{t+1}) \approx 1$,

$$\beta = \mu\,\mathbb{E}[\sigma_t^{-1}] \tag{A.11}$$

$$\begin{aligned}\alpha &= \mathbb{E}\left[f_{t+1}\right] - \beta\mu\,\mathbb{E}[\sigma_t^{-1}]\\ &= \mu^2(\mathbb{E}[\sigma_t^{-2}] - \mathbb{E}[\sigma_t^{-1}]^2)\\ &= \mu^2\,\mathrm{var}(\sigma_t^{-1}),\end{aligned} \tag{A.12}$$

which comes out to be annualized about $\alpha \approx 0.06$ with empirical moments for μ and $\mathrm{var}(\sigma_t^{-1})$. In simulations, I find the same value of alpha on average, and an information ratio of about 0.25, which is not a negligible magnitude!

In the simulations with window size $T = 12$, however, I find that the fitted values from predictive regressions with RFF don't get close to capturing all this predictability that the reciprocal volatility strategy above captures. The information ratio I find in these simulations for the RFF-based strategy is only around 0.05. So much of the predictability captured by the RFF is due to something else, not the above mechanical effect related to standardization. That said, as a general matter, that the above analysis shows that standardizing the dependent variable is a somewhat dangerous practice in a study that has the objective of documenting predictability. It's only a small part of the story in KMZ, but it could play a bigger role in others.

B. Proofs

Lemma 1. *Let $\boldsymbol{x} \sim (N)(0, \boldsymbol{I})$ with dimension P and $\boldsymbol{A}$ a deterministic symmetric matrix. Then*

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}] = \operatorname{tr}(\boldsymbol{A})\boldsymbol{I}_p + 2\boldsymbol{A} \tag{A.13}$$

Proof. For the (i, j)-entry,

$$\mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}_j] = \sum_{k,\ell} A_{k\ell}\, \mathbb{E}[x_i\, x_j\, x_k\, x_\ell] \tag{A.14}$$

Because $\boldsymbol{x} \sim \mathrm{N}(0, \boldsymbol{I})$, the fourth moment factorizes via Isserlis' theorem

$$\mathbb{E}[x_i\, x_j\, x_k\, x_\ell] = \delta_{ij}\, \delta_{k\ell} + \delta_{ik}\, \delta_{j\ell} + \delta_{i\ell}\, \delta_{jk}, \tag{A.15}$$

where $\delta_{..}$ is the Kronecker delta. Inserting into the sum yields

$$\begin{aligned}\mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}_j] &= \delta_{ij}\sum_{k,\ell}\delta_{k\ell}A_{k\ell} + \sum_{k,\ell}\delta_{ik}\,\delta_{j\ell}A_{k\ell} + \sum_{k,\ell}\delta_{i\ell}\,\delta_{jk}A_{k\ell} \\ &= \delta_{ij}\operatorname{tr}(\boldsymbol{A}) \ + \ A_{ij} \ + \ A_{ji}. \end{aligned} \tag{A.16}$$

Stacking the entries and using the symmetry of $\boldsymbol{A}$ then leads to the result. □

Lemma 2. *Let $\boldsymbol{X}$ be a $T \times P$ random matrix with IID standard normal elements and $T < P$ and let $\boldsymbol{b}$ be a deterministic $P \times 1$ vector. Then*

$$\mathbb{E}[\boldsymbol{b}'\boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}'\boldsymbol{b}] = \frac{T}{P}\boldsymbol{b}'\boldsymbol{b} \tag{A.17}$$

and

$$\operatorname{var}(\boldsymbol{b}'\boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}'\boldsymbol{b}) = (\boldsymbol{b}'\boldsymbol{b})^2\frac{2T(P-T)}{P^2(P+2)}. \tag{A.18}$$

Proof. Define $\boldsymbol{P} = \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}$. Since $\boldsymbol{X}$ has IID standard normal entries, $\boldsymbol{P}$ is a random orthogonal projector of rank T that represents the orthogonal projection in $\mathbb{R}^P$ onto a random T-dimensional subspace that is distributed uniformly on the Grassmannian of all such subspaces (see Vershynin (2025) Section 5.2.6). Lemma 5.3.2. in Vershynin (2025) then provides the stated result for the expected value. Next consider the variance. Instead of considering fixed $\boldsymbol{b}$ and random $\boldsymbol{P}$, the same random subspace can be obtained by considering fixed $\boldsymbol{P}$ and random $\frac{1}{\sqrt{\boldsymbol{b}'\boldsymbol{b}}}\boldsymbol{b}$, with $\boldsymbol{b} \sim \mathrm{N}(0, \boldsymbol{I})$ (see Vershynin (2025), proof of Lemma 5.3.2). Proceeding with this latter view, Lemma 1.5.7. in Muirhead (1982) then implies

$$\frac{\boldsymbol{b}'\boldsymbol{P}\boldsymbol{b}}{\boldsymbol{b}'\boldsymbol{b}} \sim \mathrm{Beta}\Big(\frac{T}{2}, \frac{P-T}{2}\Big). \tag{A.19}$$

Switching back to the view of $\boldsymbol{P}$ as random and $\boldsymbol{b}$ as fixed, we still have the same distribution. Evaluating the variance of the beta distribution, then gives the result on variance. □

Lemma 3. *Let $\boldsymbol{X}$ be a $T \times P$ random matrix with IID standard normal elements and $T < P - 1$. Then*

$$\mathbb{E}[(\boldsymbol{X}\boldsymbol{X}')^{-1}] = \boldsymbol{I}_T\frac{1}{P-T-1}. \tag{A.20}$$

Proof. This result follows from the fact $\boldsymbol{X}\boldsymbol{X}'$ follows a Wishart distribution and application of the properties of the inverse Wishart distribution. □

Lemma 4. *Let $\boldsymbol{X}$ be a $T \times P$ random matrix with IID standard normal elements and $T < P - 1$ and let $\boldsymbol{b}$ be a deterministic $P \times 1$ vector. Then*

$$\mathbb{E}[\boldsymbol{b}'\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-2}\boldsymbol{X}\boldsymbol{b}] = \frac{T}{P(P-T-1)}\boldsymbol{b}'\boldsymbol{b}. \tag{A.21}$$

Proof. Let $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}'$ and $\boldsymbol{A} = \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-2}\boldsymbol{X}$. Using rotational invariance, i.e., $\boldsymbol{X}\boldsymbol{V} \overset{d}{=} \boldsymbol{X}$, replacing $\boldsymbol{X}$ with $\boldsymbol{X}\boldsymbol{V}$ in $\boldsymbol{A}$ then shows

$$\boldsymbol{A} \overset{d}{=} \boldsymbol{V}'\boldsymbol{A}\boldsymbol{V} \tag{A.22}$$

for all orthogonal matrices $\boldsymbol{V}$. Hence, we have $\mathbb{E}[\boldsymbol{A}] = \boldsymbol{V}'\,\mathbb{E}[\boldsymbol{A}]\boldsymbol{V}$ for every orthogonal matrix $\boldsymbol{V}$, which means that $\mathbb{E}[\boldsymbol{A}]$ commutes with every orthogonal matrix, which, by Schur's lemma, implies that

$$\mathbb{E}[\boldsymbol{A}] = c\boldsymbol{I} \tag{A.23}$$

for some scalar constant c. We can find c by noting that

$$\mathbb{E}[\mathrm{tr}(\boldsymbol{A})] = \mathbb{E}[\mathrm{tr}(\boldsymbol{S}^{-1})] = \frac{T}{P-T-1}. \tag{A.24}$$

where the last equality follows from Lemma 3. Hence $c = \frac{T}{P(P-T-1)}$ and the stated result follows. □

B.1. Maximum achievable Sharpe ratio

Expected market-timing return

$$\begin{aligned}
\mathbb{E}[\boldsymbol{z}_t'\boldsymbol{b}\boldsymbol{b}'\boldsymbol{z}_t|\boldsymbol{b}] &= \mathbb{E}[\mathrm{tr}(\boldsymbol{z}_t'\boldsymbol{b}\boldsymbol{b}'\boldsymbol{z}_t)|\boldsymbol{b}] \\
&= \mathbb{E}[\mathrm{tr}(\boldsymbol{z}_t\boldsymbol{z}_t'\boldsymbol{b}\boldsymbol{b}')|\boldsymbol{b}] \\
&= \mathrm{tr}(\boldsymbol{b}\boldsymbol{b}') \\
&= \boldsymbol{b}'\boldsymbol{b}
\end{aligned} \tag{A.25}$$

The expected squared market timing return is

$$\begin{aligned}
\mathbb{E}[\boldsymbol{z}_t'\boldsymbol{b}\boldsymbol{b}'\boldsymbol{z}_t\boldsymbol{z}_t'\boldsymbol{b}\boldsymbol{b}'\boldsymbol{z}_t|\boldsymbol{b}] &= \mathbb{E}[\mathrm{tr}(\boldsymbol{z}_t\boldsymbol{z}_t'\boldsymbol{b}\boldsymbol{b}'\boldsymbol{z}_t\boldsymbol{z}_t'\boldsymbol{b}\boldsymbol{b}')|\boldsymbol{b}] + \mathbb{E}[e_{t+1}^2\boldsymbol{z}_t'\boldsymbol{b}\boldsymbol{b}'\boldsymbol{z}_t|\boldsymbol{b}] \\
&= \mathrm{tr}\left\{\left[\mathrm{tr}(\boldsymbol{b}\boldsymbol{b}')\boldsymbol{I}_p + 2\boldsymbol{b}\boldsymbol{b}'\right]\boldsymbol{b}\boldsymbol{b}'\right\} + \sigma^2\,\mathbb{E}[\mathrm{tr}(\boldsymbol{z}_t'\boldsymbol{b}\boldsymbol{b}'\boldsymbol{z}_t)|\boldsymbol{b}] \\
&= \mathrm{tr}\left\{\mathrm{tr}(\boldsymbol{b}\boldsymbol{b}')\boldsymbol{b}\boldsymbol{b}' + 2\boldsymbol{b}\boldsymbol{b}'\boldsymbol{b}\boldsymbol{b}'\right\} + \sigma^2\boldsymbol{b}'\boldsymbol{b} \\
&= 3(\boldsymbol{b}'\boldsymbol{b})^2 + \sigma^2\boldsymbol{b}'\boldsymbol{b}
\end{aligned} \tag{A.26}$$

where the second equality follows from Lemma 1. Therefore,

$$\mathrm{var}(\boldsymbol{z}_t'\boldsymbol{b}\boldsymbol{b}'\boldsymbol{z}_t|\boldsymbol{b}) = 3(\boldsymbol{b}'\boldsymbol{b})^2 + \sigma^2\boldsymbol{b}'\boldsymbol{b} - (\boldsymbol{b}'\boldsymbol{b})^2 = \sigma^2\boldsymbol{b}'\boldsymbol{b} + 2(\boldsymbol{b}'\boldsymbol{b})^2 \tag{A.27}$$

and hence the squared Sharpe ratio is

$$SR_{max}^2 = \frac{(\boldsymbol{b}'\boldsymbol{b})^2}{\sigma^2\boldsymbol{b}'\boldsymbol{b} + 2(\boldsymbol{b}'\boldsymbol{b})^2} = \frac{\boldsymbol{b}'\boldsymbol{b}}{\sigma^2 + 2(\boldsymbol{b}'\boldsymbol{b})} \approx \frac{1}{\sigma^2+2} \tag{A.28}$$

where the approximation is accurate for large P.

B.2. Sharpe ratio based on estimated parameters with ridgeless regression in the $T < p$ case

Expected returns

$$\begin{aligned}
\mathbb{E}[r_{t+1}\hat{r}_t|\boldsymbol{Z},\boldsymbol{b}] &= \mathbb{E}[\boldsymbol{b}'\boldsymbol{z}_t\boldsymbol{z}_t'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{b}|\boldsymbol{Z},\boldsymbol{b}] \\
&= \boldsymbol{b}'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{b} \qquad \text{(A.29)}
\end{aligned}$$

Conditioning down and using Lemma 2, I obtain

$$\mathbb{E}[r_{t+1}\hat{r}_t|\boldsymbol{b}] = \frac{T}{P}\boldsymbol{b}'\boldsymbol{b} \approx \frac{T}{P}. \qquad \text{(A.30)}$$

To calculate the variance, let's first focus on the uncentered second moment,

$$\begin{aligned}
\mathbb{E}[r_{t+1}^2\hat{r}_t^2|\boldsymbol{Z},\boldsymbol{b}] &= \mathbb{E}[\left\{(e_{t+1}+\boldsymbol{z}_t'\boldsymbol{b})\boldsymbol{z}_t'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}(\boldsymbol{Z}\boldsymbol{b}+\boldsymbol{e})\right\}^2|\boldsymbol{Z},\boldsymbol{b}] \\
&= \mathbb{E}[(e_{t+1}\boldsymbol{z}_t'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{b})^2|\boldsymbol{Z},\boldsymbol{b}] \\
&\quad + \mathbb{E}[(e_{t+1}\boldsymbol{z}_t'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{e})^2|\boldsymbol{Z},\boldsymbol{b}] \\
&\quad + \mathbb{E}[(\boldsymbol{z}_t'\boldsymbol{b}\boldsymbol{z}_t'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{b})^2|\boldsymbol{Z},\boldsymbol{b}] \\
&\quad + \mathbb{E}[(\boldsymbol{z}_t'\boldsymbol{b}\boldsymbol{z}_t'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{e})^2|\boldsymbol{Z},\boldsymbol{b}]. \qquad \text{(A.31)}
\end{aligned}$$

The omitted cross-product terms have zero expected value due to the mutual independence of e_{t+1}, $\boldsymbol{e}$, and $\boldsymbol{z}_t$. Evaluating the conditional expectations, the cyclical property of the trace, and then in the second step Lemma 1 for the third and the last term, I find that the conditional variance is

$$\begin{aligned}
\mathrm{var}(r_{t+1}\hat{r}_t|\boldsymbol{Z},\boldsymbol{b}) &= \mathbb{E}[r_{t+1}^2\hat{r}_t^2|\boldsymbol{Z},\boldsymbol{b}] - (\mathbb{E}[r_{t+1}\hat{r}_t|\boldsymbol{Z},\boldsymbol{b}])^2 \\
&= \mathbb{E}[\mathrm{tr}\left(\boldsymbol{z}_t e_{t+1}^2\boldsymbol{z}_t'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{b}\boldsymbol{b}'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\right)|\boldsymbol{Z},\boldsymbol{b}] \\
&\quad + \mathbb{E}[\mathrm{tr}\left(\boldsymbol{z}_t e_{t+1}^2\boldsymbol{z}_t'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{e}\boldsymbol{e}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\right)|\boldsymbol{Z},\boldsymbol{b}] \\
&\quad + \mathbb{E}[\mathrm{tr}\left\{(\boldsymbol{z}_t\boldsymbol{z}_t'\boldsymbol{b}\boldsymbol{b}'\boldsymbol{z}_t\boldsymbol{z}_t'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{b}\boldsymbol{b}'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\right\}|\boldsymbol{Z},\boldsymbol{b}] \\
&\quad + \mathbb{E}[\mathrm{tr}\left(\boldsymbol{b}'\boldsymbol{z}_t\boldsymbol{z}_t'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{e}\boldsymbol{e}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{z}_t\boldsymbol{z}_t'\boldsymbol{b}\right)|\boldsymbol{Z},\boldsymbol{b}] \\
&\quad - (\boldsymbol{b}'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{b})^2 \\
&= \sigma^2\,\mathrm{tr}\left(\boldsymbol{b}\boldsymbol{b}'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\right) \\
&\quad + \sigma^4\,\mathrm{tr}(\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-2}\boldsymbol{Z}) \\
&\quad + \boldsymbol{b}'\boldsymbol{b}\,\mathrm{tr}(\boldsymbol{b}\boldsymbol{b}'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}) + 2\,\mathrm{tr}(\boldsymbol{b}\boldsymbol{b}'\boldsymbol{b}\boldsymbol{b}'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}) \\
&\quad + \sigma^2\,\mathrm{tr}\left(\boldsymbol{b}\boldsymbol{b}'\left\{\mathrm{tr}[\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-2}\boldsymbol{Z}]\boldsymbol{I}_p + 2\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-2}\boldsymbol{Z}\right\}\right) \\
&\quad - (\boldsymbol{b}'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{b})^2 \\
&= \sigma^2\,\mathrm{tr}\left(\boldsymbol{b}\boldsymbol{b}'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\right) \\
&\quad + \sigma^4\,\mathrm{tr}(\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-2}\boldsymbol{Z}) \\
&\quad + 3\boldsymbol{b}'\boldsymbol{b}\,\mathrm{tr}(\boldsymbol{b}\boldsymbol{b}'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}) \\
&\quad + \sigma^2\,\mathrm{tr}\left(\boldsymbol{b}\boldsymbol{b}'\left\{\mathrm{tr}[\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-2}\boldsymbol{Z}]\boldsymbol{I}_p + 2\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-2}\boldsymbol{Z}\right\}\right) \\
&\quad - (\boldsymbol{b}'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{b})^2 \qquad \text{(A.32)}
\end{aligned}$$

By Lemma 2 for the first and third term, Lemma 3 for the second, as well as Lemma 3 and then Lemma 4 for the fourth, and Lemma 2, adding squared mean and variance, for the fifth

$$\begin{aligned}
\mathbb{E}[\operatorname{var}(r_{t+1}\hat{r}_t|\boldsymbol{Z},\boldsymbol{b})|\boldsymbol{b}] &= \sigma^2\frac{T}{P}\boldsymbol{b}'\boldsymbol{b} + \sigma^4\frac{T}{P-T-1} + 3\frac{T}{P}(\boldsymbol{b}'\boldsymbol{b})^2 \\
&\quad + \sigma^2\left\{\frac{T}{P-T-1}\operatorname{tr}\left(\boldsymbol{b}\boldsymbol{b}'\right) + 2\,\mathbb{E}[\operatorname{tr}\left(\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-2}\boldsymbol{Z}\boldsymbol{b}\boldsymbol{b}'\right)]\right\} \\
&\quad - \frac{T^2}{P^2}(\boldsymbol{b}'\boldsymbol{b})^2 - (\boldsymbol{b}'\boldsymbol{b})^2\frac{2T(P-T)}{P^2(P+2)} \\
&= \sigma^2\frac{T}{P}\boldsymbol{b}'\boldsymbol{b} + \sigma^4\frac{T}{P-T-1} + \frac{T}{P}(\boldsymbol{b}'\boldsymbol{b})^2\left(3 - \frac{T-2}{P+2}\right) \\
&\quad + \sigma^2\boldsymbol{b}'\boldsymbol{b}\left[\frac{T}{P-T-1} + 2\frac{T}{P(P-T-1)}\right] \\
&= \sigma^2\frac{T}{P} + \sigma^4\frac{T}{P-T-1} + \frac{T}{P}(\boldsymbol{b}'\boldsymbol{b})^2\left(3 - \frac{T-2}{P+2}\right) + \sigma^2\boldsymbol{b}'\boldsymbol{b}\left[\frac{T(1+2/P)}{P-T-1}\right] \\
&\approx \sigma^2\left[\frac{T}{P} + \frac{T(1+2/P)}{P-T-1}\right] + \sigma^4\frac{T}{P-T-1} + \frac{T}{P}\left(3 - \frac{T-2}{P+2}\right) \qquad \text{(A.33)}
\end{aligned}$$

Using Lemma 2,

$$\begin{aligned}
\operatorname{var}\left(\mathbb{E}[r_{t+1}\hat{r}_t|\boldsymbol{Z},\boldsymbol{b}]|\boldsymbol{b}\right) &= \operatorname{var}\left(\boldsymbol{b}'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{b}|\boldsymbol{b}\right) \\
&= (\boldsymbol{b}'\boldsymbol{b})^2\frac{2T(P-T)}{P^2(P+2)} \\
&\approx \frac{2T(P-T)}{P^2(P+2)} \qquad \text{(A.34)}
\end{aligned}$$

Hence,

$$\begin{aligned}
\operatorname{var}\left(\mathbb{E}[r_{t+1}\hat{r}_t|\boldsymbol{b}]\right) &= \mathbb{E}[\operatorname{var}(r_{t+1}\hat{r}_t|\boldsymbol{Z},\boldsymbol{b})|\boldsymbol{b}] + \operatorname{var}\left(\mathbb{E}[r_{t+1}\hat{r}_t|\boldsymbol{Z},\boldsymbol{b}]|\boldsymbol{b}\right) \\
&= \sigma^2\left[\frac{T}{P} + \frac{T(1+2/P)}{P-T-1}\right] + \sigma^4\frac{T}{P-T-1} + \frac{T}{P}\left(3 - \frac{T+2}{P+2}\right) + \frac{2T(P-T)}{P^2(P+2)} \\
&= \sigma^2\left[\frac{T}{P} + \frac{T(1+2/P)}{P-T-1}\right] + \sigma^4\frac{T}{P-T-1} + \frac{T(3P-T)}{P^2} \qquad \text{(A.35)}
\end{aligned}$$

Combing the results for expected return and variance, we get the squared Sharpe ratio conditional on $\boldsymbol{b}$

$$\begin{aligned}
SR^2 &\approx \frac{\frac{T^2}{P^2}}{\sigma^2\left[\frac{T}{P} + \frac{T(1+2/P)}{P-T-1}\right] + \sigma^4\frac{T}{P-T-1} + \frac{T(3P-T)}{P^2}} \\
&= \left(\frac{T}{P}\right)\frac{1}{\sigma^2\left[1 + \frac{P+2}{P-T-1}\right] + \sigma^4\frac{P}{P-T-1} + \frac{3P-T}{P}} \\
&= \left(\frac{T}{P}\right)\frac{1}{\sigma^2\left[1 + \frac{P+2+\sigma^2 P}{P-T-1}\right] + 3 - \frac{T}{P}} \qquad \text{(A.36)}
\end{aligned}$$