

The Virtue Of Complexity in Return Prediction

Oualid Missaoui

Agenda

- Introduction
- Correctly Specified world
- Mis-specified world
- Empirical Results
- References

Overview

Once you admit that the true pricing signal is complex, “Occam’s razor” is misleading. Rich, high-dimensional models with shrinkage are not a vice; they’re a *virtue*—they systematically deliver better trading performance than simple, elegant regressions.

The Virtue of Complexity in Return Prediction

Swiss Finance Institute Research Paper No. 21-90

Journal of Finance, forthcoming

141 Pages • Posted: 15 Dec 2021 • Last revised: 20 Oct 2023

[Bryan T. Kelly](#)

Yale SOM; AQR Capital Management, LLC; National Bureau of Economic Research (NBER)

[Semyon Malamud](#)

Ecole Polytechnique Federale de Lausanne; Centre for Economic Policy Research (CEPR); Swiss Finance Institute

[Kangying Zhou](#)

Yale School of Management

 [There are 3 versions of this paper](#)

Date Written: December 13, 2021

Abstract

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

Keywords: Portfolio choice, machine learning, random matrix theory, benign overfit, overparameterization

JEL Classification: C3, C58, C61, G11, G12, G14

Suggested Citation:

Kelly, Bryan T. and Malamud, Semyon and Zhou, Kangying, The Virtue of Complexity in Return Prediction (December 13, 2021). Swiss Finance Institute Research Paper No. 21-90, Journal of Finance, forthcoming, Available at SSRN: <https://ssrn.com/abstract=3984925> or <http://dx.doi.org/10.2139/ssrn.3984925>

[Show Contact Information](#) >

[Source](#)

Overview

Understanding The Virtue of Complexity

Swiss Finance Institute Research Paper No. 25-96

101 Pages • Posted: 10 Jul 2025 • Last revised: 11 Nov 2025

[Bryan T. Kelly](#)

Yale SOM; AQR Capital Management, LLC; National Bureau of Economic Research (NBER)

[Semyon Malamud](#)

Ecole Polytechnique Federale de Lausanne; Centre for Economic Policy Research (CEPR); Swiss Finance Institute

Date Written: July 01, 2025

Abstract

Recent papers have challenged certain aspects of the "virtue of complexity" described by Kelly et al. (2024b) (KMZ) and related work. These challenges ultimately have little bearing on the theoretical arguments or empirical findings of KMZ. They do, however, provide a valuable opportunity to better understand the nuanced behavior of complex models. In addition to responding to recent challenges, we provide detailed discussions of how complex models learn in small samples, the roles of "nominal" and "effective" complexity, the unique effects of implicit regularization, and the importance of limits to learning. We then present new empirical and theoretical analyses that expand on KMZ. Finally, we introduce and demonstrate the virtue of ensemble complexity.

Keywords: Portfolio choice, machine learning, random matrix theory, benign overfit

JEL Classification: C58, C61, G11, G12, G14

Suggested Citation:

Kelly, Bryan T. and Malamud, Semyon, Understanding The Virtue of Complexity (July 01, 2025).
Swiss Finance Institute Research Paper No. 25-96, Available at SSRN:
<https://ssrn.com/abstract=5346842> or <http://dx.doi.org/10.2139/ssrn.5346842>

- Nominal Complexity
- Effective Complexity

[Show Contact Information](#) >

What is this paper about?

It asks a simple question: in return prediction and market timing, is model complexity a bug or a feature?

Step 1 – A clean, correctly specified world.

Assume the true conditional mean of returns is exactly linear in a high-dimensional signal vector. In this world, as the number of predictors grows relative to sample size, OLS and ridgeless regression show classic double-descent: out-of-sample R^2 can plunge to $-\infty$ at the interpolation boundary, coefficients blow up, and even with infinite data you never quite reach the oracle benchmark. Here, complexity is pure statistical cost.

Step 2 – The realistic, misspecified world.

In practice, the true DGP is unknown and highly nonlinear; any small linear model is misspecified. The paper lets the empirical model use an expanding set of signals and, using random matrix theory, shows that with properly tuned shrinkage the reduction in misspecification error dominates the extra estimation noise. Both out-of-sample R^2 and the market-timing Sharpe ratio become *increasing and concave* functions of model complexity.

Notation

- **Time & Assets**

- t : time index
- Single risky asset excess return: R_{t+1}
- Risk-free asset: return normalized to 0 (all returns are excess)

- **Predictors / Signals**

- $S_t \in \mathbb{R}^P$: vector of predictive signals at time t
- P : number of predictors (model dimension)
- T : sample size (number of time periods)
- **Complexity ratio**: $c = P/T$

Model Setup: Correct vs Misspecified (True DGP + Empirical Models)

Correctly Specified (True) Model

- We assume there is a single risky asset whose return is fully explained by the signals we observe.
- The signals we feed into the regression are exactly the ones that generate the true expected return.
- In this world, if we had infinite data, our regression model could recover the true relationship between signals and returns.
- Any gap between what we learn and the truth comes purely from **finite-sample estimation noise**, not from a wrong model.

Misspecified (Empirical) Model

- In reality, the asset's return depends on a richer set of signals than the ones we actually use.
- The full "true" signal vector can be split into:
 - signals we observe and include in the regression, and
 - signals we do not observe or simply ignore.
- Our empirical model is therefore built on an **incomplete view** of the true information set.
- Even with infinite data, this model cannot fully match the true relationship, because some relevant signals are missing — this is **specification (approximation) error**, in addition to ordinary estimation noise.

Correctly Specified Model

- **True data-generating process (DGP)**

We observe a single risky asset with excess return

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

where

- $S_t \in \mathbb{R}^P$: vector of signals (predictors),
- $\beta \in \mathbb{R}^P$: true slope vector,
- ε_{t+1} : noise with $\mathbb{E}[\varepsilon_{t+1} \mid S_t] = 0$.

The key assumption: **the empirical model sees exactly these signals** S_t , so the conditional mean $\mathbb{E}[R_{t+1} \mid S_t] = S_t' \beta$ lies inside the model class.

- **Empirical model in this world**

$$R_{t+1} \approx S_t' \hat{\beta}$$

where $\hat{\beta} \in \mathbb{R}^P$ is estimated from $\{(S_t, R_{t+1})\}_{t=1}^T$ by (possibly regularized) linear regression:

$$\hat{\beta}(z) = \arg \min_b \frac{1}{T} \sum_{t=1}^T (R_{t+1} - S_t' b)^2 + z \|b\|^2,$$

where $z \geq 0$ is the ridge penalty (with $z \rightarrow 0^+$ giving the ridgeless solution).

Misspecified Model (Realistic World)

- **True data-generating process (full signal space)**

The true DGP is still linear in a larger signal vector:

$$R_{t+1} = S_t' \beta + \varepsilon_{t+1}, \quad S_t = (S_t^{(1)}, S_t^{(2)}),$$

where

- $S_t^{(1)} \in \mathbb{R}^{P_1}$: "observed / used" signals,
- $S_t^{(2)} \in \mathbb{R}^{P_2}$: "unobserved / ignored" signals,
- $P = P_1 + P_2$, and β splits accordingly as $\beta = (\beta^{(1)}, \beta^{(2)})$.

The conditional mean depends on **all** signals:

$$\mathbb{E}[R_{t+1} \mid S_t] = S_t' \beta = S_t^{(1)'} \beta^{(1)} + S_t^{(2)'} \beta^{(2)}.$$

- **Empirical model in this world (indexed by P_1 or q)**

The analyst only uses the first block $S_t^{(1)}$ and ignores $S_t^{(2)}$:

$$R_{t+1} \approx S_t^{(1)'} \hat{\beta}^{(1)}(q), \quad q = \frac{P_1}{P}$$

For each model size P_1 (or fraction $q = P_1/P$), the empirical regression is

$$\hat{\beta}^{(1)}(z; q) = \arg \min_b \frac{1}{T} \sum_{t=1}^T (R_{t+1} - S_t^{(1)'} b)^2 + z \|b\|^2,$$

Here, the model is **misspecified whenever** $P_1 < P$: it only sees part of the true signal vector, so even with infinite data the best possible predictor in $S_t^{(1)}$ generally cannot reproduce the full conditional mean that depends on both $S_t^{(1)}$ and $S_t^{(2)}$.

Timing Strategy: Definition, Equation, and Intuition

Definition

The paper studies a *market-timing* strategy on a **single risky asset** (e.g., the market index) using predictive signals. Each period, the strategy chooses how much to invest in the risky asset based on the signals observed at that time.

1. Predictive regression (forecasting next return)

- Use signals $S_t \in \mathbb{R}^P$ to forecast the next excess return:

$$\hat{\mu}_t = S_t' \hat{\beta}$$

where $\hat{\beta}$ is estimated from past data.

2. Timing rule (portfolio weight / position)

- The *timing strategy* sets the position in the risky asset equal to the forecasted excess return:

$$\pi_t = S_t' \hat{\beta} = \hat{\mu}_t$$

This π_t is the portfolio weight or leverage in the risky asset; the remainder is implicitly held in the risk-free asset.

3. Strategy return

- The excess return of the timing strategy over the next period is:

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

where R_{t+1} is the excess return on the single risky asset.

Why call it a “timing” strategy?

- A **static** investor would hold a fixed position (e.g., always 60% in the market).
- This strategy **changes its exposure over time**:
 - If signals predict high expected return ($\hat{\mu}_t$ large and positive), it takes a **large long** position ($\pi_t \gg 0$).
 - If signals predict low or negative expected return, it takes a **small, flat, or even short** position ($\pi_t \approx 0$ or $\pi_t < 0$).
- Because it is **dynamically increasing and decreasing exposure** to the same asset based on forecasts, it is interpreted as a **market-timing strategy** rather than a stock-selection or static allocation strategy.

Experimental Design: Two Worlds

1. Correctly Specified “Toy” World (Varying Problems)

- **Key idea:**

“Let’s look at many different high-dimensional worlds (different P), but keep the amount of true predictability fixed across them. Then we ask: how well can we learn that same amount of predictability when $P/T = c$ changes?”

- For each complexity level $c = P/T$, we imagine a **different high-dimensional problem**:

- Signals $S_t \in \mathbb{R}^P$,
- True coefficients $\beta \in \mathbb{R}^P$,
- Sample size T with $P/T \rightarrow c$.

- Moving along the c -axis means moving across **different economies in the same family**:

- Each (P, T) pair defines its own correctly specified regression problem.
- In every such problem, the true conditional mean of returns is exactly linear in the observed signals.

2. Misspecified “Realistic” World (Fixed Problem, Varying Models)

- Here we fix **one underlying high-dimensional economy**:
 - A full signal vector $S_t \in \mathbb{R}^P$ and true $\beta \in \mathbb{R}^P$,
 - With a fixed complexity ratio $c = P/T$.
- What varies is the **empirical model**, not the economy:
 - We only use a subset of signals $S_t^{(1)} \in \mathbb{R}^{P_1}$ with $P_1 = qP$, $q \in [0, 1]$.
 - As q increases, we include more of the true signal space: a **nested sequence of misspecified models** on the same DGP.
- In this design, the x -axis is q (or P_1), not c : we hold the world fixed and change how rich our model is.

Comparing Findings: Correctly Specified vs Misspecified

Correctly Specified ("Toy") World

- Our signals already span the true source of predictability.
- Complexity = high P/T : more parameters per observation.
- In this world, **complexity only makes learning harder**:
 - Out-of-sample R^2 deteriorates and can explode negatively at interpolation.
 - The timing Sharpe ratio falls away from the oracle benchmark and never catches up.

Misspecified ("Realistic") World

- The true DGP uses *more* signals than the empirical model initially includes.
- We index models by how large a subset of signals they use (fraction $q \in [0, 1]$).
- Here, **complexity has both a cost and a benefit**:
 - Cost: more parameters \rightarrow more estimation noise.
 - Benefit: larger $q \rightarrow$ less misspecification (we capture more of the true signal).
- With shrinkage tuned optimally, the benefit dominates: richer models do better.

Setup & Why Random Matrix Theory?

Data & Regression

- Single risky asset excess return: R_{t+1} .
- Predictors (signals): $S_t \in \mathbb{R}^P$ for $t = 1, \dots, T$.
- Population predictive model:

$$R_{t+1} = S_t' \beta + \varepsilon_{t+1}, \quad \mathbb{E}[\varepsilon_{t+1} \mid S_t] = 0.$$

Sample matrices

- Signal covariance:

$$\hat{\Psi} = \frac{1}{T} \sum_{t=1}^T S_t S_t' \in \mathbb{R}^{P \times P}.$$

- Signal–return covariance:

$$\hat{\Sigma}_{SR} = \frac{1}{T} \sum_{t=1}^T S_t R_{t+1} \in \mathbb{R}^P.$$

Ridge / ridgeless estimator

- Ridge penalty: $z \geq 0$.
- Estimator:

$$\hat{\beta}(z) = (zI_P + \hat{\Psi})^{-1} \hat{\Sigma}_{SR}.$$

- Ridgeless (minimum-norm) estimator: limit as $z \rightarrow 0^+$ when $P \geq T$.

High-dimensional regime

- Number of predictors P and sample size T both large.
- Complexity ratio: $c = P/T$ (not "small").
- Objects that drive prediction + timing performance are traces of functions of $\hat{\Psi}$, e.g.

$$\text{tr}[(\hat{\Psi} + zI_P)^{-1}], \quad \text{tr}[(\hat{\Psi} + zI_P)^{-1} \hat{\Psi}].$$

- In this regime, classical LLN/CLT no longer control these traces
→ **Random Matrix Theory (RMT)** is needed.

Correctly Specified “Toy” World

Stieltjes transform of the sample spectrum

- In the correctly specified world, the empirical model uses the same signals as the true model.
- Define the **Stieltjes transform** of the limiting eigenvalue distribution of $\hat{\Psi}$:

$$m(-z; c) := \lim_{P \rightarrow \infty} \frac{1}{P} \text{tr} \left((\hat{\Psi} + zI_P)^{-1} \right).$$

- Using Marčenko–Pastur–type results (proved in the paper), this limit exists and depends only on:
 - the complexity ratio c , and
 - the spectral distribution of the population covariance $\Psi = \mathbb{E}[S_t S_t']$.

Mixed trace & Proposition 2 (proved in the paper)

- Many performance formulas involve the “mixed” trace:

$$\frac{1}{T} \text{tr} \left[(zI_P + \hat{\Psi})^{-1} \Psi \right].$$

- **Proposition 2** shows that, as $P, T \rightarrow \infty$ with $P/T \rightarrow c$,

$$\frac{1}{T} \text{tr} \left[(zI_P + \hat{\Psi})^{-1} \Psi \right] \rightarrow \xi(z; c) = \frac{1 - z m(-z; c)}{c^{-1} - 1 + z m(-z; c)}.$$

- This is a central RMT result in the paper: it compresses the interaction of $\hat{\Psi}$ and Ψ into the scalars $m(-z; c)$ and $\xi(z; c)$.

From traces to economic quantities (Propositions 3–4, proved in the paper)

Using $m(-z; c)$ and $\xi(z; c)$, the authors derive closed-form limits for:

- Prediction MSE and out-of-sample $R^2(z; c)$.
- Expected timing return $\mathcal{E}(z; c)$ and leverage $\mathcal{L}(z; c)$.
- Second moment and volatility $\mathcal{V}(z; c)$ of timing returns.
- Sharpe ratio $SR(z; c) = \mathcal{E}(z; c) / \sqrt{\mathcal{V}(z; c)}$.

These results (Propositions 3–4) **are proven in the paper** and generate the theoretical VoC curves in the correctly specified setting (double descent in R^2 , behavior of $\|\hat{\beta}(z)\|$, SR vs c , etc.).

Misspecified “Realistic” World

Block structure and partial use of signals

- True signal vector is partitioned:

$$S_t = (S_t^{(1)}, S_t^{(2)}), \quad S_t^{(1)} \in \mathbb{R}^{P_1}, \quad S_t^{(2)} \in \mathbb{R}^{P_2}, \quad P_1 = qP.$$

- Empirical model uses only $S_t^{(1)}$, so it operates in dimension P_1 with empirical complexity $c_q = P_1/T = cq$.
- Population covariance is block-partitioned:

$$\Psi = \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix},$$

and the sample covariance of used signals is $\hat{\Psi}_{11} = T^{-1} \sum S_t^{(1)} S_t^{(1)'}$.

RMT for the reduced covariance

- Define the Stieltjes transform for the eigenvalues of $\hat{\Psi}_{11}$:

$$m(-z; c_q; q) := \lim_{P_1 \rightarrow \infty} \frac{1}{P_1} \text{tr} \left((\hat{\Psi}_{11} + zI_{P_1})^{-1} \right).$$

- **Proposition 5** (proved in the paper) shows that, using generalized MP theory, one can again express mixed traces such as

$$\frac{1}{T} \text{tr} \left[(zI_{P_1} + \hat{\Psi}_{11})^{-1} \Psi_{11} \right]$$

in terms of scalar functions $\xi(z; c_q; q)$, themselves functions of $m(-z; c_q; q)$ and the spectrum of Ψ_{11} .

From block traces to misspecified performance (Proposition 6 & Theorem 1)

- Plugging these limits into the same MSE / timing-return decomposition yields:
 - $\mathcal{E}(z; c_q; q)$: expected timing return,
 - $\mathcal{L}(z; c_q; q)$: leverage,
 - $R^2(z; c_q; q)$: OOS R^2 ,
 - $SR(z; c_q; q)$: Sharpe ratio, now **all as functions of c, q, z** and the RMT scalars.
- **Proposition 6 and Theorem 1** (proved in the paper) analyze these expressions and show that, with **optimally chosen ridge $z_*(q)$** :
 - $R^2(z_*(q); c_q; q)$ is **strictly increasing and concave in q** .
 - $SR(z_*(q); c_q; q)$ is also **strictly increasing and concave in q** .

This is the formal “Virtue of Complexity” theorem in the misspecified world.

Big Picture: How RMT Drives the Finance Results

- **Without RMT**, we would only have simulations; we could not cleanly relate:
 - model dimension P ,
 - sample size T ,
 - ridge z , to out-of-sample R^2 and Sharpe.
- **With RMT**, the paper:
 - Replaces messy eigenvalue clouds of $\hat{\Psi}$ and $\hat{\Psi}_{11}$ by a few scalar transforms $m(\cdot)$, $\xi(\cdot)$, etc.
 - Derives **closed-form asymptotics** for prediction error and market-timing performance in both:
 - the correctly specified world (Propositions 2–4), and
 - the misspecified world (Propositions 5–6, Theorem 1).
 - Allows the authors to rigorously state:
 - When complexity is a **pure cost** (correct specification, high c).
 - When complexity becomes a **virtue** (misspecification + shrinkage, increasing q).

All of these theorems are **proved in the paper** using standard but nontrivial tools from random matrix theory (Marčenko–Pastur limits, Stieltjes transforms, and trace identities).

Correctly Specified world

Theorem A

Consider ridge / ridgeless regression in the correctly specified model as $P, T \rightarrow \infty$ with $P/T \rightarrow c > 0$.

1. Prediction: The limiting out-of-sample $R^2(z; c)$

- is always **below** the oracle $R^2(0; 0)$,
- is **decreasing in complexity** c for any fixed positive ridge z ,
- in the ridgeless case $z = 0$ shows **double descent**:
 - $R^2(0; c) \rightarrow -\infty$ as $c \uparrow 1$,
 - $R^2(0; c)$ rises again and becomes positive as $c \rightarrow \infty$.

2. Timing performance: The Sharpe ratio $SR(z; c)$

- is always **below** the oracle Sharpe $SR(0; 0)$,
- is **decreasing in** c for fixed $z > 0$,
- remains **strictly positive for all** c , even when R^2 is very negative.

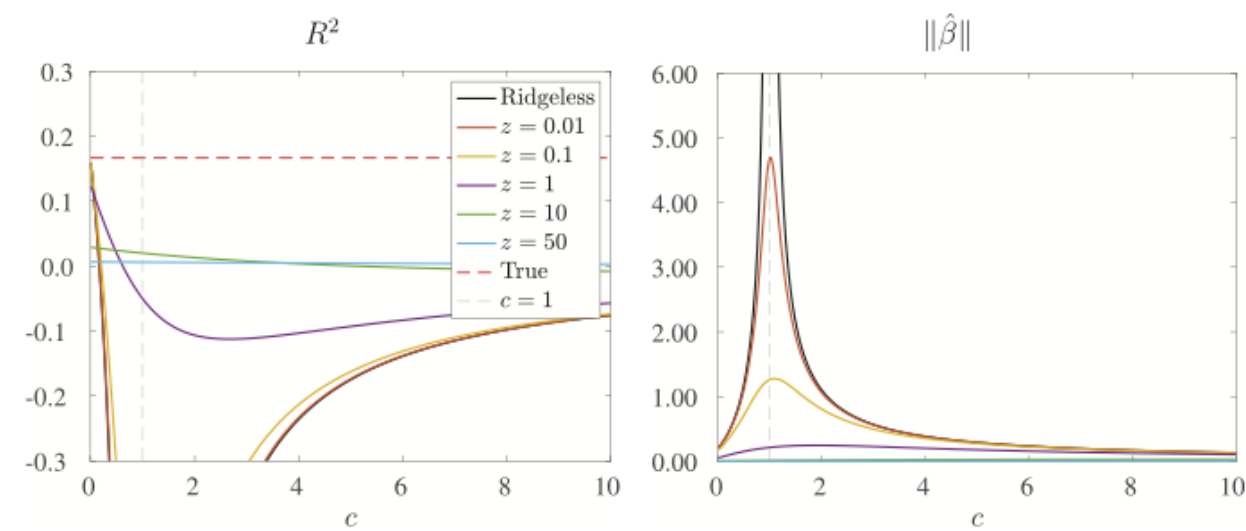


Figure 1. Expected out-of-sample R^2 and norm of least-squares coefficient. This figure shows the limiting out-of-sample R^2 and $\hat{\beta}$ norm as a function of c and z from Proposition 3 assuming Ψ is the identity matrix and $b_* = 0.2$. (Color figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com))

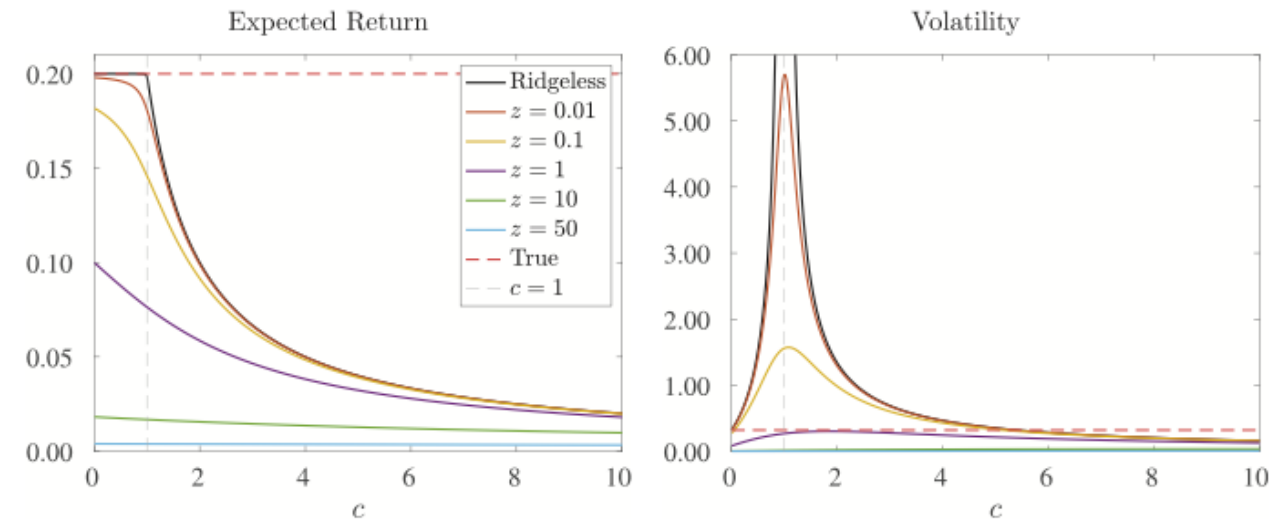


Figure 2. Expected out-of-sample risk and return of market timing. This figure shows the limiting out-of-sample expected return and volatility of the market timing strategy as a function of c and z from Proposition 3 assuming Ψ is the identity matrix and $b_* = 0.2$. (Color figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com))

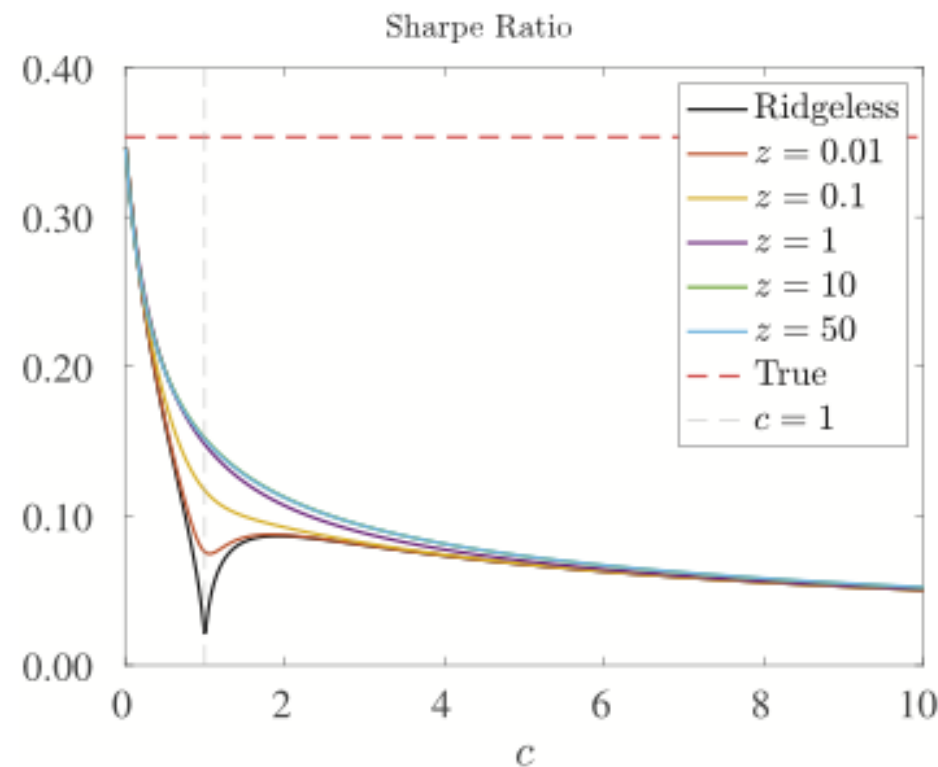


Figure 3. Expected out-of-sample Sharpe ratio of market timing. This figure shows the limiting out-of-sample Sharpe ratio of the market timing strategy as a function of c and z from Proposition 3 assuming Ψ is the identity matrix and $b_* = 0.2$. (Color figure can be viewed at wileyonlinelibrary.com)

Misspecified World

Theorem B

Fix a high-dimensional true DGP with complexity $c = P/T$, and let the empirical model use only a fraction $q \in [0, 1]$ of the true signals (so empirical complexity is $c_q = cq$). Let $z_*(q)$ be the optimal ridge level.

1. **Prediction:** The limiting out-of-sample $R^2(z_*(q); c_q; q)$

- starts near 0 when q is small (very few signals used),
- is **strictly increasing and concave in q** ,
- approaches the oracle R^2 as $q \rightarrow 1$ (we use almost all signals).

2. **Timing performance:** The Sharpe ratio $SR(z_*(q); c_q; q)$

- is also **strictly increasing and concave in q** ,
- converges to the oracle Sharpe as $q \rightarrow 1$.

This is the formal "Virtue of Complexity" result.

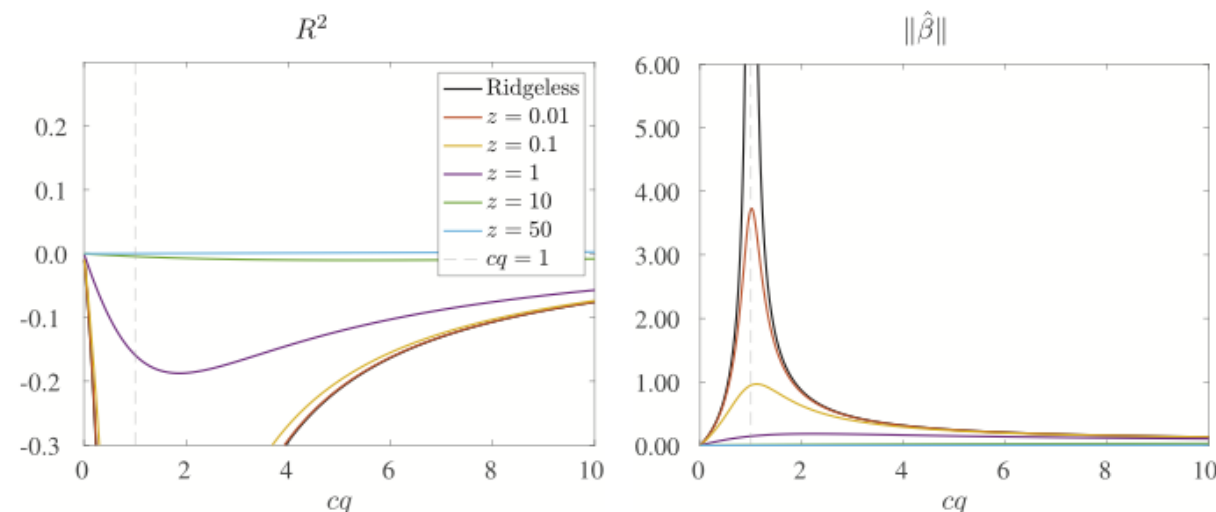


Figure 4. Expected out-of-sample prediction accuracy from misspecified models. This figure shows the limiting out-of-sample R^2 and $\hat{\beta}$ norm as a function of c and z from Proposition 6 assuming Ψ is the identity matrix, $b_* = 0.2$, and the complexity of the true model is $c = 10$. (Color figure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

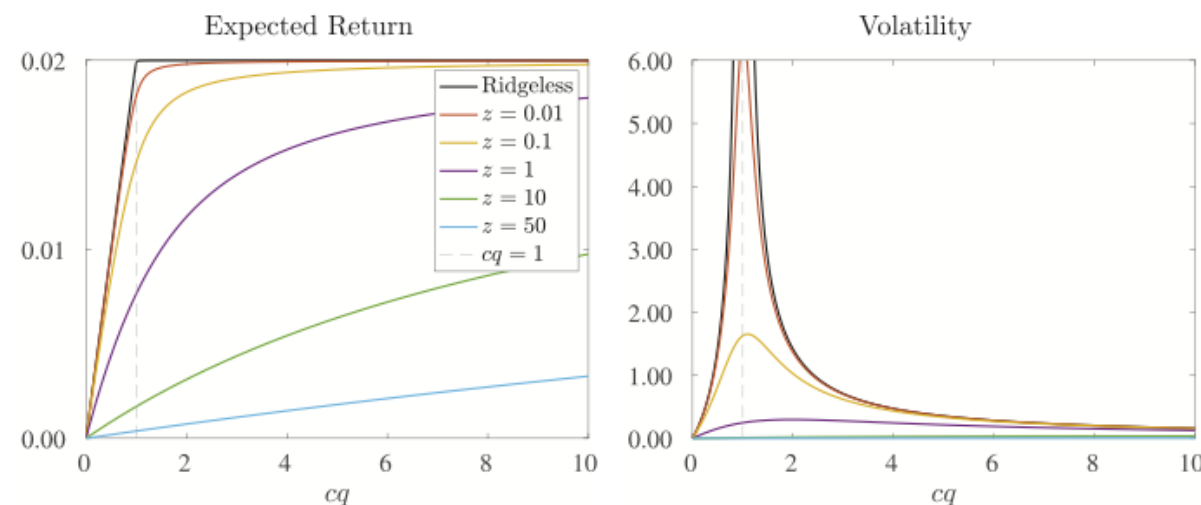


Figure 5. Expected out-of-sample risk and return from misspecified models. This figure shows the limiting out-of-sample expected return and volatility of the market timing strategy as a function of c and z from Proposition 6 assuming Ψ is the identity matrix, $b_* = 0.2$, and the complexity of the true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

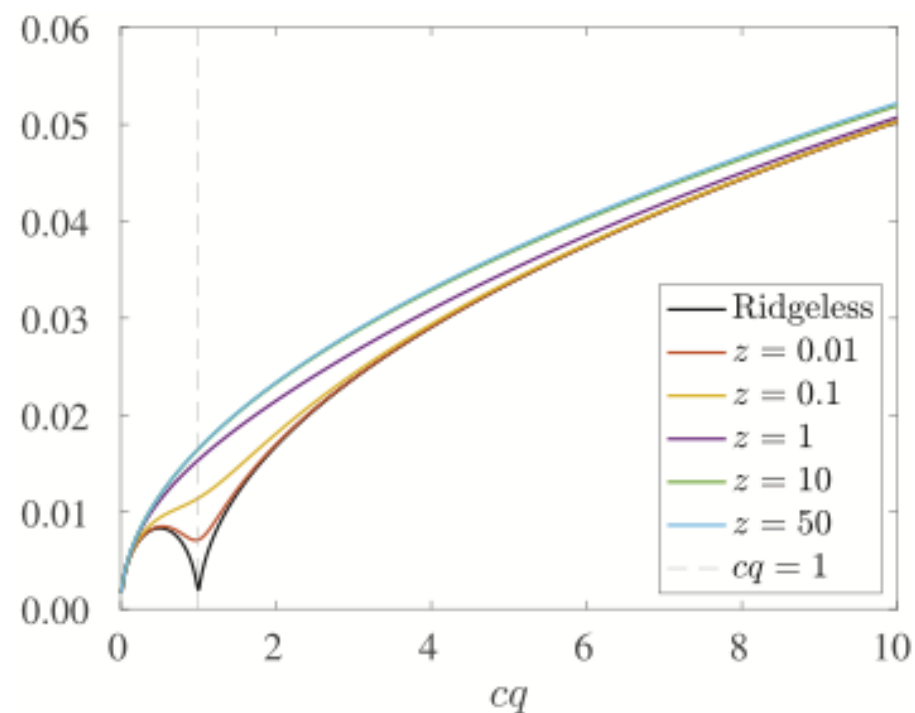


Figure 6. Expected out-of-sample Sharpe ratio from misspecified models. This figure shows the limiting out-of-sample Sharpe ratio of the market timing strategy as a function of c and z from Proposition 6 assuming Ψ is the identity matrix, $b_* = 0.2$, and the complexity of the true model is $c = 10$.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

Virtue of complexity: empirical evidence from market timing

- Data
- Random Fourier Features
- Out of Sample Performance

Data

- **Goal:**
Test the "Virtue of Complexity" predictions in a standard **market-timing** setting.
- **Target Variable:**
 - Monthly **excess return** on the **CRSP value-weighted U.S. stock market index**.
- **Predictor Set:**
 - The **15 predictive variables** from Goyal & Welch (2008) (dividend–price ratio, term spread, default spread, etc.).
 - Monthly data from **1926–2020**.
- **Standardization of Returns:**
 - Returns are scaled by their **trailing 12-month standard deviation**
→ keeps the forecast problem in excess-return units but with roughly stable conditional volatility.
- **Standardization of Predictors:**
 - Each predictor is scaled using an **expanding-window historical standard deviation** (requires at least 36 months of data → empirical sample starts in **1930**).
 - This aligns the empirical setup with the theoretical assumption of **homoskedastic signals**.
- **Robustness Note:**
 - Authors emphasize that their qualitative findings are **not sensitive** to the exact standardization choices.

Welch–Goyal Predictors Used in VoC (Core Set)

Valuation ratios

- **DP** – log dividend–price ratio
- **DY** – log dividend yield (dividends / lagged price)
- **EP** – log earnings–price ratio
- **DE** – log dividend–earnings (payout) ratio

Equity risk & corporate characteristics

- **SVAR** – stock return variance (sum of squared daily S&P 500 returns)
- **BM** – book-to-market ratio (Dow Jones book value / market value)
- **NTIS** – net equity expansion (12-month net issues / end-of-year NYSE market cap)

Interest-rate level & slope

- **TBL** – 3-month Treasury-bill rate (short rate)
- **LTY** – long-term government bond yield
- **LTR** – long-term government bond return
- **TMS** – term spread = LTY – TBL

Credit risk

- **DFY** – default yield spread = BAA – AAA corporate bond yields
- **DFR** – default return spread = long-term corporate bond return – long-term government bond return

Macro

- **INFL** – inflation rate (CPI, lagged one extra month)

A Comprehensive Look at the Empirical Performance of Equity Premium Prediction

Yale ICF Working Paper No. 04-11

59 Pages • Posted: 30 Apr 2004

[Amit Goyal](#)

University of Lausanne; Swiss Finance Institute

[Ivo Welch](#)

University of California, Los Angeles (UCLA); National Bureau of Economic Research (NBER)

 [There are 3 versions of this paper](#)

Date Written: January 11, 2006

Abstract

Economists have suggested a whole range of variables that predict the equity premium: dividend price ratios, dividend yields, earnings-price ratios, dividend payout ratios, corporate or net issuing ratios, book-market ratios, beta premia, interest rates (in various guises), and consumption-based macroeconomic ratios (cay). Our paper comprehensively reexamines the performance of these variables, both in-sample and out-of-sample, as of 2005. We find that [a] over the last 30 years, the prediction models have failed both in-sample and out-of-sample; [b] the models are unstable, in that their out-of-sample predictions have performed unexpectedly poorly; [c] the models would not have helped an investor with access only to information available at the time to time the market.

Keywords: Equity Premium, Prediction, Stock Market

JEL Classification: G12, G14

Suggested Citation:

Goyal, Amit and Welch, Ivo, A Comprehensive Look at the Empirical Performance of Equity Premium Prediction (January 11, 2006). Yale ICF Working Paper No. 04-11, Available at SSRN: <https://ssrn.com/abstract=517667>

[source](#)

Random Features

Random Features for Large-Scale Kernel Machines

Ali Rahimi
Intel Research Seattle
Seattle, WA 98105

ali.rahimi@intel.com

Benjamin Recht
Caltech IST
Pasadena, CA 91125
brecht@ist.caltech.edu

Abstract

To accelerate the training of kernel machines, we propose to map the input data to a randomized low-dimensional feature space and then apply existing fast linear methods. The features are designed so that the inner products of the transformed data are approximately equal to those in the feature space of a user specified shift-invariant kernel. We explore two sets of random features, provide convergence bounds on their ability to approximate various radial basis kernels, and show that in large-scale classification and regression tasks linear machine learning algorithms applied to these features outperform state-of-the-art large-scale kernel machines.

[source](#)

Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning

Paper #858

Abstract

Randomized neural networks are immortalized in this AI Koan:

In the days when Sussman was a novice, Minsky once came to him as he sat hacking at the PDP-6.

"What are you doing?" asked Minsky. "I am training a randomly wired neural net to play tic-tac-toe," Sussman replied. "Why is the net wired randomly?" asked Minsky. Sussman replied, "I do not want it to have any preconceptions of how to play."

Minsky then shut his eyes. "Why do you close your eyes?" Sussman asked his teacher. "So that the room will be empty," replied Minsky. At that moment, Sussman was enlightened.

We analyze shallow random networks with the help of concentration of measure inequalities. Specifically, we consider architectures that compute a weighted sum of their inputs after passing them through a bank of arbitrary randomized nonlinearities. We identify conditions under which these networks exhibit good classification performance, and bound their test error in terms of the size of the dataset and the number of random nonlinearities.

[source](#)

Random Fourier Features

- **Base predictors:**

- Let $G_t \in \mathbb{R}^{15}$ be the vector of the 15 Welch–Goyal predictors at time t .

- **Idea:**

- Use **Random Fourier Features** to build a *much higher-dimensional* feature vector S_t from G_t .
- This lets us smoothly move from very low-dimensional to very high-dimensional models while always using the same economic inputs.

- **Construction (per random feature):**

- Draw a random weight vector $\omega_i \sim N(0, I_{15})$.
- Form a random linear combination of the predictors: $\omega_i' G_t$.
- Pass it through trigonometric functions with a scale parameter γ :

$$S_{i,t} = [\sin(\gamma \omega_i' G_t), \cos(\gamma \omega_i' G_t)].$$

- Each ω_i gives **two new features** (sine and cosine).

- **Controlling model complexity:**

- If we draw **1** random vector ω_i , we get $P = 2$ features \rightarrow a very low-dimensional model.
- If we draw **5,000** random vectors, we get $P = 10,000$ features \rightarrow a very high-dimensional model.
- Thus, by changing how many ω_i 's we sample, we can choose any desired feature dimension P .

- **Interpretation:**

- RFF approximates a general nonlinear function $f(G_t)$ for $\mathbb{E}[R_{t+1} \mid G_t]$.
- It is equivalent to a **two-layer neural network**:
 - first layer: fixed random weights ω_i and nonlinearities (sin/cos),
 - second layer: learned linear weights β in the regression.
- This is the bridge between the 15 economic predictors and the large- P linear models analyzed in the theory.

Empirical Experiments – How They Build “VoC Curves”

Goal:

Mimic the theoretical VoC analysis using real data: CRSP VW index returns + 15 Welch–Goyal predictors.

Rolling training windows (T):

- Use **1-year, 5-year, 10-year** rolling windows: $T \in \{12, 60, 120\}$ months.
- Short windows (e.g. $T = 12$) allow:
 - very high complexity ratios $c = P/T$ with moderate P ,
 - a stress test that shows virtue of complexity even in small samples.

Model complexity and shrinkage grid:

- Number of Random Fourier Features (RFFs):

$$P \in \{2, \dots, 12,000\}$$

.

- Ridge penalty grid:

$$\log_{10}(z) \in \{-3, -2, -1, 0, 1, 2, 3\}$$

Procedure for one run (fixed T, P, z):

1. Generate **12,000 RFFs** from the 15 predictors G_t (choose bandwidth γ).
2. Select the **first** P RFFs as the predictor vector S_t .
3. Run a **recursive rolling regression**:
 - For each time t in the OOS period (roughly 1,091 months), estimate the regression using the past T observations $\{(R_{t-\tau+1}, S_{t-\tau}), \dots\}$.
 - Use the estimated $\hat{\beta}$ to form:
 - out-of-sample return forecast $\hat{\beta}' S_t$,
 - timing return $\hat{\beta}' S_t \cdot R_{t+1}$.
4. From the OOS sequence, compute:
 - average $\|\hat{\beta}\|^2$ over windows,
 - OOS prediction R^2 ,
 - timing strategy's average return, volatility, and Sharpe ratio.

Monte Carlo over RFF randomness:

- Because RFFs are random, small- P models can be noisy.
- Repeat steps above **1,000 times** with independent RFF draws and **average** performance statistics.
- Plot these averaged statistics vs. model complexity $c = P/T$ to obtain the empirical **VoC curves**.

Summary of Empirical Findings (vs. Theory)

Object	Theoretical prediction (misspecified world)	Empirical evidence from VoC (CRSP + GW predictors)	Aligned with theory?
Out-of-sample R^2	Double-descent: very negative near interpolation $c \approx 1$, recovery and improvement in high-complexity regime.	Figure 7 & IA3: $\hat{\beta}$ and OOS R^2 blow up near $c = 1$; zoomed plots show R^2 turning positive again at large c when shrinkage is used.	Yes – same qualitative shape (explosion + recovery).
Coefficient norm $\ \hat{\beta}\ $	Peaks at $c \approx 1$, then shrinks as complexity increases (implicit shrinkage of ridgeless).	Empirical $\ \hat{\beta}\ ^2$ spikes at $c = 1$ and declines for large c (Figure 7 & IA1–IA2).	Yes – very close match.
Expected timing return	In misspecified case, increasing in model complexity (more of true signal captured), then flattening as $q \rightarrow 1$.	Figure 7: clear upward pattern in OOS average return as c rises; nearly flat beyond $c = 1$ for almost-ridgeless z , more gradual rise for higher z .	Yes – monotone increase + flattening.
Volatility of timing returns	Spike near interpolation, then decline with high complexity and shrinkage.	Volatility shows a sharp spike around $c = 1$ and falls as c increases, especially under stronger ridge (Figure 7).	Yes – same spike-and-decline pattern.
Sharpe ratio of timing strategy	With tuned shrinkage, Sharpe should increase with complexity and be concave; always positive.	Figure 8: Sharpe generally rises with c ; minor dip near $c = 1$ at very low z , then levels off in high c .	Mostly yes – small dip at interpolation, but overall increasing & concave.
Alpha & Information Ratio	Alpha and IR inherit the same monotone shape as expected return and Sharpe.	Figure 8–9: in high c , $IR \approx 0.3(T = 12)$ or $0.25(T = 60, 120)$ with t-stats > 2 ; patterns mirror Sharpe.	Yes – same monotone behavior.
Behavior of positions	High-complexity, shrunk models should look like smooth, mostly long positions that adjust around macro states.	Figure 10: positions are mostly long-only, negative bets rare and small; strategy cuts exposure before 14/15 NBER recessions, all out-of-sample.	Yes – positions behave as predicted.

Key Empirical Numbers at High Complexity (Illustrative)

Quantity	Order of magnitude / qualitative takeaway
Max OOS Sharpe (high c , tuned ridge)	$SR \gtrsim 0.4$
Information ratio (alpha vs buy-and-hold)	$IR \approx 0.3(T = 12); IR \approx 0.25(T = 60, 120)$
Alpha t-statistics	Roughly $2.6\text{--}2.9(T = 12); > 2.0(T = 60, 120)$
Behavior near interpolation $c \approx 1$	$\ \hat{\beta}\ $ and R^2 extremely unstable; volatility spikes; SR dips.
Behavior for large c with shrinkage	R^2 mostly positive; volatility low; SR, alpha, and IR all strong .
Qualitative shape of empirical VoC curves	Very close match to theoretical curves under misspecification (Figures 4–6).

OOS R^2 and $\|\hat{\beta}\|^2$ ($T = 12$)

- Plots out-of-sample R^2 and coefficient norm $\|\hat{\beta}\|^2$ vs. complexity c .
- R^2 :
 - Crashes near the interpolation boundary $c \approx 1$ (very negative).
 - Recovers and turns positive again in the high-complexity regime, especially with ridge.
- $\|\hat{\beta}\|^2$:
 - Spikes sharply at $c \approx 1$.
 - Declines as c grows beyond 1.
- Match to theory:** Yes – mirrors the misspecified VoC curves (Figure 4): double-descent in R^2 and peak-then-shrink in $\|\hat{\beta}\|^2$.

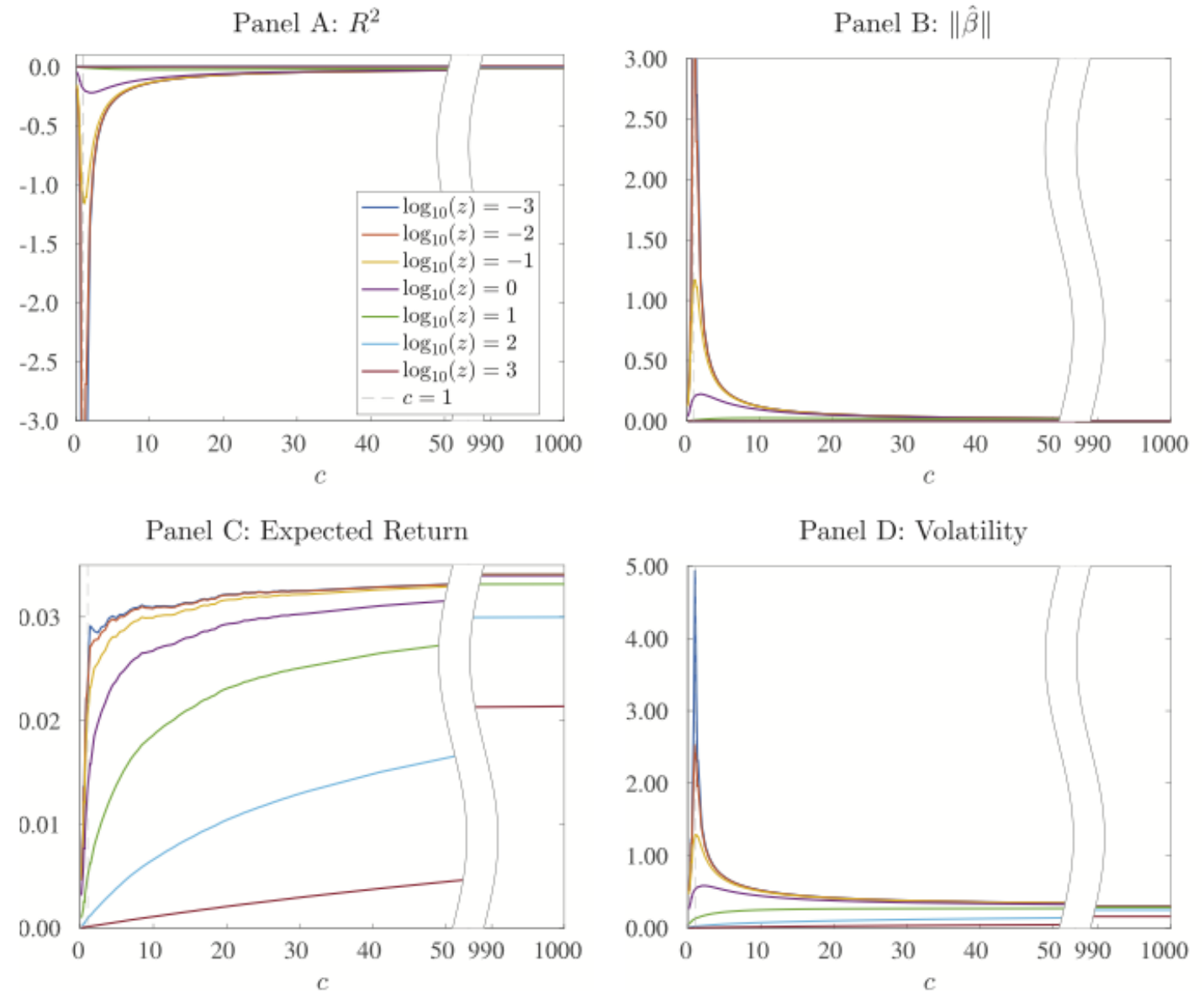


Figure 7. Out-of-sample market timing performance ($T = 12$). This figure shows the out-of-sample prediction accuracy and portfolio performance estimates for the empirical analysis described in Section V.C. The training window is $T = 12$ months and RFF count P (or cT) ranges from 2 to 12,000 with $\gamma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

Expected Return, Volatility, Sharpe, Alpha & IR

- Shows OOS average timing return, volatility, Sharpe ratio, and alpha/IR vs. c .
- Expected timing return:
 - Increases with model complexity; nearly flat beyond $c = 1$ in almost-ridgeless case.
 - With stronger ridge, the increase continues into very high c .
- Volatility:
 - Spikes near $c = 1$, then falls as c increases.
- Sharpe & IR:
 - Generally rise with complexity, with a small dip near $c = 1$ at very low ridge.
 - Reach values around $SR \geq 0.4$, $IR \approx 0.3$ in high-complexity region.
- **Match to theory:** Yes – richer models + shrinkage improve economic performance; Sharpe/IR are increasing and concave in complexity.

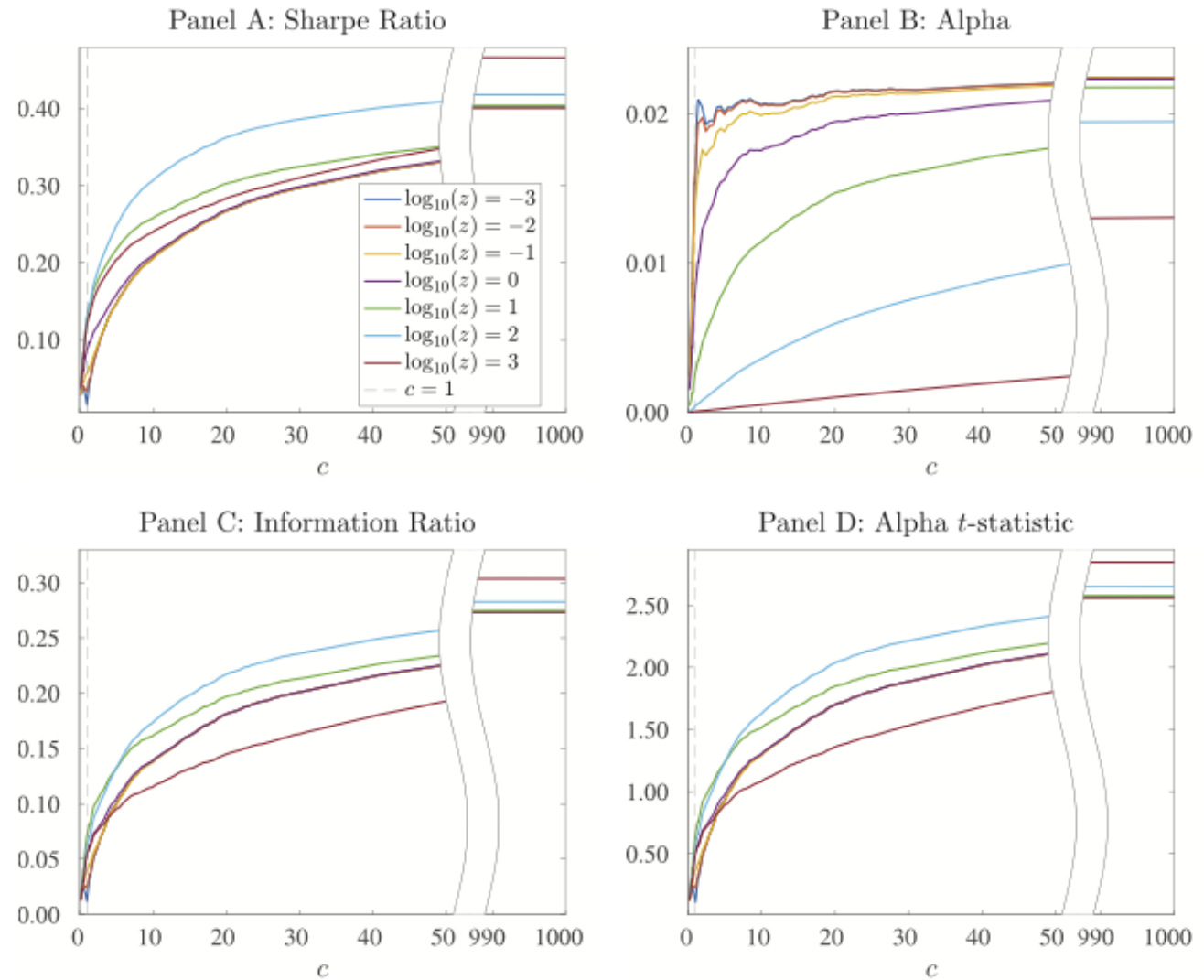


Figure 8. Out-of-sample market timing performance ($T = 12$). This figure shows the out-of-sample prediction accuracy and portfolio performance estimates for the empirical analysis described in Section V.C. The training window is $T = 12$ months and RFF count P (or cT) ranges from 2 to 12,000 with $\gamma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

Longer Training Windows ($T = 60, 120$)

- Repeats alpha and IR analysis for 5-year and 10-year training windows.
- Patterns:
 - $IR \approx 0.25$ in high-complexity regime.
 - Alpha t-stats above 2.0 across shrinkage levels.
 - Same rising-then-flattening shape in complexity as in Figure 8.
- **Match to theory:** Yes – confirms that VoC patterns depend on $c = P/T$, not on a particular short window; behavior is robust across T .

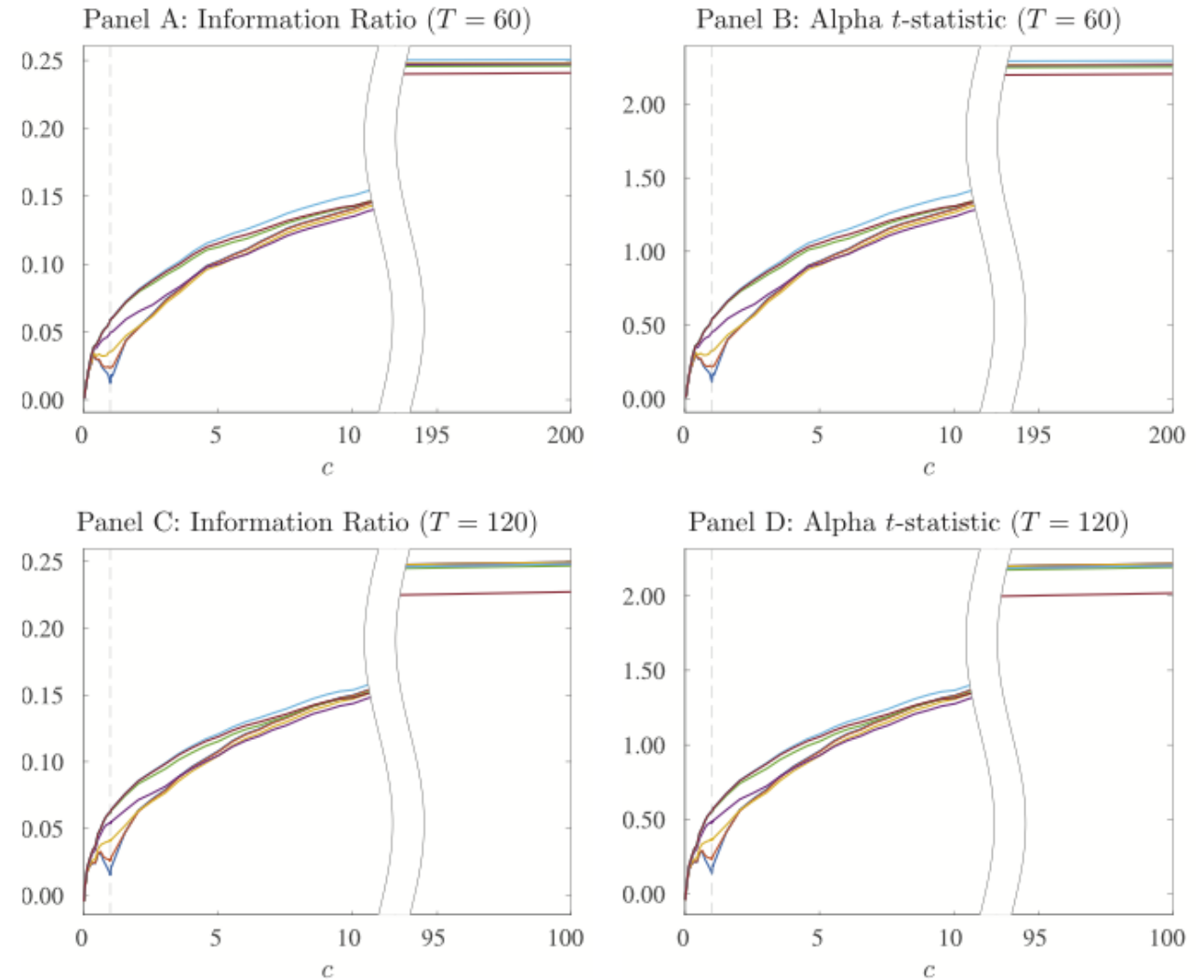


Figure 9. Out-of-sample market timing performance ($T = 60, 120$). This figure shows the out-of-sample prediction accuracy and portfolio performance estimates for the empirical analysis described in Section V.C. The training window is $T = 60$ or 120 months and RFF count P (or cT) ranges from 2 to 12,000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

Shape of High-Complexity Timing Positions

- Plots timing positions $\hat{\pi}_t(z, c)$ for a very high-complexity, high-shrinkage model (e.g. $P = 12,000$, large z) for $T = 12, 60, 120$.
- Positions:
 - Highly correlated across different training windows (90–97% correlations).
 - Mostly **long-only in spirit**: negative bets are rare and small.
- Business-cycle behavior:
 - Strategy systematically cuts exposure before NBER recessions.
 - For 14 out of 15 recessions, positions are substantially reduced ahead of the downturn, all out-of-sample.
- **Match to theory**: Yes – high-dimensional shrunk strategies behave like smooth, sign-stable market-timing rules that de-risk in bad states, as predicted.

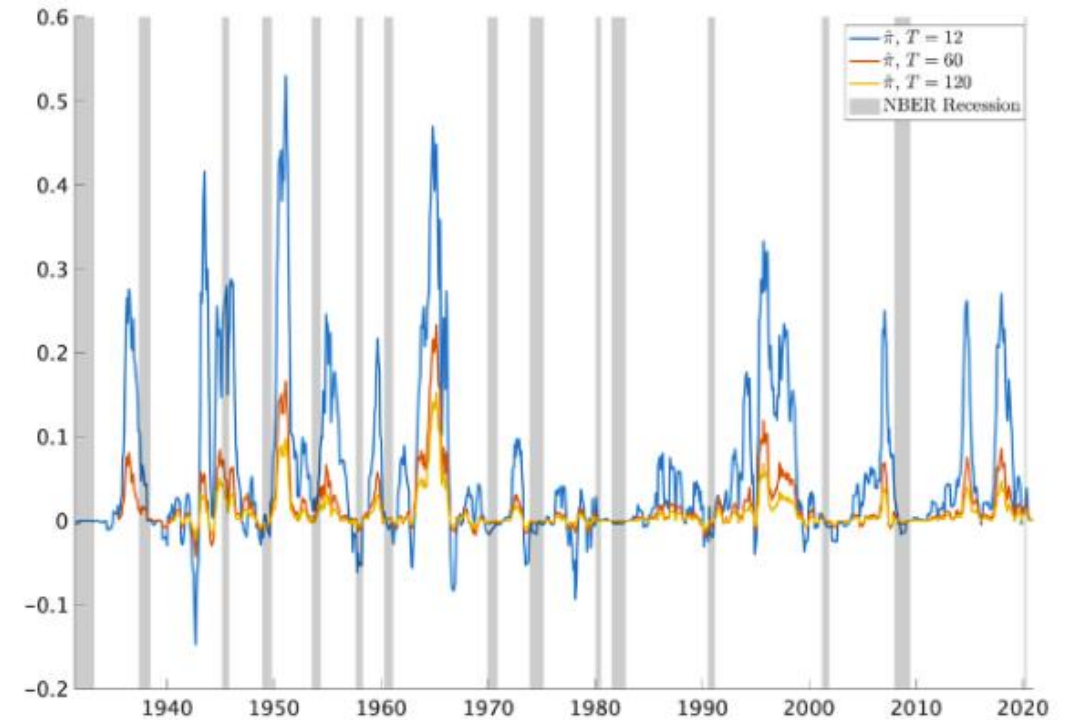


Figure 10. Market timing positions. This figure shows the out-of-sample prediction accuracy and portfolio performance estimates for the empirical analysis described in Section V.C. The training window is $T = 12, 60$, or 120 months with $P = 12,000$, $z = 10^3$, and $\gamma = 2$. Positions are averaged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

Goyal & Welch (2008) vs. Virtue of Complexity

Aspect	Goyal & Welch (2008) "Kitchen Sink"	Kelly–Malamud–Zhu VoC (Revisit of GW Data)
Information set	15 monthly GW predictors	Same 15 predictors
Model class	Linear OLS ("kitchen-sink" regression)	Linear ridge + high-dimensional RFF (nonlinear)
Complexity regime	$P = 15$, short windows already near $c \approx 1$	Explicitly views GW as near interpolation ; explores very high c with shrinkage
Performance metric emphasized	Out-of-sample prediction R^2	Out-of-sample timing performance (Sharpe, alpha, IR)
Baseline finding ($T = 12$)	OOS R^2 large negative ; timing $SR \approx -0.1 \rightarrow$ "return prediction fails"	Confirms bad R^2 , but interprets it as interpolation pathology, not lack of signal
Effect of ridge on GW kitchen sink	Not emphasized	With strong ridge (e.g. $z = 10^3$): R^2 still slightly negative, but timing $SR \approx 0.46, IR \approx 0.33, t \approx 3.1$ vs market
High-complexity nonlinear model	Not considered	RFF model with $c \approx 1000, z = 10^3$: OOS $R^2 \approx 1\%$ /month, $SR \approx 0.46, IR \approx 0.31, t \approx 2.5$ vs best linear model
Overall message	OOS R^2 is persistently negative \rightarrow market timing with these predictors isn't useful	Same data do support profitable timing once you (i) control interpolation with shrinkage and (ii) exploit the predictors in a high-dimensional nonlinear way

Variable Importance

- **Question:** How can high-complexity models learn useful patterns with only 12 months of data, when many predictors are highly persistent?
- **Method (Variable Importance, VI):**
 - Re-estimate the machine-learning timing model **15 times**, each time **dropping one predictor**.
 - Define VI for predictor i as the **change in performance** (OOS R^2 or Sharpe) when we go from the full 15-variable model to the 14-variable model that excludes i .
- **Key findings ($T = 12$, $P = 12,000$, $z = 10^3$):**
 - The **most important variables** are those with **high short-horizon variation** (least persistent): e.g. stock variance "svar", long-term bond return "ltr", default return "dfr". Dropping them reduces OOS R^2 by about 1.9%, 1.3%, and 0.8% per month.
 - VI measured using Sharpe tells the same story: these volatile predictors are crucial information sources for the high-complexity model.
- **Nonlinear benefits beyond simple linear effects:**
 - Linear models using the same predictors can have decent individual performance, but the **machine-learning model generates significant alpha over all of them**.
 - Its information ratio against every linear univariate timing strategy is large and significant ($IR \approx 0.32$ vs "All" combined, $t \approx 2.9$), showing that gains are not just from picking the right linear variable but from **exploiting nonlinear combinations**.

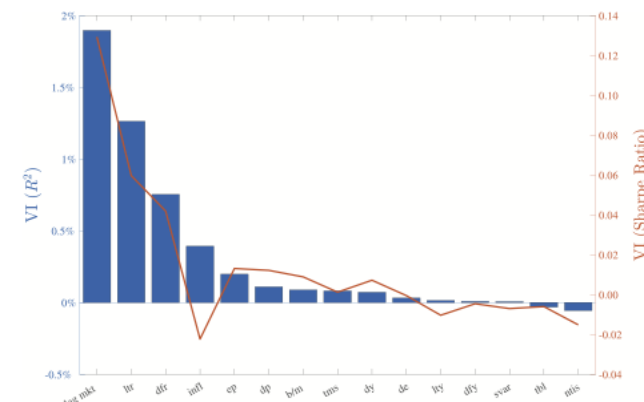


Figure 11. Variable importance. This figure shows the variable importance (VI) for the i^{th} predictor that is the change in performance, defined as out-of-sample R^2 or Sharpe ratio, moving from the full model with 15 variables to the reestimated model using 14 variables (excluding variable i). (Color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

Extent of Nonlinearity & Robustness

- **Linear vs Nonlinear RFF Models**

- RFF bandwidth γ controls **nonlinearity**:
 - As $\gamma \rightarrow 0$, $\sin(\gamma\omega'G_t) \approx \gamma\omega'G_t \rightarrow$ effectively a **linear random-features model**, equivalent to a transformed "kitchen sink" regression.
 - Larger γ introduces stronger nonlinearities in the mapping from predictors to returns.
- There is **no single optimal** γ : in high-complexity regimes, different γ 's approximate different projections of the true (unknown) nonlinear DGP, and none strictly dominates the others. Linear and nonlinear models contain **complementary information**.

- **Robustness Checks**

- Varying nonlinearity: results are robust when changing γ (e.g. 0.5, 1 vs baseline $\gamma = 2$).
- Volatility standardization: dropping return standardization does not alter the qualitative VoC patterns.
- Subsamples: splitting the sample (e.g. 1930–1974 vs 1975–2020) yields **similar shapes** for expected return, volatility, and IR vs complexity; magnitudes are about half as large in the later sample, consistent with fewer buying opportunities and smaller positions.
- Comparison with **12-month time-series momentum**:
 - If predictors were purely persistent, a 12-month high-complexity regression would behave like TS momentum.
 - The authors show instead that their machine-learning timing strategy has **economically large and statistically significant alpha over TS momentum**, and is driven by **higher-frequency fluctuations** in the predictors, not by slow trends alone.

Conclusion

- **Big picture:**
 - Asset pricing and asset management are rapidly adopting machine learning, but the behavior of **highly parameterized portfolios** is not well understood.
 - This paper provides a **theoretical and empirical foundation** for how such portfolios behave in out-of-sample market timing.
- **Main contributions:**
 - Show that **ridgeless least squares** and related ML models can achieve **positive Sharpe improvements even at very high complexity**, grounded in random matrix theory.
 - Demonstrate that **out-of-sample R^2** is a poor proxy for economic value: timing strategies can earn large profits even when predictive R^2 is small or negative.
 - Compare correctly specified vs misspecified worlds and show that, under misspecification with proper shrinkage, **larger models have a true “virtue of complexity.”**
- **Empirical confirmation:**
 - Using classic GW predictors and CRSP market returns, find out-of-sample **IRs around 0.3** versus buy-and-hold—statistically and economically significant.
 - High-complexity models behave like **long-only strategies that de-risk before recessions**, and they learn this behavior without hand-crafted constraints.
- **Implications for practice:**
 - Not a license to add arbitrary junk predictors; instead:
 - include **all plausibly relevant signals**, and
 - use **rich nonlinear models with shrinkage**, rather than forcing simple linear specs.
 - Even when the raw predictor set is small, using it inside **highly parameterized nonlinear models** can yield substantial gains.
- **Occam’s razor vs Occam’s blunder:**
 - The traditional view (Box, 1976) warns against overparameterization and favors parsimonious models.
 - This paper (and related ML theory) shows that small models are only preferable if **truly correctly specified**—a condition that is almost never met.
 - Logical conclusion: under broad conditions, **large models are preferable**, and the same lesson likely applies widely in finance and economics.

Replication repo

https://github.com/omroot/TheVirtueOfComplexity_PaperReplication

References

- Bryan Kelly , Semyon Malmud, Kangying Zhou (2023), [The Virtue of Complexity in Return Prediction](#)