The Virtue of Complexity

Bryan Kelly

Yale University and AQR Capital Management

In collaboration with: Semyon Malamud, Kangying Zhou, Antoine Didisheim, Barry Ke

“Principle of Parsimony” (Tukey, 1961)

Textbook Rule #1

> “It is important, in practice, that we employ the **smallest possible** number of parameters for adequate representations” (Box and Jenkins, *Time Series Analysis: Forecasting and Control*)

“Principle of Parsimony” (Tukey, 1961)

Textbook Rule #1

> “It is important, in practice, that we employ the **smallest possible** number of parameters for adequate representations” (Box and Jenkins, *Time Series Analysis: Forecasting and Control*)

Principle clashes with massive parameterizations adopted by modern ML algorithms

- Leading edge GPT-3 language model (Brown et al., 2020) uses 175 billion parameters
- Return prediction neural networks (Gu, Kelly, and Xiu, 2020) use 30,000+ parameters
- To Box-Jenkins econometrician, seems profligate, prone to overfit, and likely disastrous out-of-sample...

“Principle of Parsimony” (Tukey, 1961)

Textbook Rule #1

> “It is important, in practice, that we employ the **smallest possible** number of parameters for adequate representations” (Box and Jenkins, *Time Series Analysis: Forecasting and Control*)

Principle clashes with massive parameterizations adopted by modern ML algorithms

- Leading edge GPT-3 language model (Brown et al., 2020) uses 175 billion parameters
- Return prediction neural networks (Gu, Kelly, and Xiu, 2020) use 30,000+ parameters
- To Box-Jenkins econometrician, seems profligate, prone to overfit, and likely disastrous out-of-sample...

...But this is incorrect!

- Image/NLP models with astronomical parameterization—and *exactly fit* training data—are best performing models out-of-sample (Belkin, 2021)
- Evidently, modern machine learning has turned the principle of parsimony on its head

… And It's Happening In Finance Too

- Finance lit: Rapid advances in return prediction/portfolio choice using ML
- Large empirical gains over simple models
- Little theoretical understanding of why, and significant skepticism from old guard

What We Do: Building the "Case" for Financial ML

- **Main theoretical result**
  - Portfolio performance (Sharpe ratio) generally *increasing* in model complexity
- Explain the intuition, answer the skeptics
  - Prior evidence of empirical gains from ML are *what we should expect*
- Provide direct empirical support for theory

Problem Formulation

True Model: $R_{t+1} = f(G_t) + \epsilon_{t+1}$

- Predictors G may be known to the analyst, but the **prediction function f is unknown**
- Analyst cannot know true model, so instead she approximates f with large neural network:

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

- Each $S_{i,t} = \tilde{f}(w_i' G_t)$ is a known nonlinear function of original predictors

Problem Formulation

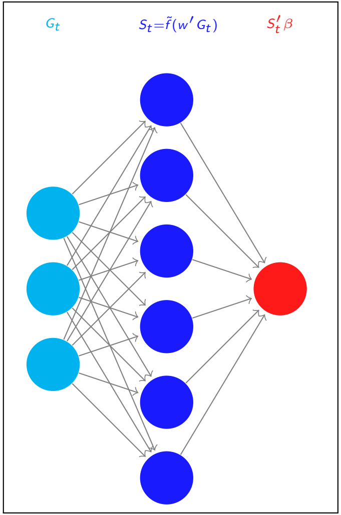
True Model: $R_{t+1} = f(G_t) + \epsilon_{t+1}$

- Predictors G may be known to the analyst, but the **prediction function f is unknown**
- Analyst cannot know true model, so instead she approximates f with large neural network:

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

- Each $S_{i,t} = \tilde{f}(w_i' G_t)$ is a known nonlinear function of original predictors

Empirical Model: $R_{t+1} = \sum_{i=1}^{P} S_{i,t}\beta_i + \tilde{\epsilon}_{t+1}$



Problem Formulation

True Model: $R_{t+1} = f(G_t) + \epsilon_{t+1}$

Empirical Model: $R_{t+1} = \sum_{i=1}^{P} S_{i,t}\beta_i + \tilde{\epsilon}_{t+1}$, where $S_{i,t} = \tilde{f}(w_i' G_t)$

The Choice:

- Given T data points, decide on "complexity" (number of features P) to use in approximating model

The Tradeoff:

- Simple model ($P << T$) has low variance thanks to parsimony, but is coarse approximator of f
- Complex model ($P > T$) is good approximator, but may behave poorly (and requires shrinkage)

Our Central Research Question:

- Which P should analyst opt for? Does benefit of more parameters justify their cost?

Problem Formulation

True Model: $R_{t+1} = f(G_t) + \epsilon_{t+1}$

Empirical Model: $R_{t+1} = \sum_{i=1}^{P} S_{i,t}\beta_i + \tilde{\epsilon}_{t+1}$, where $S_{i,t} = \tilde{f}(w_i' G_t)$

The Choice:

- Given T data points, decide on "complexity" (number of features P) to use in approximating model

The Tradeoff:

- Simple model ($P << T$) has low variance thanks to parsimony, but is coarse approximator of f
- Complex model ($P > T$) is good approximator, but may behave poorly (and requires shrinkage)

Our Central Research Question:

- Which P should analyst opt for? Does benefit of more parameters justify their cost?

Answer:

- Use the largest P you can compute

Environment

Model and Strategy

Model

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

- Single asset, R_{t+1}
- $P \times 1$ vector of predictor variables, S_t
- Linearity is without loss of generality
- Assumptions on β
  - Predictability identically distributed across signals (in expectation)
  - Total predictability is fixed

Timing Strategy

$$R^{\pi}_{t+1} = \pi_t R_{t+1}, \quad \pi_t = \beta' S_t.$$

- π_t: Timing weight scales asset position up/down to exploit time variation in expected return
- (Approximately) optimal for unconditional Sharpe maximization, convenient to analyze
- Results not sensitive to details of π function

Environment

Big Data + Big Model Limits

Goals of Theoretical Analysis

1. Characterize expected **out-of-sample** behaviors (prediction and portfolio performance)
   - All moments reported in "expected out-of-sample" form, *nothing in-sample*
2. Emphasize behavior of **machine learning** models, i.e., when number of parameters P is large
   - Differentiate between **correctly specified** versus **mis-specified** models

Tools

- Joint limits as numbers of observations and parameters are large, $T, P \to \infty$
- **Model complexity**, defined as $c = P/T$, arises as primary determinant of out-of-sample behaviors
- We leverage limiting results of random matrix theory

Why Do Big Models "Work"? Background From Least Squares

$$\boxed{R_{t+1} = \beta' S_t + \epsilon_{t+1}}$$

- Estimator when $P \leq T$: OLS

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

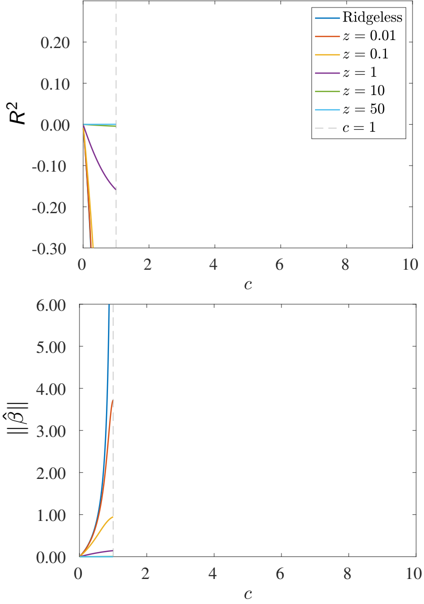
  - T equations in P unknowns ⇒ Unique solution for $\hat{\beta}$

- Estimator when $P > T$: Ridge Regression

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

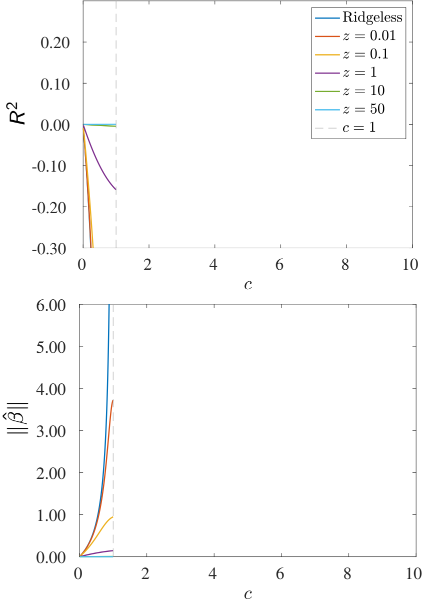
  - More unknowns (P) than equations (T) ⇒ Multiple solutions for $\hat{\beta}$
  - "Ridgeless" regression, $\lim_{z \to 0} \hat{\beta}(z) \equiv \hat{\beta}(0^+)$. Smallest variance solution that exactly fits training data

Why Do Big Models “Work”? Background From Least Squares



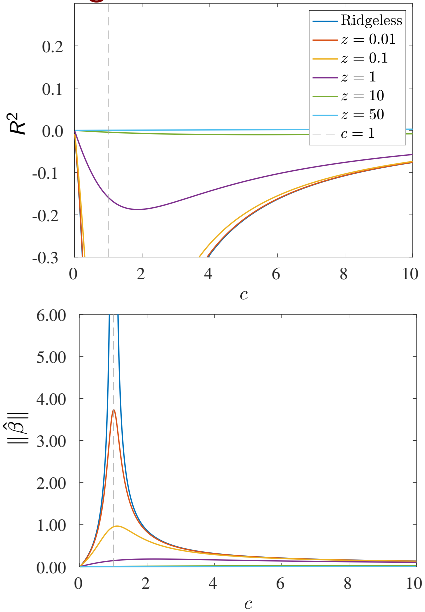
- P, $T \to \infty$ and $P/T \to c$
- $c = 0$: “Standard” asymptotics

Why Do Big Models "Work"? Background From Least Squares



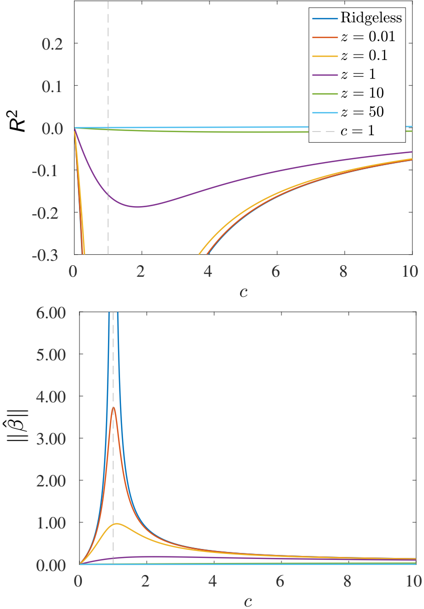
- $P, T \to \infty$ and $P/T \to c$
- $c = 0$: "Standard" asymptotics
- As $c \to 1$, expected out-of-sample R^2 tends to $-\infty$
  - Wild variance of estimates
  - Common interpretation is overfit: Exactly fit training data, but poor generalization out-of-sample
- Worrisome for trading strategy!
- Regularization helps mitigate

Why Do Big Models “Work”? Background From Least Squares



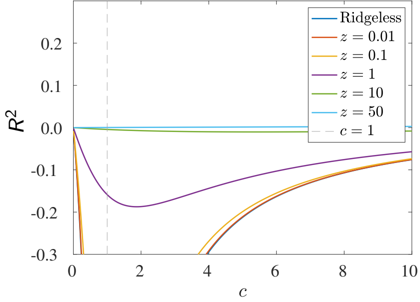
- When $c > 1$, “ridgeless” is $\lim_{z\to 0}\beta(z)$
- Counter-intuitively, OOS R^2 begins to *rise* with model complexity! Why?

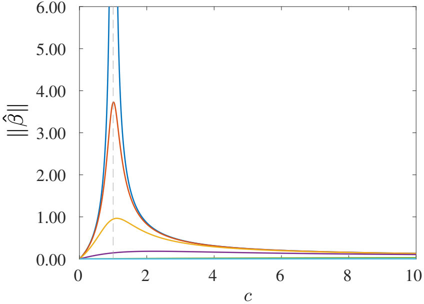
Why Do Big Models "Work"? Background From Least Squares



- When $c > 1$, "ridgeless" is $\lim_{z \to 0} \beta(z)$
- Counter-intuitively, OOS R^2 begins to *rise* with model complexity! Why?
- Many β's exactly fit training data, ridgeless selects one with smallest $\|\beta\|$
- Higher $c \Rightarrow$ more solutions to search over $\Rightarrow$ smaller $\|\beta\|$ with perfect training fit
- Shrinking β estimate despite $z \to 0 \Rightarrow$ forecast variance drops, R^2 improves

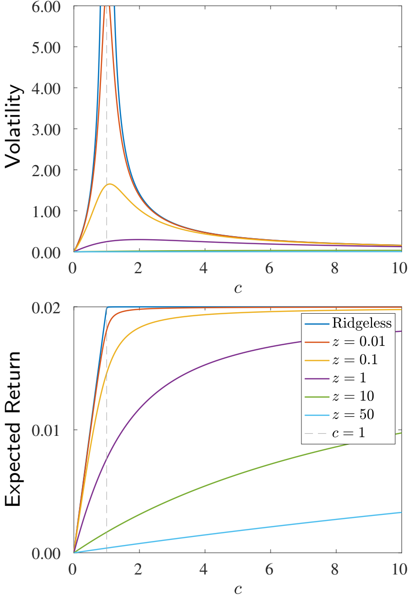
Why Do Big Models "Work"? Background From Least Squares





- When $c > 1$, "ridgeless" is $\lim_{z \to 0} \beta(z)$
- Counter-intuitively, OOS R^2 begins to *rise* with model complexity! Why?
- Many β's exactly fit training data, ridgeless selects one with smallest $\|\beta\|$
- Higher $c \Rightarrow$ more solutions to search over $\Rightarrow$ smaller $\|\beta\|$ with perfect training fit
- Shrinking β estimate despite $z \to 0 \Rightarrow$ forecast variance drops, R^2 improves
- Active topic of research in ML literature ("benign overfit," "double descent," ...)
- Challenges dogma of parsimony

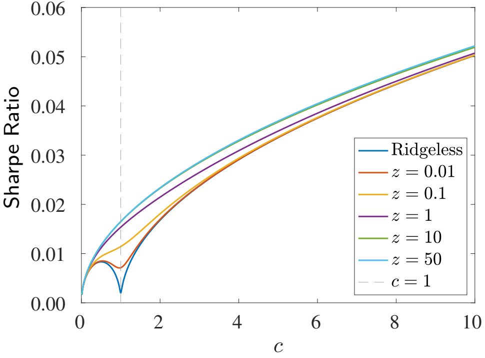
Why Do Big Models "Work"? The Trading Strategy Perspective



- $c = P/T$

1. Strategy variance
   - As $c \to 1$, strategy variance blows up. One β exactly fits training data, but it has high variance
   - When $c > 1$, variance *drops* with model complexity! Why?
   - Many β's exactly fit training data, ridge selects one with small variance
2. Strategy expected returns
   - ER low for $c \approx 0$ due to poor approximation of true model
   - Raising model complexity monotonically increases ER
   - Note the contrast with R^2!

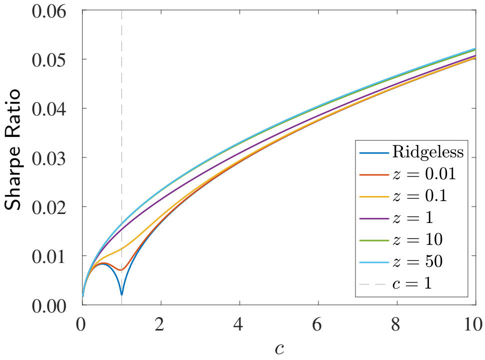
Why Do Big Models "Work"? The Trading Strategy Perspective



Main theory result

- Expected return always rises with model complexity (benefit of improved approximation)
- At same time, complex models have surprisingly low variance
- As a result, Sharpe ratio strictly increases with complexity

Why Do Big Models “Work”? The Trading Strategy Perspective



Main theory result

- Expected return always rises with model complexity (benefit of improved approximation)
- At same time, complex models have surprisingly low variance
- As a result, Sharpe ratio strictly increases with complexity

Complexity is a virtue. Approximation benefits dominate costs of heavy parameterization

- Paper provides general, rigorous theoretical statements and proofs that underlie plots
- Plots calculated from our theorems in a reasonable calibration

Empirical Analysis

- Analyze exact empirical analogues to theoretical comparative statics
- Focus on a cornerstone of empirical finance research—forecasting aggregate market return
- To make conclusions as easy to digest as possible, study conventional setting with conventional data
    - Forecast target is monthly return of CRSP value-weighted index 1926–2020
    - Info set consists of 15 predictor variables[†] from Welch and Goyal (WG, 2008)

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

Empirical Analysis

Random Fourier Features

- Empirical model: $R_{t+1} = S_t'\beta + \epsilon_{t+1}$
- Need framework to smoothly transition from low to high complexity

Empirical Analysis

Random Fourier Features

- Empirical model: $R_{t+1} = S_t'\beta + \epsilon_{t+1}$
- Need framework to smoothly transition from low to high complexity
- Adopt ML method known as "random Fourier features" (RFF)
  - Let G_t be 15×1 predictors. RFF converts G_t into

$$S_{i,t} = \sin(\omega_i' G_t), \quad \omega_i \sim iidN(0, \gamma I)$$

  - $S_{i,t}$: Random lin-combo of G_t fed through non-linear activation
- For fixed inputs, can create arbitrarily large (or small) feature set
  - Low-dim model (say $P = 1$) draw a single random weight
  - High-dim model (say $P = 10{,}000$) draw many weights

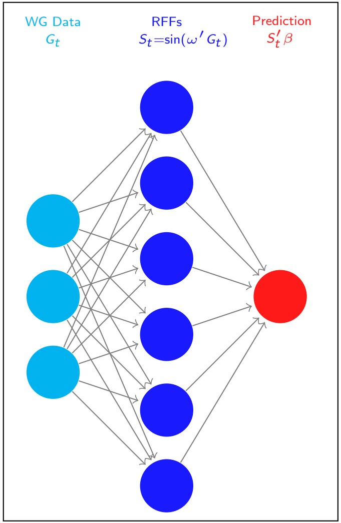
Empirical Analysis

Random Fourier Features

- Empirical model: $R_{t+1} = S_t'\beta + \epsilon_{t+1}$
- Need framework to smoothly transition from low to high complexity
- Adopt ML method known as "random Fourier features" (RFF)
  - Let G_t be 15×1 predictors. RFF converts G_t into

$$S_{i,t} = \sin(\omega_i' G_t), \quad \omega_i \sim iidN(0, \gamma I)$$

  - $S_{i,t}$: Random lin-combo of G_t fed through non-linear activation
- For fixed inputs, can create arbitrarily large (or small) feature set
  - Low-dim model (say $P = 1$) draw a single random weight
  - High-dim model (say $P = 10{,}000$) draw many weights
- In fact, RFF is two-layer neural network with fixed weights (ω_i) in first layer and optimized weights (regression β) in second layer



Empirical Analysis

Training and Testing

- One-year rolling training window ($T = 12$) and large set of RFFs
  1. Reach extreme levels of model complexity with smaller P and thus less computing burden
  2. Demonstrates virtue of complexity can be enjoyed in shockingly small samples
- Draw plots with model complexity $P = 1, \dots, 12{,}000$ and shrinkage of $\log_{10}(z) = -3, \dots, 3$

Empirical Analysis

Training and Testing

- One-year rolling training window ($T = 12$) and large set of RFFs
  1. Reach extreme levels of model complexity with smaller P and thus less computing burden
  2. Demonstrates virtue of complexity can be enjoyed in shockingly small samples
- Draw plots with model complexity $P = 1, ..., 12{,}000$ and shrinkage of $\log_{10}(z) = -3, ..., 3$

Empirical Procedure

i. Generate 12,000 RFFs

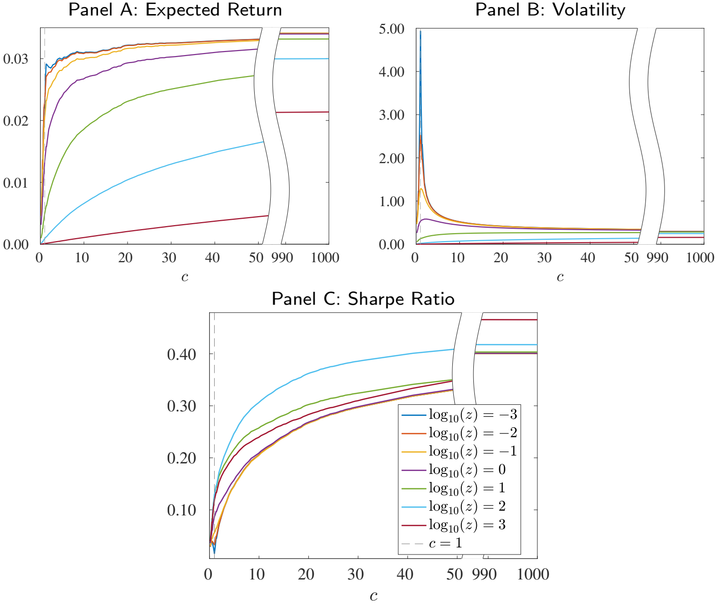
ii. Fix model defined by choice of (P, z)

iii. For each model (P, z), conduct recursive OOS prediction/timing strategy

iv. From OOS predictions, calculate ER, vol, and Sharpe of timing strategy

Out-of-sample Market Timing Performance

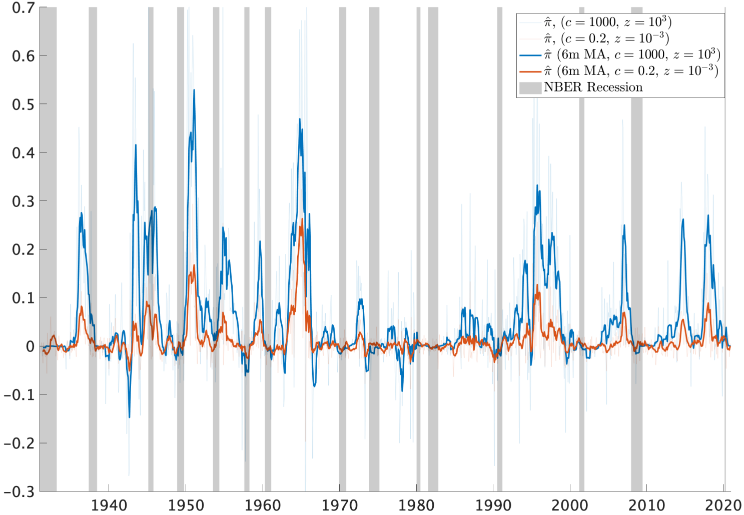
- Broadly: OOS behavior of ML predictions closely matches theory
- Variance explodes at $c \approx 1$ and recovers in high complexity regime
- Most importantly: OOS ER is increasing in complexity
- Sharpe of 0.4 p.a. for high complexity model. Mostly alpha/IR versus buy-and-hold with $t(\alpha) = 2.9$

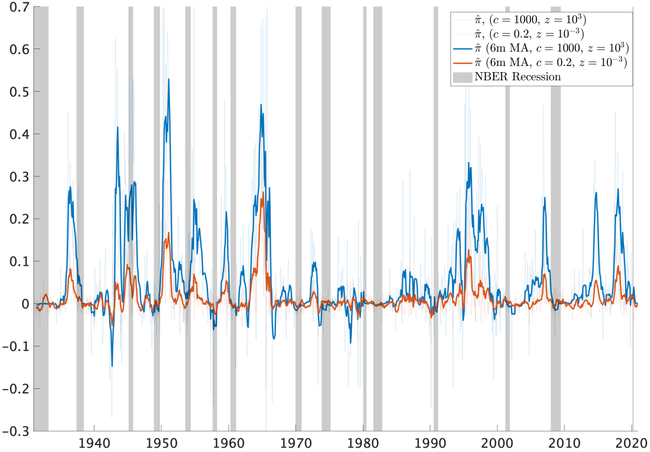

Panel A: Expected Return

Panel B: Volatility

Panel C: Sharpe Ratio

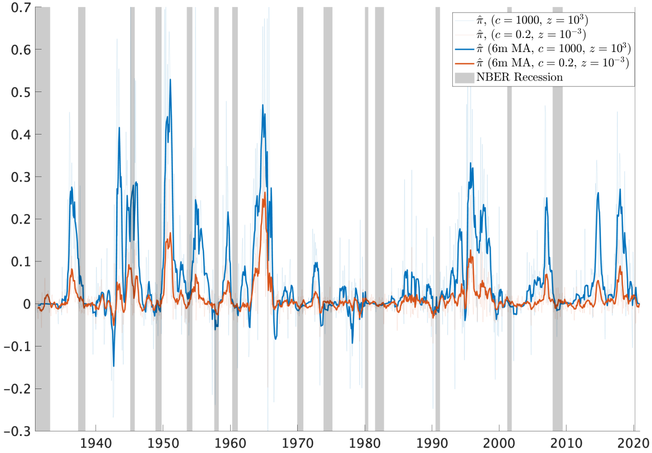
What Do ML Timing Bets Look Like?





1. **ML strategy is long-only at heart:** Almost never bets on market downturn

- Campbell and Thompson (2008) *"many predictive regressions beat the historical average return, once weak restrictions are imposed on the signs of coefficients and return forecasts"*
- Machine learns this without requiring explicit restriction



2. **Machine learns to divest ahead of recessions**

- For 14 of 15 recessions in OOS sample, essentially zeros out market position prior to recession (the exception is the 8-month recession of 1945)

Reconciling With The Literature

- Our results seem at odds with primary conclusion of Welch and Goyal (WG 2008), who argue market return prediction is a failed endeavor:

  "these models seem unstable, as diagnosed by their out-of-sample predictions and other statistics; and these models would not have helped an investor with access only to available information to profitably time the market."

- We use same predictive info. What is source of discrepancy?

1. **WG conclusions based on findings of consistently negative OOS R^2. They do not analyze implications for portfolio Sharpe ratios**
2. **Their c is close to 1 but doesn't use shrinkage**

Extensions

Virtue of Complexity Everywhere (Kelly, Malamud, and Zhou, 2022)

- Document identical pattern—OOS Sharpe ratio increasing in model complexity—in many asset classes
- US equities, international equities, bonds, commodities, currencies, and interest rates

Complexity in the Cross Section (Didisheim, Ke, Kelly, and Malamud, 2022)

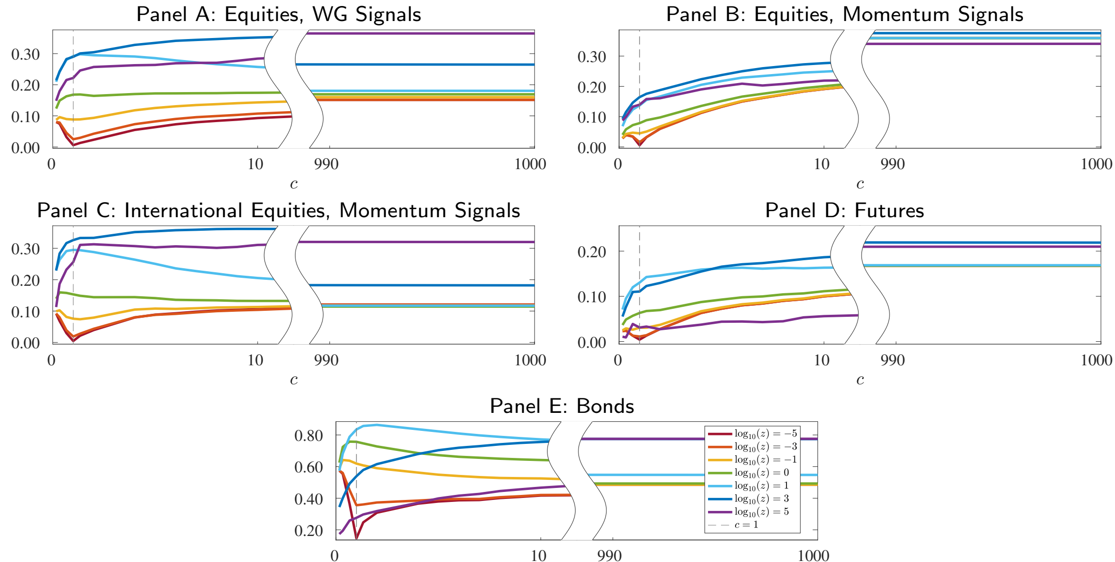
- Panel prediction problem, new results on virtue of complexity for stock selection

Deep Regression Ensembles (Didisheim, Kelly, Malamud, Kachman, and Rood, 2022)

- Introduce "deep" VoC models and apply to image recognition

“The Virtue of Complexity Everywhere”

Kelly, Malamud, and Zhou



Panel A: Equities, WG Signals

Panel B: Equities, Momentum Signals

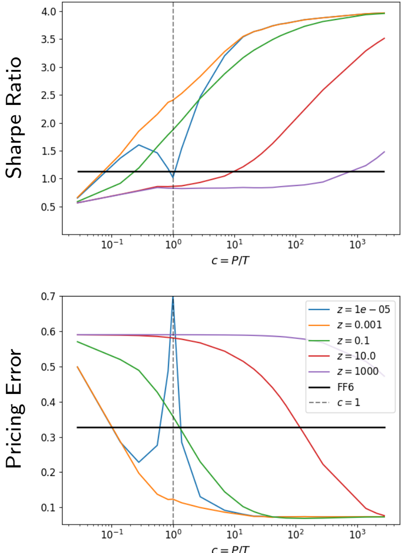
Panel C: International Equities, Momentum Signals

Panel D: Futures

Panel E: Bonds

“Complexity In Factor Pricing Models”

Kelly, Malamud, Didisheim, and Ke



Main Theoretical Result

- Best models have enormous numbers of factors (more than their are time series observations or base assets)
- Intuition: Complexity is necessary to approximate true SDF

Main Empirical Result

- OOS behavior of ML-based SDF closely matches theory
- High complexity models
  - Improve over simple models by a factor of 3
  - Dominate popular benchmarks like FF6

Conclusions, I

- Asset pricing and asset management in midst of boom in ML research
- We provide new, rigorous theoretical insight into the behavior of ML models/portfolios
- Contrary to conventional wisdom: Higher complexity improves model performance

Virtue of Complexity: Performance of ML portfolios can be improved by pushing model parameterization far beyond the number of training observations

Conclusions, I

- Asset pricing and asset management in midst of boom in ML research
- We provide new, rigorous theoretical insight into the behavior of ML models/portfolios
- Contrary to conventional wisdom: Higher complexity improves model performance

Virtue of Complexity: Performance of ML portfolios can be improved by pushing model parameterization far beyond the number of training observations

- *Not* license to add arbitrary predictors to model. Instead, we recommend
  1. including all plausibly relevant predictors
  2. using rich non-linear models rather than simple linear specifications
  - Doing so confers prediction/portfolio benefits, even when training data is scarce and particularly when accompanied by shrinkage

Conclusions, I

- Asset pricing and asset management in midst of boom in ML research
- We provide new, rigorous theoretical insight into the behavior of ML models/portfolios
- Contrary to conventional wisdom: Higher complexity improves model performance

Virtue of Complexity: Performance of ML portfolios can be improved by pushing model parameterization far beyond the number of training observations

- *Not* license to add arbitrary predictors to model. Instead, we recommend
  1. including all plausibly relevant predictors
  2. using rich non-linear models rather than simple linear specifications
  - Doing so confers prediction/portfolio benefits, even when training data is scarce and particularly when accompanied by shrinkage
- In canonical empirical problem—market prediction and timing—we find
  - OOS Sharpe nearly doubles relative to buy-and-hold strategy (highly significant)

Conclusions, II

- Clashes with philosophy of parsimony frequently espoused by economists
- Two oft-repeated quotes from famed statistician George Box:

All models are wrong, but some are useful.

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

Conclusions, II

- Clashes with philosophy of parsimony frequently espoused by economists
- Two oft-repeated quotes from famed statistician George Box:

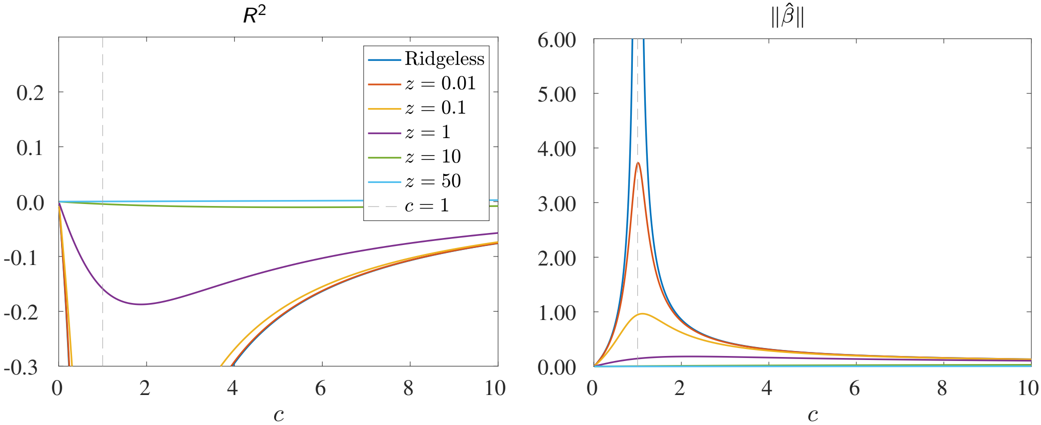
All models are wrong, but some are useful.

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

Occam’s Blunder? Small model is preferable only if it is correctly specified. But models are never correctly specified. Logical conclusion?

Appendix Slides

Out-of-Sample R^2 and Estimator Variance



R^2
Ridgeless
$z = 0.01$
$z = 0.1$
$z = 1$
$z = 10$
$z = 50$
$c = 1$
0.2
0.1
0.0
-0.1
-0.2
-0.3
0
2
4
6
8
10
c
$\|\hat{\beta}\|$
6.00
5.00
4.00
3.00
2.00
1.00
0.00
0
2
4
6
8
10
c