The Virtue of Complexity Everywhere

Bryan Kelly
Yale, AQR, NBER

Semyon Malamud
SFI, EPFL, and CEPR

Kangying Zhou
Yale

“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 in **US equities**, **international equities**, **futures**, and **bonds** markets

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

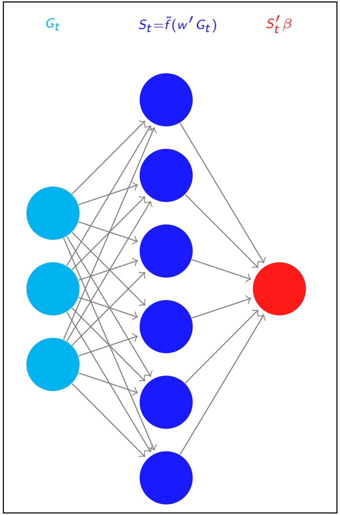
True Model: $R_{t+1} = f(G_t) + \epsilon_{t+1}$

- 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

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

$$\boxed{R_{t+1} = \beta' S_t + \tilde{\epsilon}_{t+1}}$$

- Estimator when $P \leq T$: OLS

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

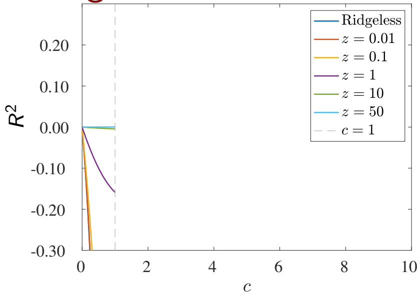
  - T equations in P unknowns $\Rightarrow$ Unique solution for $\hat{\beta}$

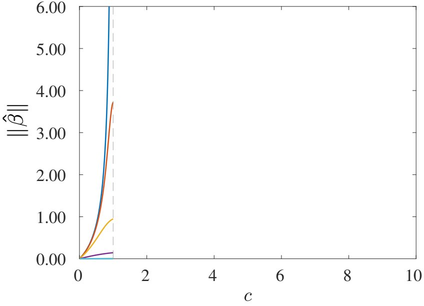
- Estimator when $P > T$: Ridge Regression

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

  - More unknowns (P) than equations (T) $\Rightarrow$ 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

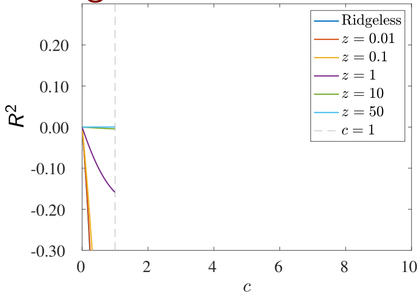
Why Do Big Models "Work"? Background From Least Squares

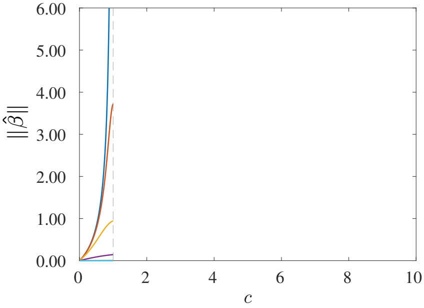




- $c = P/T$
- $c = 0$: "Standard" asymptotics

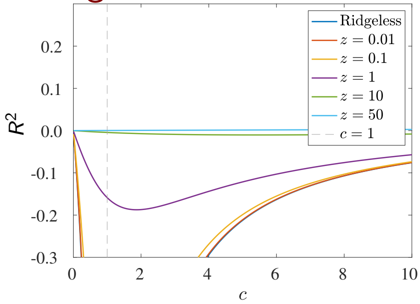
Why Do Big Models "Work"? Background From Least Squares

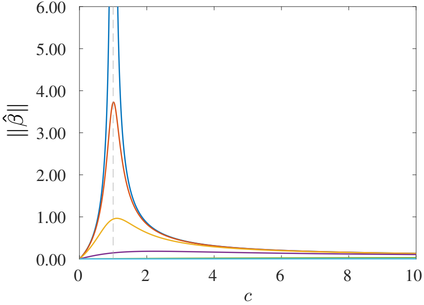




- $c = P/T$
- $c = 0$: "Standard" asymptotics
- As $c \to 1$, expected out-of-sample R^2 tends to $-\infty$
  - Wild variance of estimates
  - Common interpretation is overfit: Exactly fit training data, but poor generalization out-of-sample
- Worrisome for trading strategy!
- Regularization helps mitigate

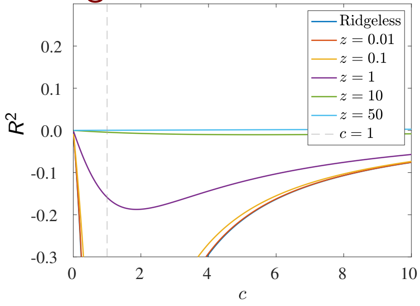
Why Do Big Models “Work”? Background From Least Squares

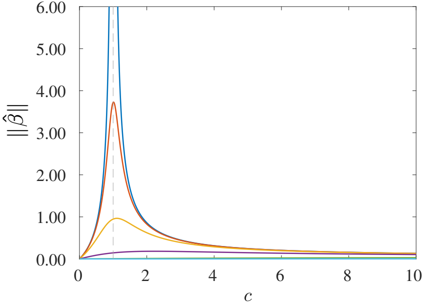




- When $c > 1$, “ridgeless” is $\lim_{z \to 0} \beta(z)$
- Counter-intuitively, OOS R^2 begins to *rise* with model complexity! Why?

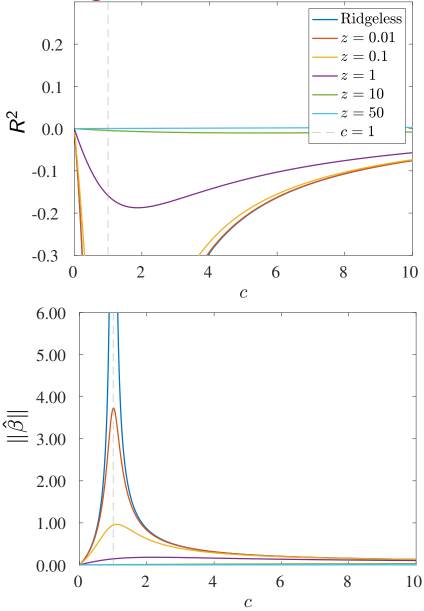
Why Do Big Models "Work"? Background From Least Squares





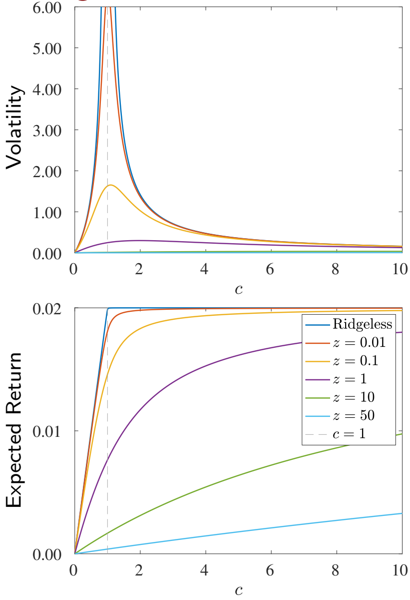
- When $c > 1$, "ridgeless" is $\lim_{z \to 0} \beta(z)$
- Counter-intuitively, OOS R^2 begins to *rise* with model complexity! Why?
- Many β's exactly fit training data, ridgeless selects one with smallest $\|\beta\|$
- Higher $c \Rightarrow$ more solutions to search over $\Rightarrow$ smaller $\|\beta\|$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

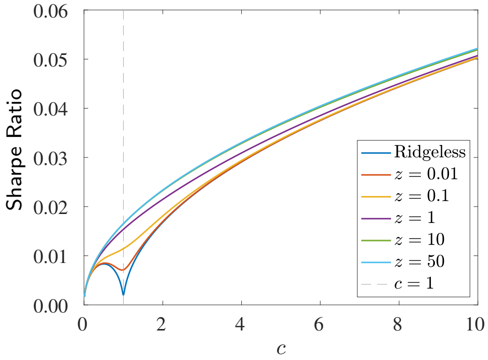
Why Do Big Models "Work"? The Trading Strategy Perspective



- Timing strategy: $R^{\pi}_{t+1} = \pi_t R_{t+1}, \quad \pi_t = \beta' S_t$

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 expected strategy returns due to better approximation of DGP

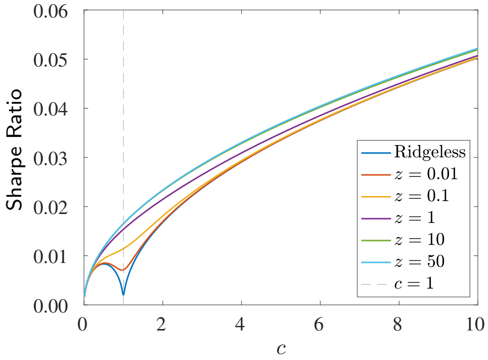
Why Do Big Models "Work"? The Trading Strategy Perspective



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

- There are 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
- Is the virtue of complexity **universal** in all assets?
- Analyze **US equities**, **international equities**, **futures**, and **bonds** markets

Empirical Analysis

US Equities

Forecast targets

- Monthly Fama/French 5 factors (1964–2021) and momentum (1927–2021) from Kenneth R. French Data Library
- 153 US equity factors from Jensen, Kelly, and Petersen (JKP, 2021), 1926–2020
- Trade monthly

Two Separate Information Sets

1. 15 predictor variables[†] from Welch and Goyal (WG, 2008)
2. Lag 1 and 2-12 factor momentum based on Gupta and Kelly (GK, 2019)

[†] 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

International Equities

Forecast targets

- Excess market returns of 93 countries from Jensen, Kelly, and Petersen (JKP, 2021), 1926–2020
- Trade monthly

Information Set

- Lag 1 and 2-12 factor momentum based on Gupta and Kelly (GK, 2019)

Empirical Analysis

Futures

Forecast targets

- Daily futures returns for 44 U.S. and international futures contracts traded on CME and CBOE exchanges, 1959–2021. The data is from the Stevens Continuous Futures (SCF) database feed
- The 44 futures contracts include 21 commodities, 9 equity indexes, 8 currencies, and 6 interest rates
- Trade weekly

Information Set

- The momentum: rolling average of daily returns from lag 1 to lag T
- Carry momentum: rolling average of carry from lag 1 to lag T, where the carry is computed according to Koijen, Moskowitz, Pedersen, and Vrugt (KMPV, 2018)
- The number of lags T is chosen as 21, 63, 126, 252 trading days

Empirical Analysis

Bonds

Forecast targets

- Average annual returns across maturity from Cochrane and Piazzesi (CP, 2005), 1952–2020.
- Trade annually

Information Set

- Log forward rates from Cochrane and Piazzesi (CP, 2005)

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 raw predictors. RFF converts G_t into

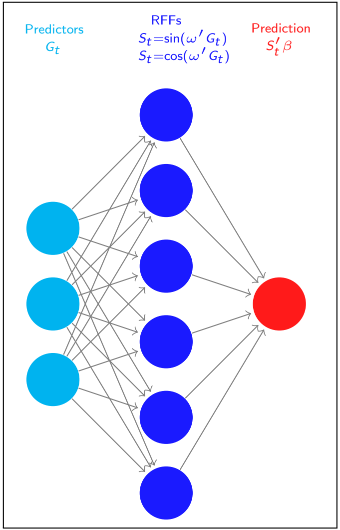
$$S_{i,t} = \left[\sin(\omega_i' G_t), \cos(\omega_i' G_t)\right]', \quad \omega_i \sim iidN(0, \gamma I)$$

  - $S_{i,t}$: 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
- Draw RFFs from a set of bandwidth parameter $\gamma = \{0.1, 0.5, 1, 2, 4, 8, 16\}$ and estimate with the **mixed RFFs**

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 raw predictors. RFF converts G_t into
$$S_{i,t} = \left[\sin(\omega_i' G_t), \cos(\omega_i' G_t)\right]', \quad \omega_i \sim iidN(0, \gamma I)$$
  - $S_{i,t}$: 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
- Draw RFFs from a set of bandwidth parameter $\gamma = \{0.1, 0.5, 1, 2, 4, 8, 16\}$ and estimate with the **mixed RFFs**
- 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

- 12-period 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

- 12-period 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 $P = 1, \ldots, 12{,}000$ and shrinkage of $\log_{10}(z) = -3, \ldots, 3$

Empirical Procedure

i. Generate 12,000 RFFs

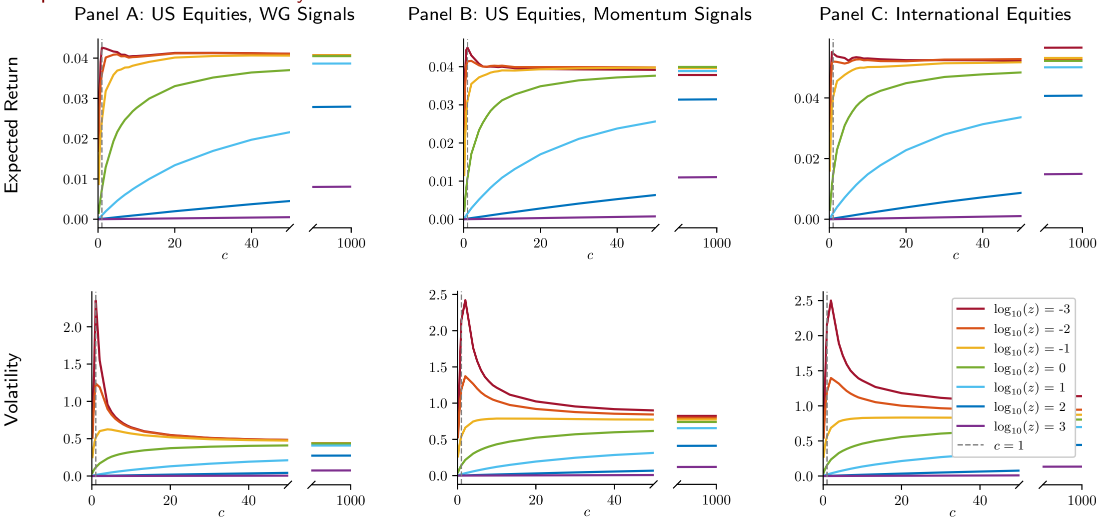
ii. Fix model defined by choice of (P, z)

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

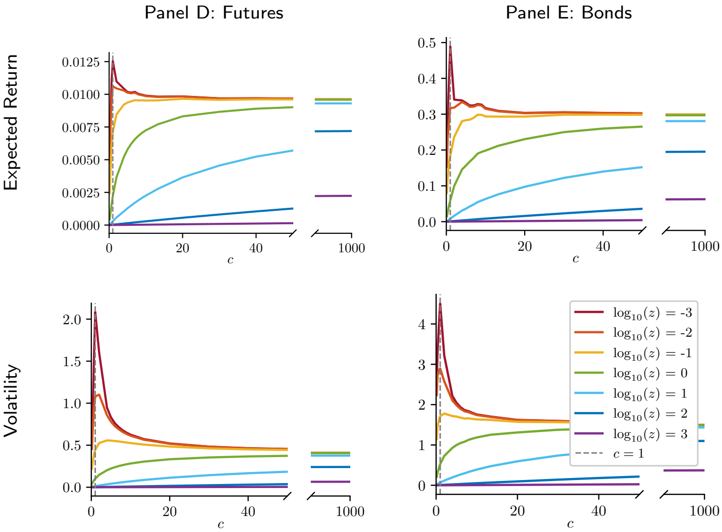
OOS Expected Return and Volatility



Out-of-sample Market Timing Performance

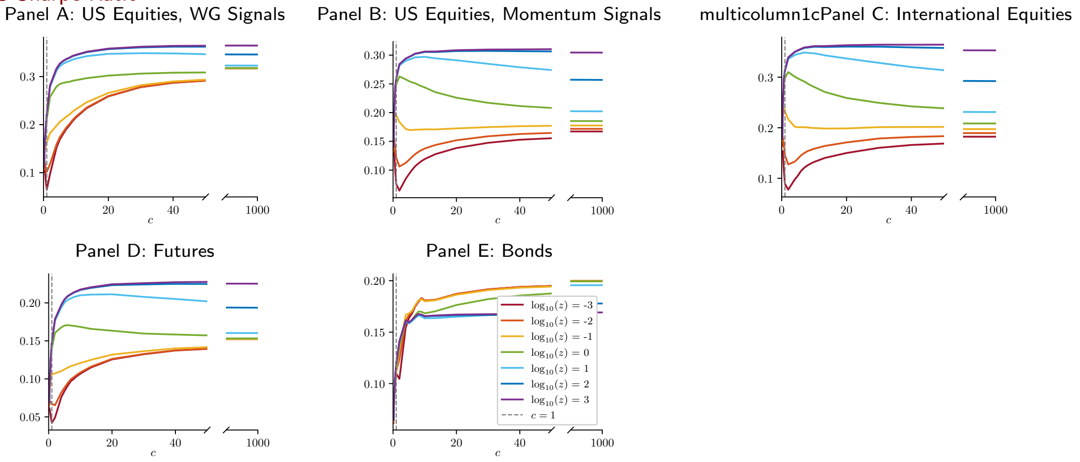
OOS Expected Return and Volatility

- 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



Out-of-sample Market Timing Performance

OOS Sharpe Ratio



- Increasing OOS ER and decreasing vol (for $c > 1$) $\Rightarrow$ increasing OOS SR
- Increasing pattern in OOS SR as complexity rises is **universal** for all the asset classes

Conclusions

- 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

- 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
  i. including all plausibly relevant predictors
  ii. 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

- 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
  - i. including all plausibly relevant predictors
  - ii. 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—prediction and timing—we find
  - The virtue of complexity is **universal**: SR increases with the model complexity in **US equities**, **international equities**, **futures**, and **bonds** markets
  - The high-complexity predictability mainly comes from nonlinear prediction effects

Conclusions

- 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

- 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