Complexity in the Factor Pricing Models

Antoine Didisheim
Uni. Melbourne

Barry Ke
Yale

Bryan Kelly
Yale

Semyon Malamud
EPFL

“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 (GPT-4 has, apparently, 1.76 trillion 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 (GPT-4 has, apparently, 1.76 trillion 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—that *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

Building the "Case" for Financial ML

- Finance lit: Rapid advances in return prediction/portfolio choice using ML
- Little theoretical understanding of why (and healthy skepticism)

"Virtue of Complexity in Return Prediction" (Kelly, Malamud, Zhou, forthcoming JF)

- Main theoretical result: Out-of-sample univariate timing strategy performance generally *increasing* in model complexity (# of parameters). Bigger models are better. Verified in data.

… And It's Happening In Finance Too

Building the "Case" for Financial ML

- Finance lit: Rapid advances in return prediction/portfolio choice using ML
- Little theoretical understanding of why (and healthy skepticism)

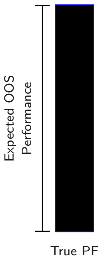
"Virtue of Complexity in Return Prediction" (Kelly, Malamud, Zhou, forthcoming JF)

- Main theoretical result: Out-of-sample univariate timing strategy performance generally *increasing* in model complexity (# of parameters). Bigger models are better. Verified in data.

This Paper: ML in Cross-sectional Asset Pricing

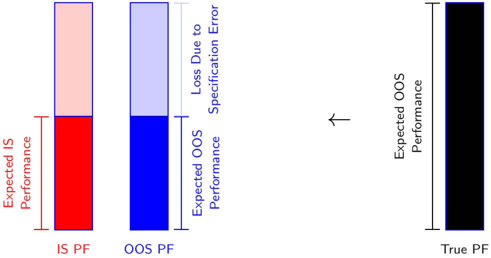
- Main theoretical result: PF performance generally *increasing* in model complexity
  - Higher portfolio Sharpe ratio
  - Smaller pricing errors
- Prior evidence of empirical gains from ML are *what we should expect*
- **Direct empirical support for theory**

Complexity in the Cross Section: Machine Learning Perspective



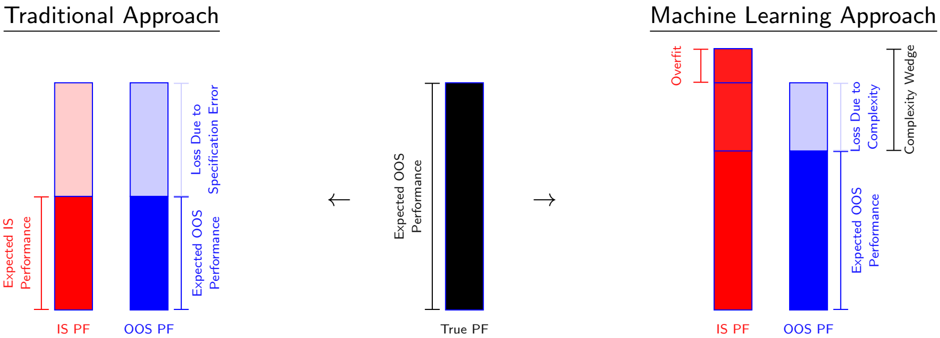
Complexity in the Cross Section: Machine Learning Perspective

Traditional Approach



- Restrict specification so $P/T \approx 0$
- Aligns IS and OOS performance
- May get lucky with spec, but can't be lucky on average
- Like shrinking *before seeing data*

Complexity in the Cross Section: Machine Learning Perspective



- Restrict specification so $P/T \approx 0$
- Aligns IS and OOS performance
- May get lucky with spec, but can't be lucky on average
- Like shrinking *before seeing data*

- $P/T \to \infty$ eliminates specification error
- IS overfit *improves* OOS performance
- Loss due to limits on learning (breakdown of LLN, high variance)
- Mitigate with shrinkage *after seeing data*

Intoduction to Asset Pricing I

- assets $i = 1, \cdots, N$ have prices $P_{i,t}$ and *excess* returns

$$R_{i,t+1} = \frac{P_{i,t+1} + D_{t+1}}{P_{i,t}} - \underbrace{R_{f,t}}_{risk \;\; free \;\; rate} \tag{1}$$

- if you invest fraction $\pi_{i,t}$ of your wealth W_t into security i, the rest stays on your bank account and grows at the rate $R_{f,t}$:

$$W_t = \sum_i \underbrace{\pi_{i,t} W_t}_{investment \;\; in \;\; stock \;\; i} + \underbrace{(W_t - \sum_i \pi_{i,t} W_t)}_{bank \;\; account} \tag{2}$$

Intoduction to Asset Pricing II

and then you sell your investments at time t and collect dividends so that

$$\begin{aligned} W_{t+1} &= \sum_i W_t \pi_{i,t} \frac{P_{i,t+1} + D_{t+1}}{P_{i,t}} + (W_t - \sum_i \pi_{i,t} W_t) R_{f,t} \\ &= W_t R_{f,t} + W_t \sum_i \pi_{i,t} R_{i,t+1} \end{aligned} \tag{3}$$

- Thus, the excess return on your wealth is

$$\frac{W_{t+1}}{W_t} - R_{f,t} = \sum_i \pi_{i,t} R_{i,t+1} = \pi_t' R_{t+1} \tag{4}$$

- Thus, we want π_t that gives good returns. But what is the criterion?

Intoduction to Asset Pricing III

- mean-variance optimization:

$$\pi_t \;=\; \arg\max_{\pi_t} \left(E_t[\pi_t' R_{t+1}] \;-\; 0.5 \underbrace{\gamma}_{risk\ \ aversion} E_t[(\pi_t' R_{t+1})^2] \right) \tag{5}$$

  and hence the Mean-Variance Efficient (MVE) portfolio is

$$\underbrace{\pi_t}_{tangency\ \ portfolio} \;=\; \gamma^{-1} \underbrace{(E_t[R_{t+1} R_{t+1}])^{-1}}_{N\times N\ \ covariance\ \ matrix} \; N\times 1 \; \underbrace{E_t[R_{t+1}]}_{expected\ \ returns} \tag{6}$$

- Now comes the big question: How do we measure the conditional expectations, $E_t[R_{t+1}]$ and $E_t[R_{t+1} R'_{t+1}]$?
- Once can start with a simple prediction problem: measure $E_t[R_{t+1}]$ by running a regression on observables (economic variables) S_t using past data (time series prediction)
- Virtue of Complexity in Return Prediction (Kelly, Malamud, and Zhou (2022):

Intoduction to Asset Pricing IV

- $$R_{t+1} = \sum_i \beta_i S_{i,t} + \varepsilon_{t+1} \tag{7}$$

  estimate

  $$\hat{\beta} = \left(\underbrace{z}_{ridge\ \ penalty} I + \frac{1}{T}\sum_t S_t S_t' \right)^{-1} \frac{1}{T}\sum_t S_t R_{t+1} \tag{8}$$

  with $S_t \in \mathbb{R}^P$ = vector of random features $S_t = f(X_t)$ and the prediction

  $$\pi_t = \hat{\beta}' S_t \tag{9}$$

- you want the strategy to work. Build a timing strategy

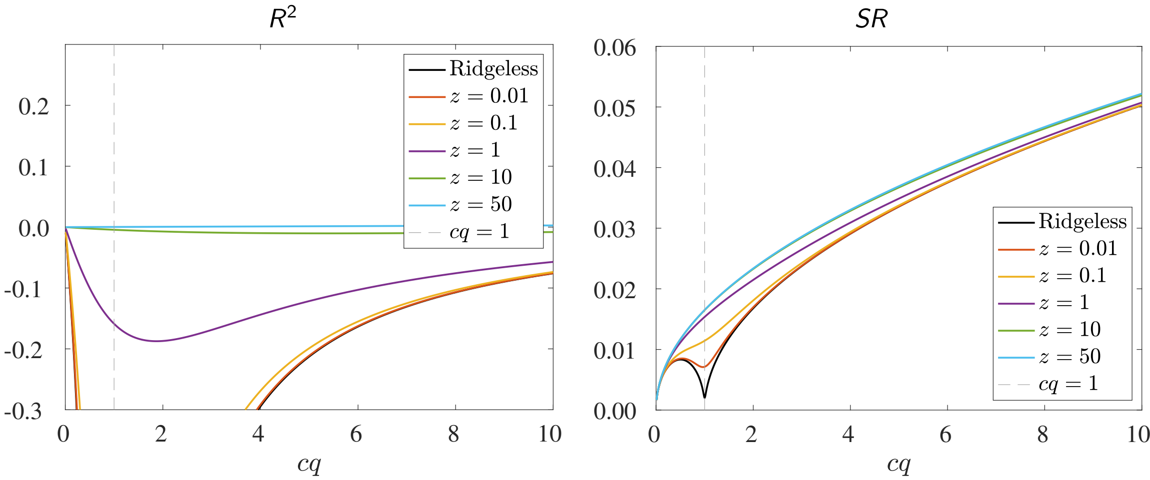
  $$R^{\pi}_{t+1} = \pi_t R_{t+1} \tag{10}$$

- complexity $c = P/T$, when $P > T$ we have overparametrization

Intoduction to Asset Pricing V

- Theorem: Virtue of complexity. Out-of-sample (OOS) performance monotone increasing in c for z_*=optimal shrinkage

There is no double acscent, only permanent ascent



Many Assets = Cross-Section

- N assets (stocks) with returns $R_{i,t+1},\ i = 1, \cdots, N$
- And characteristics $X_{i,t} \in \mathbb{R}^d$ (d characteristics per stock)
- So, we want to build the best portfolio π_t so that

$$\pi_t' R_{t+1} \;=\; \sum_{i=1}^{N} \underbrace{\pi_{i,t}}_{portfolio\ \ weight\ \ for\ \ stock\ \ i} R_{i,t+1} \tag{11}$$

  has a high Sharpe Ratio
- Each security i (e.g., a stock) comes with characteristics $X_{i,t} \in \mathbb{R}^d$, d is about $100 - 200$
- It is intuitive to search for

$$\pi_{i,t} \;=\; w(X_{i,t}) \tag{12}$$

- how do we find the function w ?

Complexity in the Cross Section: A Brief History I

- Standard solution: Restrict w
  - E.g., Fama-French:

$$X_{i,t} = (\text{Size}_{i,t}, \ \text{Value}_{i,t}) \tag{13}$$

    and **linear** $w_{i,t} = b_0 + b_1 \text{Size}_{i,t} + b_2 \text{Value}_{i,t}$
  - As a result

$$\begin{aligned} \pi_t' R_{t+1} &= \underbrace{\sum_{i=1}^{N} w_{i,t} R_{i,t+1}}_{\textit{sum over stocks}} \\ &= \sum_{i=1}^{N} (b_0 + b_1 \text{Size}_{i,t} + b_2 \text{Value}_{i,t}) R_{i,t+1} \\ &= b_0 \underbrace{\sum_{i=1}^{N} R_{i,t+1}}_{MKT} + b_1 \underbrace{\sum_{i=1}^{N} \text{Size}_{i,t} R_{i,t+1}}_{SMB} + b_2 \underbrace{\sum_{i=1}^{N} \text{Value}_{i,t} R_{i,t+1}}_{HML} \end{aligned} \tag{14}$$

Complexity in the Cross Section: A Brief History II

- **Factor Zoo:** d is large (Jensen, Kelly, and Pedersen (2022): $d \geq 153$)

$$
\begin{aligned}
F_{k,t+1} &= \underbrace{\sum_{i=1}^{N} X_{i,t}(k) R_{i,t+1}}_{Characteristics-Managed\ \ Portfolio} \\
\pi_t' R_{t+1} &= \sum_{k=1} \underbrace{\lambda_k}_{factor\ \ weight} F_{k,t+1} \\
w(X_{i,t}) &= \underbrace{\lambda' X_{i,t}}_{linear\ \ function} = \lambda_k X_{i,t}(k)
\end{aligned}
\tag{15}
$$

Complexity in the Cross Section: Machine Learning Perspective I

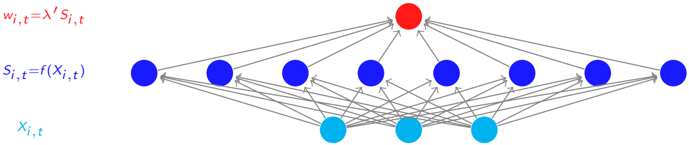
Rather than restricting $w(X_t)$....

- ...expand parameterization, saturate with conditioning information
- build many **non-linear** transformations: For a large P,

$$S_{i,t}(j) = \underbrace{f_j(X_{i,t})}_{nonlinear\ feature\ j\ for\ stock\ i}, \quad j = 1, \cdots, P \tag{16}$$

$X_t \rightarrow S_t$ **embedding** of $\mathbb{R}^d$ to $\mathbb{R}^P$.

Complexity in the Cross Section: Machine Learning Perspective II



- Implies that empirical PF is a high-dimensional factor model with factors F_{t+1} :

$$
\begin{aligned}
\pi_t' R_{t+1} &= \lambda' S_t' R_{t+1} \\
&= \sum_i (\lambda' S_{i,t} R_{i,t+1}) = \lambda' \underbrace{\sum_i S_{i,t} R_{i,t+1}}_{=F_{t+1} \in \mathbb{R}^{P\times 1}} = \lambda' \underbrace{F_{t+1}}_{vector\ \ of\ \ P\ \ factor\ \ returns}
\end{aligned}
\tag{18}
$$

Complexity in the Cross Section: Machine Learning Perspective

The Objective:

- Maximize out-of-sample Sharpe ratio

Complexity in the Cross Section: Machine Learning Perspective

The Objective:

- Maximize out-of-sample Sharpe ratio

The Choice:

- Fix T data points. Decide on "complexity" (number of factors P) to use in approximating model

Complexity in the Cross Section: Machine Learning Perspective

The Objective:

- Maximize out-of-sample Sharpe ratio

The Choice:

- Fix T data points. Decide on "complexity" (number of factors P) to use in approximating model

The Tradeoff:

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

Complexity in the Cross Section: Machine Learning Perspective

The Objective:

- Maximize out-of-sample Sharpe ratio

The Choice:

- Fix T data points. Decide on "complexity" (number of factors P) to use in approximating model

The Tradeoff:

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

The Central Research Question:

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

Complexity in the Cross Section: Machine Learning Perspective

The Objective:

- Maximize out-of-sample Sharpe ratio

The Choice:

- Fix T data points. Decide on "complexity" (number of factors P) to use in approximating model

The Tradeoff:

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

The Central Research Question:

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

Answer:

- Use the largest factor model (largest P) that you can compute

Theory Environment

Model

- n assets with returns R_{t+1}
- Empirical PF $M_{t+1} = 1 - \lambda' S_t' R_{t+1}$
  - Think of S_t as "generated features" in neural net with input X_t
  - $P \times 1$ vector of instruments, S_t (i.e., P factors F_{t+1})
- (Ridge-penalized) objective

$$\min_{\lambda} E[(1 - \lambda' S_t' R_{t+1})^2] + z\lambda'\lambda$$

Solution:

$$\hat{\lambda}(z) = \left(zI + \frac{1}{T} \sum_t F_t F_t' \right)^{-1} \frac{1}{T} \sum_t F_t$$

Theory Environment

Model

- n assets with returns R_{t+1}
- Empirical PF $M_{t+1} = 1 - \lambda' S_t' R_{t+1}$
  - Think of S_t as "generated features" in neural net with input X_t
  - $P \times 1$ vector of instruments, S_t (i.e., P factors F_{t+1})
- (Ridge-penalized) objective

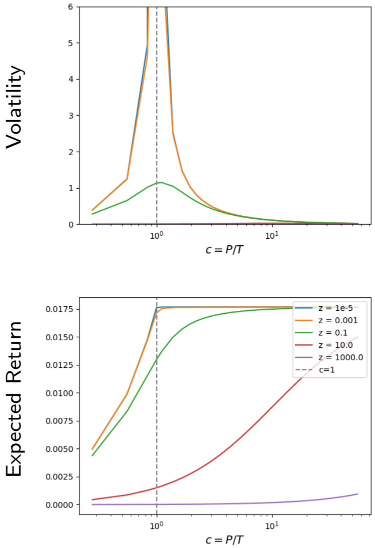
$$\min_{\lambda} E[(1 - \lambda' S_t' R_{t+1})^2] + z\lambda'\lambda$$

  Solution:

$$\hat{\lambda}(z) = \left(zI + \frac{1}{T} \sum_t F_t F_t' \right)^{-1} \frac{1}{T} \sum_t F_t$$

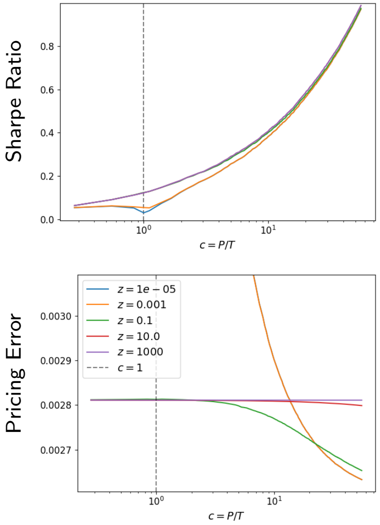
- Goal: Characterize **out-of-sample** behaviors, contrast **simple** (small P) models vs. **complex** models
- Tools: Joint limits as numbers of observations and parameters are large, $T, P \to \infty$, RMT

Complexity and the PF



1. PF variance
   - As $c \to 1$, λ variance blows up. A unique λ produces max SR, but it has high variance
   - When $c > 1$, variance *drops* with model complexity! Why?
   - Many λ's exactly fit training data, ridge selects one with small variance
2. PF expected returns
   - Low for $c \approx 0$ due to poor approximation of true model
   - Monotonically increases with model complexity

Complexity and the PF



Main theory result

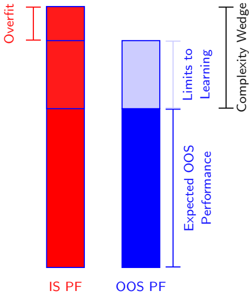
- Complexity is a virtue—biggest model wins
  - Approximation benefits dominate costs of heavy parameterization
  - For moderate complexity ($c \approx 1$), ridge shrinkage is beneficial
  - For high complexity ($c >> 1$), ridge shrinkage has small benefit (the important shrinkage is implicit)
- Paper provides general, rigorous theoretical statements and proofs that underlie plots
- Plots calculated from our theorems in a reasonable calibration

Complexity and the PF: Other Theoretical Results

1. "Complexity wedge" $=$ IS Performance – Expected OOS Performance

$$= \underbrace{\text{IS} - \text{True}}_{\text{"Overfit"}} + \underbrace{\text{True} - \text{OOS}}_{\text{"Limits to Learning"}}$$

- Quantifiable based on training data
- Can infer performance of true PF and how far you are from it, but cannot recover it!



2. Show how to infer optimal shrinkage, z^*, from training data
3. There is no low-rank rotation of complex factors that preserves model performance (cf. Kozak, Nagel, and Santosh, 2020)

Empirical Analysis

- Analyze empirical analogs to theoretical comparative statics
- Study conventional setting with conventional data
  - Forecast target is a monthly return of US stocks from CRSP 1963–2021
  - Conditioning info ($X_{i,t}$) is 130 stock characteristics from Jensen, Kelly, and Pedersen (2022)
- Out-of-sample performance metrics are:
  - PF Sharpe ratio
  - Mean squared pricing errors (factors as test assets)

Empirical Analysis

Random Fourier Features

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

Empirical Analysis

Random Fourier Features

- Empirical model: $\lambda' S_t' R_{t+1}$
- Need framework to smoothly transition from low to high complexity
- Adopt ML method known as "random Fourier features" (RFF)
  - Let $X_{i,t}$ be 130×1 predictors. RFF converts $X_{i,t}$ into

$$S_{\ell,i,t} = \sin(\gamma_\ell' X_{i,t}), \quad \gamma_\ell \sim iidN(0, \gamma I)$$

  - $S_{\ell,i,t}$: Random lin-combo of $X_{i,t}$ fed through non-linear activation

Empirical Analysis

Random Fourier Features

- Empirical model: $\lambda' S_t' R_{t+1}$
- Need framework to smoothly transition from low to high complexity
- Adopt ML method known as "random Fourier features" (RFF)
  - Let $X_{i,t}$ be 130×1 predictors. RFF converts $X_{i,t}$ into

$$S_{\ell,i,t} = \sin(\gamma_\ell' X_{i,t}), \quad \gamma_\ell \sim iidN(0, \gamma I)$$

  - $S_{\ell,i,t}$: Random lin-combo of $X_{i,t}$ fed through non-linear activation
- For fixed inputs can create an 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

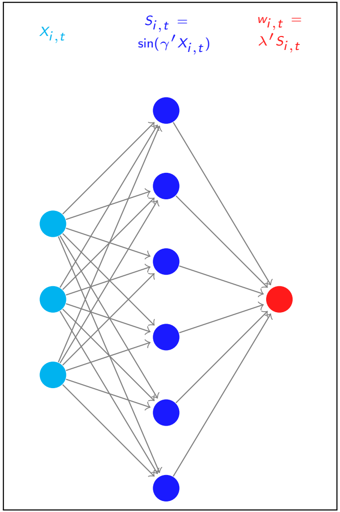
Empirical Analysis

Random Fourier Features

- Empirical model: $\lambda' S_t' R_{t+1}$
- Need framework to smoothly transition from low to high complexity
- Adopt ML method known as "random Fourier features" (RFF)
  - Let $X_{i,t}$ be 130×1 predictors. RFF converts $X_{i,t}$ into

$$S_{\ell,i,t} = \sin(\gamma_\ell' X_{i,t}), \quad \gamma_\ell \sim iidN(0, \gamma I)$$

  - $S_{\ell,i,t}$: Random lin-combo of $X_{i,t}$ fed through non-linear activation
- For fixed inputs can create an 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 a two-layer neural network with fixed weights (γ) in the first layer and optimized weights (λ) in the second layer

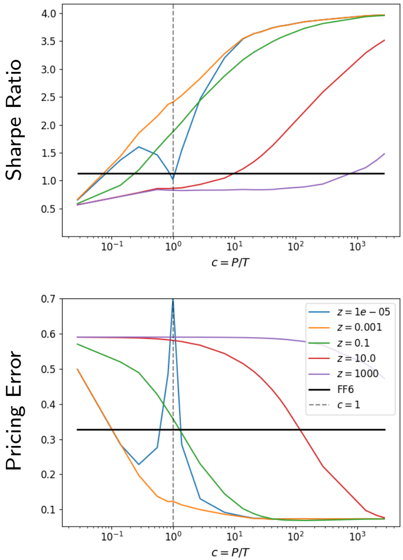


Empirical Analysis

Training and Testing

- We estimate out-of-sample PF with:
  i. Thirty-year rolling training window ($T = 360$)
  ii. Various shrinkage levels, $\log_{10}(z) = -12, ..., 3$
  iii. Various complexity levels $P = 10^2, ..., 10^6$
- For each level of complexity $c = P/T$, we plot
  i. Out-of-sample Sharpe ratio of the kernels and
  ii. Pricing errors on 10^6 "complex" factors: $F_{t+1} = S_t' R_{t+1}$
- Also report Sharpe ratio and pricing errors of FF6 to benchmark our results

Out-of-sample PF Performance

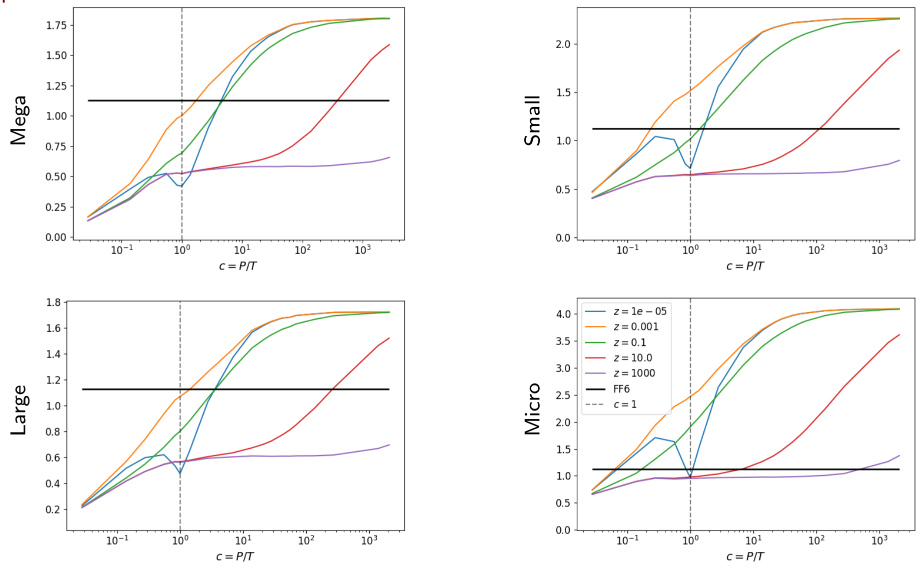


Main Empirical Result

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

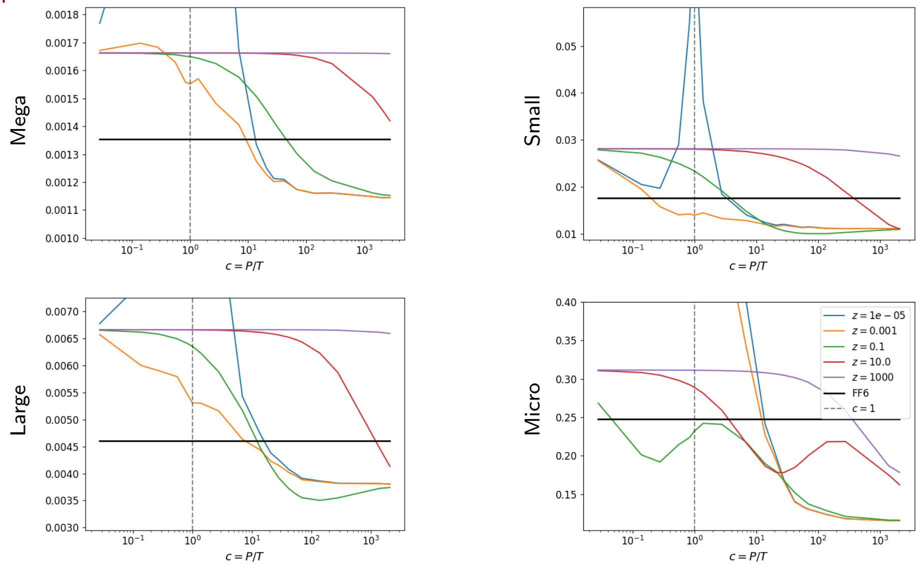
PF Performance in Restricted Samples: Sharpe Ratio

Market Capitalization Subsamples

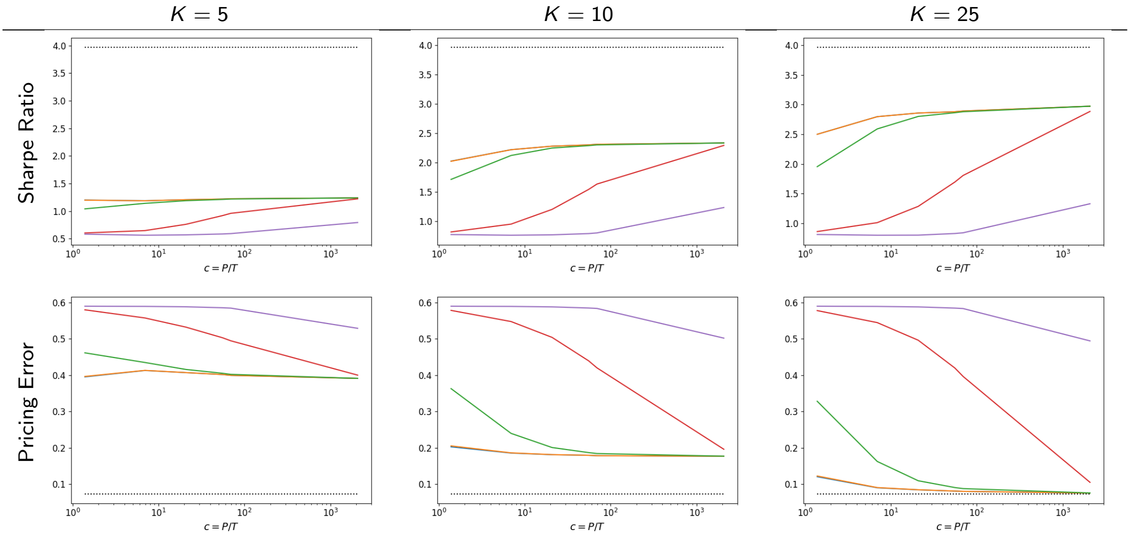


PF Performance in Restricted Samples: Pricing Errors

Market Capitalization Subsamples



What About "Shrinking" With PCA?



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
  - 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—pricing the cross section of returns—we find
  - OOS Sharpe rise by factor of 4 relative to FF6 model, pricing errors reduced by a factor of 3

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

Understanding The Theory

- Suppose $c = P/T \approx 0$. Then, we know

$$\lambda \;=\; E[FF']^{-1}E[F] \;=\; \frac{1}{1+MaxSR^2}\,\mathrm{Var}[F]^{-1}E[F]\,, \tag{19}$$

where we have defined

$$MaxSR^2 \;=\; E[F]'\,\mathrm{Var}[F]^{-1}E[F] \tag{20}$$

$$E[\lambda' F_{t+1}] \;=\; E[(\lambda' F_{t+1})^2] \;=\; E[F]'E[FF']^{-1}E[F] \;=\; \frac{MaxSR^2}{1+MaxSR^2} \tag{21}$$

Principal Components and Ridge I

- $\mathrm{Var}[F] \; = \; U \,\mathrm{diag}(\mu) U'$, and we can define PC_i to be the i-th column of $U'F$; and

$$\theta \; = \; U'E[F] \tag{22}$$

- $$R(PC_i) \; = \; PC_i' F_{t+1}$$

$$E[R(PC_i)] \; = \; \theta_i, \; \mathrm{Var}[R(PC_i)] \; = \; \mu_i, \; (SR(PC_i))^2 \; = \; \frac{\theta_i^2}{\mu_i}$$

  and

$$\begin{aligned} MaxSR^2 \; &= \; E[F]' \,\mathrm{Var}[FF']^{-1} E[F] \; = \; E[F]' U \,\mathrm{diag}(\mu^{-1}) U' E[F] \\ &= \; \theta' \,\mathrm{diag}(\mu^{-1}) \theta \; = \; \sum_i \frac{\theta_i^2}{\mu_i} \; = \; \sum_i (SR(PC_i))^2 \,. \end{aligned} \tag{23}$$

Principal Components and Ridge II

- Define

$$\lambda(z) \;=\; (zI + E[FF'])^{-1}E[F] \tag{24}$$

and

$$R^{infeasible}(z) \;=\; F'_{t+1}\lambda(z) \tag{25}$$

- The first moment is

$$\mathcal{R}_1^{infeas}(z) \;=\; E[R^{infeasible}(z)] \;=\; E[F]'(zI + E[FF'])^{-1}E[F] \;=\; \frac{A(z)}{1+A(z)} \tag{26}$$

where

$$A(z) \;=\; E[F]'(zI + \mathrm{Var}[F])^{-1}E[F] \;=\; \sum_i (SR(PC_i))^2 \frac{\mu_i}{\mu_i + z}\,. \tag{27}$$

- and

$$\mathcal{R}_2^{infeas}(z) \;=\; E[(R^{infeasible}(z))^2] \;=\; \frac{d}{dz}\left(\frac{zA(z)}{1+A(z)}\right)\,. \tag{28}$$

Principal Components and Ridge III

In this case,

$$SR^{infeas}(z) \;=\; \frac{\mathcal{R}_1^{infeas}(z)}{(\mathcal{R}_2^{infeas}(z))^{1/2}} \tag{29}$$

is monotone decreasing in z.

Random Matrix Theory and Implicit Regularization I

- When $c = P/T > 0$, estimating $E[FF']$ and $E[F]$ becomes infeasible and

$$\hat{\lambda}(z) = \left(zI + \frac{1}{T} \sum_t F_t F_t' \right)^{-1} \frac{1}{T} \sum_t F_{t+1} \;\not\approx\; (zI + E[FF'])^{-1} E[F] \tag{30}$$

because

$$B_T \;=\; \frac{1}{T} \sum_t F_t F_t' \;\not\approx\; E[FF'] \; and \; \bar{F}_T \;=\; \frac{1}{T} \sum_t F_{t+1} \;\not\approx\; E[F] \tag{31}$$

- Stieltjes transforms

$$\begin{aligned} m(-z) \;&=\; P^{-1} \operatorname{tr}((zI + \operatorname{Var}[FF'])^{-1}) \;=\; P^{-1} \sum_i (z + \mu_i)^{-1} \\ m(-z; c) \;&=\; P^{-1} \operatorname{tr}\left((zI + B_T)^{-1} \right) \end{aligned} \tag{32}$$

Random Matrix Theory and Implicit Regularization II

- $$\xi(z;c) \;=\; \frac{1}{T} F'_{T+1} \left(zI + B_T\right)^{-1} F_{T+1} \;\leq\; c\, z^{-1} \tag{33}$$
- The implicit shrinkage function
$$Z_*(z;c) \;=\; z\,(1 + \xi(z;c)) \tag{34}$$
- **Theorem** When $P \to \infty,\; P/T \to c$:
$$m(-z;c) \;=\; \frac{Z_*(z;c)}{z}\, m(-Z_*(z;c)) \tag{35}$$

Implicit Regularization and Expected Return

Recall that

$$\mathcal{R}_1^{infeas}(z) \;=\; E[R^{infeasible}(z)] \;=\; E[F]'(zI + E[FF'])^{-1}E[F] \;=\; \frac{A(z)}{1 + A(z)} \tag{36}$$

Our goal is to understand

$$\mathcal{R}_1(z; c) \;=\; E[\hat{\lambda}(z)' F_{t+1}] \tag{37}$$

where

$$\mathcal{R}_1^{infeas}(z) \;=\; \underbrace{\mathcal{R}_1(z; 0)}_{zero\;\;complexity} \tag{38}$$

Theorem When $P \to \infty,\;\; P/T \to c:$

$$\mathcal{R}_1(z; c) \;=\; \mathcal{R}_1^{infeas}(Z_*(z)) \tag{39}$$

The Risk Of Doing ML

Theorem Suppose that $E[F] = 0$. Then,

$$\lim_{P\to\infty,\ P/T\to c} E[R^F_{t+1}(z)] \;=\; 0. \tag{40}$$

Yet,

$$\lim_{P\to\infty,\ P/T\to c} E[(R^F_{t+1}(z))^2] \;=\; G(z;c) \;>\; 0\,, \tag{41}$$

where

$$G(z;c) \;=\; \lim_{T\to\infty, P/T\to c} \frac{1}{T} E[(F'_{t_1}(zI + B_T)^{-1} F_{t_2})^2] \tag{42}$$

for any $t_1 \neq t_2$ is given by

$$G(z;c) \;=\; (\xi(z;c)(1+\xi(z;c)) \;+\; z\xi'(z;c) + (\xi(z;c))^2)/(1+\xi(z;c))^2. \tag{43}$$

In particular, $G(z;c)$ is monotone decreasing in z and increasing in c.

Where Does The Risk Of Doing ML Come From?

To understand how the big data regime produces this intrinsic noise, consider a simple portfolio strategy that invests proportionally to the historical mean returns:

$$R_{t+1}^{M} \;=\; \bar{F}_T'\, F_{T+1}\,. \tag{44}$$

Then,

$$E[R_{t+1}^{M}] \;=\; E[\bar{F}_T'\, F_{T+1}] \;=\; E[\bar{F}_T]\, E[F_{T+1}] \;=\; 0\,, \tag{45}$$

under the assumption that $E[F] = 0$. Yet,

$$\begin{aligned} E[(R_{t+1}^{M})^2] \;&=\; E[(\bar{F}_T'\, F_{T+1})^2] \;=\; \operatorname{tr} E[\bar{F}_T \bar{F}_T' F_{T+1} F_{T+1}'] \\ &=\; \operatorname{tr} E[\bar{F}_T \bar{F}_T' \Psi] \;=\; \frac{1}{T^2} \sum_t \operatorname{tr} E[F_t F_t' \Psi] \;=\; \frac{1}{T} \operatorname{tr}(\Psi^2) \end{aligned} \tag{46}$$

If, for example, $\Psi = I$, this quantity equals $P/T \to c$. Thus, many minor estimation errors accumulate and generate non-trivial risk for the portfolio.

The Second Moment

Theorem

We have

$$E[(R^F_{T+1}(z))^2] \;\to\; \underbrace{\mathcal{R}_2^{infeas}(Z^*(z;c))}_{implicit\;\; regularization} \;+\; \underbrace{G(z;c)(1-2\mathcal{R}_1^{infeas}(Z^*(z;c))+\mathcal{R}_2^{infeas}(Z^*(z;c)))}_{estimation\;\; risk}, \tag{47}$$

where

$$\mathcal{R}_2^{infeas}(z) = \mathcal{R}_2(z;0) \;=\; \frac{d}{dz}\left(\frac{zA(z)}{1+A(z)}\right) \tag{48}$$

is the second moment of the return on the infeasible portfolio, $F'_{T+1}(\mathbf{z}I+E[FF'])^{-1}E[F]$, estimated using $T=\infty$.