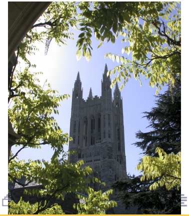
Structural Deep Learning in Conditional Asset Pricing

Jianqing Fan

Princeton University

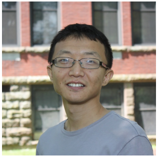
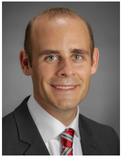
https://fan.princeton.edu/

with **Tracy Ke**, **Yuan Liao**, **Andreas Neuhierl**



Outlines

1. Introduction
2. Interpretable Asset Pricing model
3. Methods of Estimation
4. Asymptotic Theory
5. Empirical Application

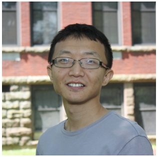
  

Tracy Ke Yuan Liao Andreas Neuhierl

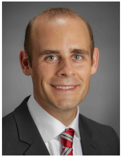
Outlines

1. Introduction
2. Interpretable Asset Pricing model
3. Methods of Estimation
4. Asymptotic Theory
5. Empirical Application

Tracy Ke



Yuan Liao



Andreas Neuhierl

Introduction

Motivation

- Returns are noisy
- Understanding **expected returns** and **volatility** is all we can do
- Understanding **expected returns** is still **really hard**!

Why so hard?

★ **High-dim** set of possible signals (firm characteristics).

★ Theory is silent of how **signals relate** to returns

- Unknown functional forms
- Unknown dynamics (risks can change over time)
- Unknown interpretation (risk vs. mispricing)

★ Curse of dimensionality of nonparametric modeling

Why so hard?

★ **High-dim** set of possible signals (firm characteristics).

★ Theory is silent of how **signals relate** to returns

- Unknown functional forms
- Unknown dynamics (risks can change over time)
- Unknown interpretation (risk vs. mispricing)

★ Curse of dimensionality of nonparametric modeling

Why so hard?

★ **High-dim** set of possible signals (firm characteristics).

★ Theory is silent of how **signals relate** to returns

- Unknown functional forms
- Unknown dynamics (risks can change over time)
- Unknown interpretation (risk vs. mispricing)

★ Curse of dimensionality of nonparametric modeling

Why so hard?

- ★ **High-dim** set of possible signals (firm characteristics).
- ★ Theory is silent of how **signals relate** to returns
  - Unknown functional forms
  - Unknown dynamics (risks can change over time)
  - Unknown interpretation (risk vs. mispricing)
- ★ Curse of dimensionality of nonparametric modeling

Classical Solutions

★ **Additive model**: $E[y|x_1, ..., x_d] = \sum_{i=1}^{d} f_i(x_i)$ **rate**: $N^{-2/5}$

★ **Single index models** $E[y|\mathrm{X}] = f(\mathrm{X}^T \beta)$ **rate**: $N^{-2/5}$

★ Structural rather than algorithmic solutions.

■Statistical Machine Learning such as **Neural networks** come to rescue

Classical Solutions

★ **Additive model**: $E[y|x_1, \ldots, x_d] = \sum_{i=1}^{d} f_i(x_i)$ **rate**: $N^{-2/5}$

★ **Single index models** $E[y|\mathrm{X}] = f(\mathrm{X}^T \boldsymbol{\beta})$ **rate**: $N^{-2/5}$

★ Structural rather than algorithmic solutions.

■Statistical Machine Learning such as **Neural networks** come to rescue

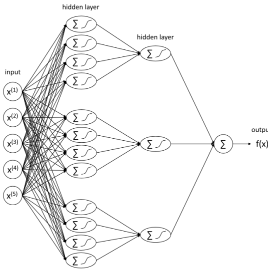
Why Deep Neural Networks?

★ **Univ. approx**: any sup-smooth function can be appr. by an NN (*Barron, 93*):

$$f(x) \overset{\text{Inv Fourier}}{=} \int_{R^d} \exp(i\omega^T x)\tilde{f}(\omega)d\omega = E_{\omega\sim g}\exp(i\omega^T x)\frac{\tilde{f}(\omega)}{g(\omega)}$$

$$\approx N^{-1}\sum_{j=1}^{N}\underbrace{\exp(i\omega_j^T x)}_{\sigma(\omega_j^T x)}\underbrace{\frac{\tilde{f}(\omega_j)}{g(\omega_j)}}_{weight\ \beta_j} + \underbrace{O_P(N^{-1/2})}_{\text{no curse-of-dim.}}, \quad \text{if } E_{\omega\sim g}\left|\frac{\tilde{f}(\omega)}{g(\omega)}\right|^2 < C.$$

★ A tooth function w/ $\mathcal{O}(2^k)$ oscillations (Telgarsky, 16) is
- •a ReLU-DNN with depth $\mathcal{O}(k)$ and width $\mathcal{O}(1)$,
- •a one-layer ReLU-DNN but with $\Omega(2^k)$ nodes.
- •extendable to Lipchitz functions.



From Bauer and Kohler, 2019, AOS

★ **Adapt to unknown comp**: Compositions of DNN is a deeper DNN. •No **COD** in implementation.

Why Deep Neural Networks?

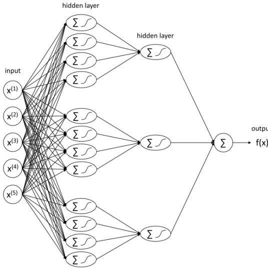
★ **Univ. approx**: any sup-smooth function can be appr. by an NN (*Barron, 93*):

$$f(x) \overset{\text{Inv Fourier}}{=} \int_{R^d} \exp(i\omega^T x)\tilde{f}(\omega)d\omega = E_{\omega\sim g} \exp(i\omega^T x)\frac{\tilde{f}(\omega)}{g(\omega)}$$

$$\approx N^{-1}\sum_{j=1}^{N} \underbrace{\exp(i\omega_j^T x)}_{\sigma(\omega_j^T x)} \underbrace{\frac{\tilde{f}(\omega_j)}{g(\omega_j)}}_{weight\ \beta_j} + \underbrace{O_P(N^{-1/2})}_{\text{no curse-of-dim.}}, \quad \text{if } E_{\omega\sim g}\left|\frac{\tilde{f}(\omega)}{g(\omega)}\right|^2 < C.$$

★ A tooth function w/ $\mathcal{O}(2^k)$ oscillations (Telgarsky, 16) is

- •a ReLU-DNN with depth $\mathcal{O}(k)$ and width $\mathcal{O}(1)$,
- •a one-layer ReLU-DNN but with $\Omega(2^k)$ nodes.
- •extendable to Lipchitz functions.



From Bauer and Kohler, 2019, AOS

★ **Adapt to unknown comp**: Compositions of DNN is a deeper DNN. •No **COD** in implementation.

Why Deep Neural Networks?

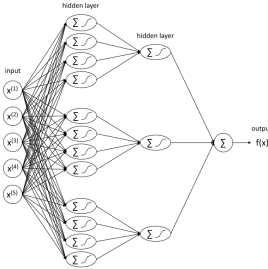
★ **Univ. approx**: any sup-smooth function can be appr. by an NN (*Barron, 93*):

$$f(x) \overset{\text{Inv Fourier}}{=} \int_{R^d} \exp(i\omega^T x)\tilde{f}(\omega)d\omega = E_{\omega\sim g}\exp(i\omega^T x)\frac{\tilde{f}(\omega)}{g(\omega)}$$

$$\approx N^{-1}\sum_{j=1}^{N}\underbrace{\exp(i\omega_j^T x)}_{\sigma(\omega_j^T x)}\underbrace{\frac{\tilde{f}(\omega_j)}{g(\omega_j)}}_{\text{weight } \beta_j} + \underbrace{O_P(N^{-1/2})}_{\text{no curse-of-dim.}}, \quad \text{if } E_{\omega\sim g}\left|\frac{\tilde{f}(\omega)}{g(\omega)}\right|^2 < C.$$

★ A tooth function w/ $\mathcal{O}(2^k)$ oscillations (Telgarsky, 16) is

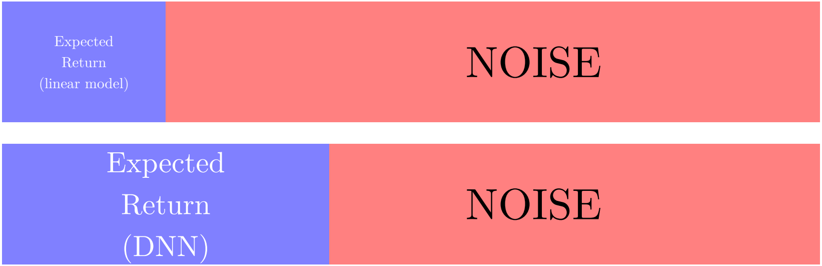
- a ReLU-DNN with depth $\mathcal{O}(k)$ and width $\mathcal{O}(1)$,
- a one-layer ReLU-DNN but with $\Omega(2^k)$ nodes.
- extendable to Lipchitz functions.



From Bauer and Kohler, 2019, AOS

★ **Adapt to unknown comp**: Compositions of DNN is a deeper DNN. • No **COD** in implementation.

What neural networks can do



- ML can explain more expected returns
- Produce good estimates of

$$\hat{y}_{it} = \mathbb{E}[\text{excess return}|\text{high dimensional info}]$$

Understanding source of returns

Expected Return (DNN)

Risk Premium	Factor Component	Mispricing

- Provide an interpretation of $\widehat{y}_{i,t+1|t}$
- Decomposition into **compensation for risk** and **mispricing**
- Evolution of return components over time

What is this paper?

★ Impose mild **economic structure**

★ Expected returns from **deep neural network**

★ **In-sample** Expected return: mispricing + risk premium + factor realization

★ **Out-of-sample** prediction: mispricing + risk premium + factor innovation

★ Provide novel methods to estimate each of 3 component

★ Derive rigorous asymptotic theory

★ Factor accounts 95% (riskP: 1/5) and mispricing 5% (Sharpe: 1.23, $\downarrow$ in t)

★ Predictions in asset pricing are very noisy due to factor realizations;
•should be removed for predictions.

What is this paper?

- ★ Impose mild **economic structure**
- ★ Expected returns from **deep neural network**
- ★ **In-sample** Expected return: mispricing + risk premium + factor realization
- ★ **Out-of-sample** prediction: mispricing + risk premium + factor innovation
- ★ Provide novel methods to estimate each of 3 component
- ★ Derive rigorous asymptotic theory
- ★ Factor accounts 95% (riskP: 1/5) and mispricing 5% (Sharpe: 1.23, $\downarrow$ in t)
- ★ Predictions in asset pricing are very noisy due to factor realizations;
  - should be removed for predictions.

What is this paper?

★ Impose mild **economic structure**

★ Expected returns from **deep neural network**

★ **In-sample** Expected return: mispricing + risk premium + factor realization

★ **Out-of-sample** prediction: mispricing + risk premium + factor innovation

★ Provide novel methods to estimate each of 3 component

★ Derive rigorous asymptotic theory

★ Factor accounts 95% (riskP: 1/5) and mispricing 5% (Sharpe: 1.23, ↓ in t)

★ Predictions in asset pricing are very noisy due to factor realizations;
- should be removed for predictions.

What is this paper?

- ★ Impose mild **economic structure**
- ★ Expected returns from **deep neural network**
- ★ **In-sample** Expected return: mispricing + risk premium + factor realization
- ★ **Out-of-sample** prediction: mispricing + risk premium + factor innovation
- ★ Provide novel methods to estimate each of 3 component
- ★ Derive rigorous asymptotic theory
- ★ Factor accounts 95% (riskP: 1/5) and mispricing 5% (Sharpe: 1.23, $\downarrow$ in t)
- ★ Predictions in asset pricing are very noisy due to factor realizations;
  - should be removed for predictions.

Related Literature

★ Machine Learning in Finance

- Freyberger et al. (2020), Gu et al. (2020) Bali et al. (2021), Bianchi et al. (2021), , Avramov et al. (2021), Chen et al. (2020), Cong et al. (2021), DeMiguel et al. (2020), Bryzgalova et al. (2020), Rossi and Utkus (2020), Li and Rossi (2020)

★ (Conditional) Factor pricing models

- Giglio and Xiu (2021); Giglio et al. (2021); Kim et al. (2021)
- Shanken (1990); Ferson and Harvey (1999); Lettau and Ludvigson (2001); Ghysels (1998); Gagliardini et al. (2016); Kelly et al. (2019, 2020); Gu et al. (2019)

★ Panel data / factor estimation

- Connor and Korajczyk (1986); Bai (2003); Stock and Watson (2002b); Connor et al. (2012); Fan et al. (2016)

Interpretable Asset Model

Conditional factor model:

$$y_{i,t} = \alpha_{i,t-1} + \boldsymbol{\beta}'_{i,t-1}\boldsymbol{\lambda}_t + \boldsymbol{\beta}'_{i,t-1}(\mathrm{f}_t - \mathbb{E}\mathrm{f}_t) + u_{it}$$

- α – potential mispricing
- β – risk exposure
- $\boldsymbol{\lambda}_t$ – factor risk premia
- f_t – factor realization

■all **time-varying**

Impact of Firm Characteristics

■ $\mathrm{x}_{i,t}$ carry information about alphas and betas:

$$\begin{aligned}\alpha_{i,t-1} &= g_{\alpha,t}(\mathrm{x}_{i,t-1}) + \gamma_{\alpha,i,t-1}\\ \boldsymbol{\beta}_{i,t-1} &= \mathrm{g}_{\beta,\mathrm{t}}(\mathrm{x}_{\mathrm{i,t-1}}) + \boldsymbol{\gamma}_{\beta,i,t-1}\end{aligned}$$

- $g_{\alpha,t}$ – mispricing function
- $g_{\beta,t}$ – factor loading function
- $g_{\alpha,t}(\cdot)$ and $g_{\beta,t}(\cdot)$ can change over time
- α and $\boldsymbol{\beta}$ may vary in high-frequency due to γ_t

Structural machine learning predictions

★ Obtain $\widehat{m}_t(\cdot)$ by DNN on $\{(y_{i,t}, \mathrm{x}_{i,t-1})\}_{i=1}^N$ for each t, estimating

$$\mathrm{m}_\mathrm{t}^0(\mathrm{x}_{\mathrm{i,t-1}}) = \mathbb{E}(\mathrm{y}_{\mathrm{i,t}} | \mathrm{x}_{\mathrm{i,t-1}}, \mathrm{f}_\mathrm{t}), \qquad y_{i,t} = m_t^0(\mathrm{x}_{i,t-1}) + e_{i,t}.$$

★ Predict out-of-sample by "pluging in new data"

$$\widehat{y}_{i,t+1|t} = \widehat{m}_t(\mathrm{x}_{i,t})$$

★ Little interpretation about source of predictability.

Aim to open the black box

Structural machine learning predictions

★ Obtain $\widehat{m}_t(\cdot)$ by DNN on $\{(y_{i,t}, \mathrm{x}_{i,t-1})\}_{i=1}^N$ for each t, estimating

$$\mathrm{m}_t^0(\mathrm{x}_{i,t-1}) = \mathbb{E}(\mathrm{y}_{i,t} | \mathrm{x}_{i,t-1}, \mathrm{f}_t), \qquad y_{i,t} = m_t^0(\mathrm{x}_{i,t-1}) + e_{i,t}.$$

★ Predict out-of-sample by "pluging in new data"

$$\widehat{y}_{i,t+1|t} = \widehat{m}_t(\mathrm{x}_{i,t})$$

★ Little interpretation about source of predictability.

Aim to open the black box

Interpretation

★ **In-sample decomposition**: **Spot** expected return

$$\widehat{m}_t(\mathrm{x}_{i,t-1}) \approx \underbrace{g_{\alpha,t}(\mathrm{x}_{i,t-1})}_{\text{mispricing}} + \underbrace{g_{riskP,t}(\mathrm{x}_{i,t-1})}_{g_{\beta,t}(\mathrm{x})'\boldsymbol{\lambda}_t} + g_{\beta,t}(\mathrm{x}_{i,t-1})'(\mathrm{f}_t - \mathbb{E}\mathrm{f}_t)$$

—Enable us to asset contributions of each component.

- **Out-of-sample decomposition:** plug-in the "new" $\mathrm{x}_{i,t}$:

$$\widehat{m}_t(\mathrm{x}_{i,t}) \approx g_{\alpha,t}(\mathrm{x}_{i,t}) + g_{riskP,t}(\mathrm{x}_{i,t}) + \underbrace{g_{\beta,t}(\mathrm{x}_{i,t})'(\mathrm{f}_t - \mathbb{E}\mathrm{f}_t)}_{\text{factor realization}}$$

$$y_{i,t+1} \approx g_{\alpha,t}(\mathrm{x}_{i,t+1}) + g_{riskP,t}(\mathrm{x}_{i,t}) + \underbrace{g_{\beta,t}(\mathrm{x}_{i,t})'(\mathrm{f}_{t+1} - \mathbb{E}\mathrm{f}_{t+1})}_{\text{factor innovation}} + e_{i,t+1}$$

—Enable us to understand/improve predictability.

Interpretation

★ **In-sample decomposition**: **Spot** expected return

$$\widehat{m}_t(\mathrm{x}_{i,t-1}) \approx \underbrace{g_{\alpha,t}(\mathrm{x}_{i,t-1})}_{\text{mispricing}} + \underbrace{g_{riskP,t}(\mathrm{x}_{i,t-1})}_{g_{\beta,t}(\mathrm{x})'\boldsymbol{\lambda}_t} + g_{\beta,t}(\mathrm{x}_{i,t-1})'(\mathrm{f}_t - \mathbb{E}\mathrm{f}_t)$$

—Enable us to asset contributions of each component.

- **Out-of-sample decomposition:** plug-in the "new" $\mathrm{x}_{i,t}$:

$$\begin{aligned} \widehat{m}_t(\mathrm{x}_{i,t}) &\approx g_{\alpha,t}(\mathrm{x}_{i,t}) + g_{riskP,t}(\mathrm{x}_{i,t}) + \underbrace{g_{\beta,t}(\mathrm{x}_{i,t})'(\mathrm{f}_t - \mathbb{E}\mathrm{f}_t)}_{\text{factor } \textbf{realization}} \\ y_{i,t+1} &\approx g_{\alpha,t}(\mathrm{x}_{i,t+1}) + g_{riskP,t}(\mathrm{x}_{i,t}) + \underbrace{g_{\beta,t}(\mathrm{x}_{i,t})'(\mathrm{f}_{t+1} - \mathbb{E}\mathrm{f}_{t+1})}_{\text{factor } \textbf{innovation}} + e_{i,t+1} \end{aligned}$$

—Enable us to understand/improve predictability.

Methodology

Building block

$$
\begin{aligned}
& \mathbb{E}(Y_s|X_{s-1}, f_s) - \mathbb{E}\left(\mathbb{E}(Y_s|X_{s-1}, f_s)\middle| X_{s-1}\right) \\
\approx\quad & G_{\beta,t}(X_{t-1})(f_s - \mathbb{E}f_s), \qquad \text{for } s \approx t.
\end{aligned}
$$

■ **Locally**, $G_{\beta,t}(X_{t-1})$ is the eigenvector.

Estimation Steps:

1. Apply DNN **cross-sectionally** to estimate $\mathbb{E}(Y_s|X_{s-1}, f_s)$.
2. Apply **time-domain** kernel smoothing to estimate $\mathbb{E}(Y_s|X_{s-1})$.
3. Apply **local PCA** to estimate $G_{\beta,t}(X_{t-1})$ and f_t

Building block

$$
\begin{aligned}
& \mathbb{E}(Y_s | X_{s-1}, f_s) - \mathbb{E}\left(\mathbb{E}(Y_s | X_{s-1}, f_s) \middle| X_{s-1}\right) \\
\approx \quad & G_{\beta,t}(X_{t-1})(f_s - \mathbb{E}f_s), \qquad \text{for } s \approx t.
\end{aligned}
$$

■**Locally**, $G_{\beta,t}(X_{t-1})$ is the eigenvector.

Estimation Steps:

1. Apply DNN **cross-sectionally** to estimate $\mathbb{E}(Y_s | X_{s-1}, f_s)$.
2. Apply **time-domain** kernel smoothing to estimate $\mathbb{E}(Y_s | X_{s-1})$.
3. Apply **local PCA** to estimate $G_{\beta,t}(X_{t-1})$ and f_t

Step 1 – Cross sectional DNN

■ Apply period-by-period DNN

$$\widehat{m}_t = \arg \min_{m \in DNN} \sum_{i=1}^{N} (y_{i,t} - m(\mathrm{x}_{i,t-1}))^2$$

1. Captures nonlinearity
2. adapts to low-dim structure, avoiding curse of dim.
3. behavior is insensitive to tuning

Double-Descent Risk: miracle of overfitting

★ We use period-by-period DNN, so it can be overfitting.

★ DNN performs well in the overfitting regime.

★ Pred risks admit a **double descent curve** as complexity increases.

(Belkin et al., 2020; Mei and Montanari, 2019; Belkin et al., 2020; Hastie et al., 2019)

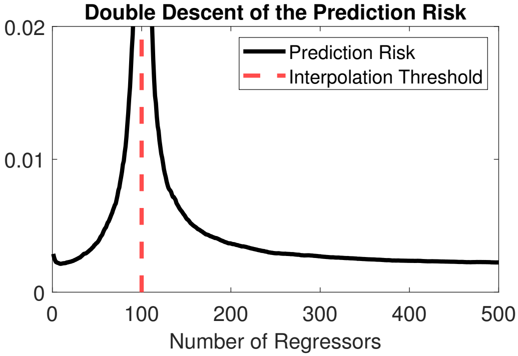
★ Illustrate using factor regression (Stock and Watson (2002a)).

$$R_t = b' f_t + e_t, \quad X_{t,p} = B' f_t + u_t$$

predicted by overparatermized linear model

Illustration of Double Decent

- Simulation DGP is calibrated using actual monthly returns.
- Evaluate out-sample risk $Risk(p) = \frac{1}{25}\sum_{s=1}^{25}(R_{T+s} - X'_{T+s,p}\widehat{\theta}_p)^2$.



- Interpretations: **diversification effect**

Step 2 - Time-domain Kernel Smoothing

- With bandwidth $h \to 0$, local kernel weight is

$$w_{s,t} = \frac{1}{h} K\left(\frac{s-t}{Th}\right) \Bigg/ \frac{1}{Th} \sum_{k=1}^{T} \frac{1}{h} K\left(\frac{k-t}{Th}\right).$$

- Kernel smoothing:

$$\widehat{E(y_{it}|\mathrm{x}_{i,t-1})} = \frac{1}{T} \sum_{s=1}^{T} \underbrace{\widehat{\mathbb{E}(y_{is}|\mathrm{x}_{i,s-1}, \mathrm{f}_s)}}_{\text{from DNN}} w_{s,t}$$

$$\approx g_{\alpha,t}(\mathrm{x}_{i,t-1}) + g_{riskP,t}(\mathrm{x}_{i,t-1}).$$

Step 3 - Local PCA

$$\mathbb{E}(y_{is}|\mathrm{x}_{i,s-1},\mathrm{f}_s) - \mathbb{E}(y_{is}|\mathrm{x}_{i,s-1}) \approx g_{\beta,t}(\mathrm{x}_{i,t-1})'(\mathrm{f}_s - \mathbb{E}\mathrm{f}_s), \quad \forall \frac{s}{T} \approx \frac{t}{T}$$

- At each t, estimate $g_{\beta,t}(\mathrm{x}_{i,t-1})$ as the eigenvectors of

$$\frac{1}{T}\sum_{s=1}^{T} \Delta_s \Delta_s' w_{s,t}, \qquad \Delta_s = \mathbb{E}(\widehat{y_{is}|\mathrm{x}_{i,s}}_{-1}) - \mathbb{E}(y_{is}\widehat{|\mathrm{x}_{i,s-1}},\mathrm{f}_s)$$

and $\widehat{\mathrm{f}}_t = \widehat{\mathrm{G}}'_{\beta,t-1}(\widehat{\mathrm{m}}_t(\mathrm{X}_{t-1}) - \bar{\mathrm{m}}_t)$.

- Conventional Fama-MacBeth type regression

$$\widehat{\lambda}_t = \frac{1}{N}\widehat{\mathrm{G}}'_{\beta,t-1}\bar{\mathrm{m}}_t, \quad \widehat{\mathrm{G}}_{\alpha,t-1} := \bar{\mathrm{m}}_t - \widehat{\mathrm{G}}_{\beta,t-1}\widehat{\lambda}_t.$$

Step 3 - Local PCA

$$\mathbb{E}(y_{is}|\mathrm{x}_{i,s-1}, \mathrm{f}_s) - \mathbb{E}(y_{is}|\mathrm{x}_{i,s-1}) \quad \approx \quad g_{\beta,t}(\mathrm{x}_{i,t-1})'(\mathrm{f}_s - \mathbb{E}\mathrm{f}_s), \qquad \forall \frac{s}{T} \approx \frac{t}{T}$$

- At each t, estimate $g_{\beta,t}(\mathrm{x}_{i,t-1})$ as the eigenvectors of

$$\frac{1}{T}\sum_{s=1}^{T} \Delta_s \Delta_s' w_{s,t}, \qquad \Delta_s = \mathbb{E}(\widehat{y_{is}|\mathrm{x}_{i,s-1}}) - \mathbb{E}(y_{is}\widehat{|\mathrm{x}_{i,s-1}}, \mathrm{f}_s)$$

  and $\widehat{\mathrm{f}}_t = \widehat{\mathrm{G}}'_{\beta,t-1}(\widehat{\mathrm{m}}_t(\mathrm{X}_{t-1}) - \bar{\mathrm{m}}_t)$.

- Conventional Fama-MacBeth type regression

$$\widehat{\boldsymbol{\lambda}}_t = \frac{1}{N}\widehat{\mathrm{G}}'_{\beta,t-1}\bar{\mathrm{m}}_t, \quad \widehat{\mathrm{G}}_{\alpha,t-1} := \bar{\mathrm{m}}_t - \widehat{\mathrm{G}}_{\beta,t-1}\widehat{\boldsymbol{\lambda}}_t.$$

Step 4 – Out-of-sample Predictions

★ For predictions, we need to know how alphas and risks depend on $\mathrm{x}_{i,T}$ by use DNN:

$$\widehat{g}_{\mathrm{riskP},T}(\cdot) = \arg\min_{r\in DNN}\sum_{i=1}^{N}(\underbrace{\widehat{g_{\beta,t}(\mathrm{x}_{i,T-1})}'\boldsymbol{\lambda}_T}_{\hat{g}_{riskP,T,i}} - r(\mathrm{x}_{i,T-1}))^2$$

$$\widehat{g}_{\alpha,T}(\cdot) = \arg\min_{g\in DNN}\sum_{i=1}^{N}(\underbrace{\widehat{g_{\alpha,T}(\mathrm{x}_{i,T-1})}}_{\hat{g}_{\alpha,T,i}} - g(\mathrm{x}_{i,T-1}))^2$$

★ Not only can predict returns, but also can predict α and β.

Theory

In sample theory

Theorem 1. Under some ▸ technical cond.

Spot E-return: $\mathbb{E}[\widehat{m}_t(x_{i,t-1}) - \mathbb{E}(y_{it}|x_{i,t-1}, f_t)]^2 = O_P(\delta_T^2 + \varphi_T^2)$.

long-term: $\mathbb{E}[\bar{m}_{i,t} - \mathbb{E}(y_{it}|x_{i,t-1})]^2 = O_P(\delta_T^2 + \varphi_T^2 + \eta_T^2)$.

1. Aapproximation error of DNN: ▸ Details about DNN

$$\varphi_T = \max_t \inf_{r \in DNN} \|g_{\alpha,t} - r\| + \max_t \inf_{r \in DNN} \|g_{\beta,t} - r\|$$

2. Complexity of the DNN: $\delta_T = \sqrt{\frac{\log(NT)p(DNN)}{N}}$
3. Smoothing error: $\eta_T = \frac{1}{\sqrt{Th}} + h^2$

In sample theory

Theorem 1. Under some technical cond.

Spot E-return: $\mathbb{E}[\widehat{m}_t(\mathrm{x}_{i,t-1}) - \mathbb{E}(y_{it}|\mathrm{x}_{i,t-1}, \mathrm{f}_t)]^2 = O_P(\delta_T^2 + \varphi_T^2).$

long-term: $\mathbb{E}[\bar{m}_{i,t} - \mathbb{E}(y_{it}|\mathrm{x}_{i,t-1})]^2 = O_P(\delta_T^2 + \varphi_T^2 + \eta_T^2).$

1. Aapproximation error of DNN: Details about DNN

$$\varphi_T = \max_t \inf_{r \in DNN} \|g_{\alpha,t} - r\| + \max_t \inf_{r \in DNN} \|g_{\beta,t} - r\|$$

2. Complexity of the DNN: $\delta_T = \sqrt{\frac{\log(NT)p(DNN)}{N}}$
3. Smoothing error: $\eta_T = \frac{1}{\sqrt{Th}} + h^2$

Statistical errors in function estimation

Theorem 2. For each t,

$$
\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}[\widehat{g}_{\alpha,t-1,i} - g_{\alpha,t}(\mathrm{x}_{i,t-1})]^2 &= O_P(\delta_T^2 + \varphi_T^2 + \eta_T^2), \\
\frac{1}{N}\sum_{i=1}^{N}[\widehat{g}_{\mathrm{riskP},t,i} - g_{\mathrm{riskP},t}(\mathrm{x}_{i,t-1})]^2 &= O_P(\delta_T^2 + \varphi_T^2 + \eta_T^2) \\
\frac{1}{N}\sum_{i=1}^{N}[\widehat{g}_{\mathrm{factor},t,i} - g_{\mathrm{factor},t}(\mathrm{x}_{i,t-1})]^2 &= O_P(\delta_T^2 + \varphi_T^2 + \eta_T^2).
\end{aligned}
$$

Out of sample Theory

Theorem 3. For the new observations $\mathrm{x}_{i,T}$

$\mathbb{E}(\xi_{i,T+1}|\mathcal{F}_T) = 0$, where $\mathcal{F}_T = \sigma(X_1, ..., X_T)$,

$$y_{i,T+1} = \widehat{g}_{\alpha,T}(\mathrm{x}_{i,T}) + \underbrace{\widehat{g}_{\mathrm{riskP},T}(\mathrm{x}_{i,T})}_{g_\beta' \lambda} + \xi_{i,T+1} + O_P(a_{NT}),$$

1. We estimate both "g" functions that have predictive power
2. $a_{NT} = \delta_T + \varphi_T + \eta_T$ is the "statistical error".

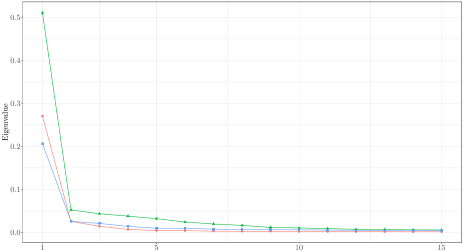
Empirical Analysis

Data

- ★ 62 firm characteristics from CRSP and Compustat (Freyberger et al., 2020)
- ★ Sample Period: 1965 - 2018, 648 months, 4261 firms on ave.
- ★ **Characteristics**: market cap, book-to-market, profitability, investment, beta, idiosyncratic volatility, turnover, bid-ask spread, short-term reversal, momentum, intermediate momentum, long-run reversal, total assets, cash over assets, D&A over assets, fixed costs to assets, capex to assets, operating leverage, price-to-cost margin, return-on-equity, operating accruals, free-cash flow to book value of equity, Tobin's Q, net payout ratio, assets-to-market cap, total assets, capital turnover, capital intensity, change in PP&E, earnings to price, and others

Implementation Details

- 60 months estimation window
- FNN with 3 layers (32-16-8),
- Bandwidth for kernel $h = 0.75$.
- Choose 5 factors from eigen-ratios to be safe



- Out-of-sample: 1970 - 2018, one-month ahead pred.

In-sample Decomposition

	1970 - 2018				1970 - 1999				2000 - 2018			
	R^2_{total}	R^2_{risk}	$g'_{\alpha,t}y_t$	$\frac{g'_{\alpha,t}y_t}{\sigma(g'_{\alpha,t}y_t)}$	R^2_{total}	R^2_{risk}	$g'_{\alpha,t}y_t$	$\frac{g'_{\alpha,t}y_t}{\sigma(g'_{\alpha,t}y_t)}$	R^2_{total}	R^2_{risk}	$g'_{\alpha,t}y_t$	$\frac{g'_{\alpha,t}y_t}{\sigma(g'_{\alpha,t}y_t)}$
All Firms												
	11.89	95.86	2.04	1.23	11.43	95.40	2.03	1.40	12.78	96.75	2.08	1.02
Large Firms												
	16.00	95.38	1.92	0.78	14.84	94.89	1.80	0.91	18.28	96.32	2.15	0.66
Small Firms												
	11.08	95.59	2.08	1.05	10.55	95.05	2.08	1.24	12.14	96.65	2.06	0.83

★ Large firms: top 20% market CAP

★ Most of the explained variation in returns due to risk ($\geq 95\%$)

★ Mispricing economically meaningful (2% monthly, Sharpe ratio > 1)

Risk Premia vs. Factor Shock

$$\widehat{y}_{it} = \beta_{\mathrm{rp}} \widehat{r}_{\mathrm{risk},t,i} + \beta_{\mathrm{factor}} \widehat{g}_{\mathrm{factor},t,i} + \varepsilon_{it}$$

	1970 - 2018		1970 - 1999		2000 - 2018	
	$\beta_{\text{risk premium}}$	$\beta_{\text{factor shock}}$	$\beta_{\text{risk premium}}$	$\beta_{\text{factor shock}}$	$\beta_{\text{risk premium}}$	$\beta_{\text{factor shock}}$
All Firms						
	0.172	0.973	0.180	0.968	0.157	0.984
Large Firms						
	0.127	0.972	0.126	0.967	0.129	0.981
Small Firms						
	0.170	0.973	0.175	0.968	0.161	0.983

★ Factor shock takes lions share

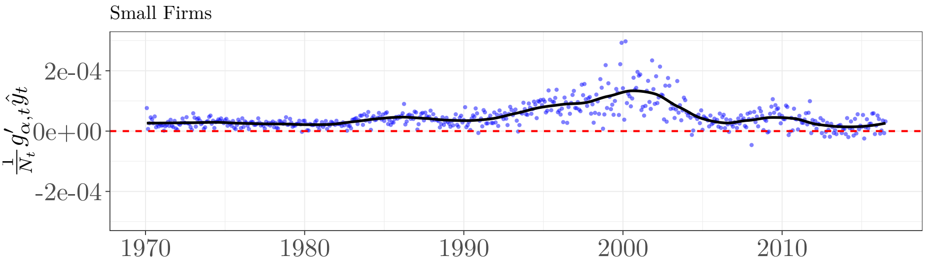
Out-of-sample Decomposition

	1970 - 2018				1970 - 1999				2000 - 2018			
	$R^2_{\hat{y}}$	$R^2_{g_\alpha}$	$R^2_{g_\beta}$	$R^2_{g_\beta,g_\alpha}$	$R^2_{\hat{y}}$	$R^2_{g_\alpha}$	$R^2_{g_\beta}$	$R^2_{g_\beta,g_\alpha}$	$R^2_{\hat{y}}$	$R^2_{g_\alpha}$	$R^2_{g_\beta}$	$R^2_{g_\beta,g_\alpha}$
All Firms												
	1.619	0.826	1.801	2.376	1.520	0.620	1.627	2.027	1.778	1.154	2.079	2.931
Large Firms												
	3.106	1.132	3.746	4.571	2.608	0.926	3.066	3.768	3.900	1.460	4.829	5.849
Small Firms												
	1.485	0.743	1.557	2.087	1.386	0.525	1.355	1.708	1.643	1.091	1.880	2.690

★ Greater predictive accuracy by focusing only on the risk-premium and mispricing, than plugging in new data

$$
\begin{aligned}
y_{i,t+1} &\approx \widehat{g}_{\alpha,t}(\mathrm{x}) + \widehat{r}_{risk,t}(\mathrm{x}_{i,t}) + g_{\beta,t}(\mathrm{x}_{it})'(\mathrm{f}_{t+1} - \mathbb{E}\mathrm{f}_{t+1}) + e_{i,t+1} \\
\widehat{y}_{i,t+1|t} &\approx \widehat{g}_{\alpha,t}(\mathrm{x}) + \widehat{r}_{risk,t}(\mathrm{x}_{i,t}) + g_{\beta,t}(\mathrm{x}_{it})'(\mathrm{f}_{t} - \mathbb{E}\mathrm{f}_{t})
\end{aligned}
$$

Temporal Evolution of Pricing error



- Smallest 80% firms, pricing error decays in recent sample
- Evidence of allowing time-varying $g_{\alpha,t}(x)$:

$$\frac{1}{N_t} g'_\alpha \widehat{y}_t \to^P \int_0^1 g_{\alpha,t}(x)^2 dx$$

- Economic interpretation: "learn and arbitrage away".

Simulation model

- Generate five characteristics: for each i, t,

$$x_{i,t,k} = \frac{1}{N+1} \text{rank}(\bar{x}_{i,t,k}), \quad \bar{x}_{i,t,k} = 0.98^k \bar{x}_{i,t,k-1} + \epsilon_{x,i,t,k},$$

- Set $g_{\alpha,t}(\mathrm{x}) = [\phi_1(\mathrm{x}), ..., \phi_5(\mathrm{x})]\boldsymbol{\theta}_{\alpha,t}$, where:

$$\min_{\boldsymbol{\theta}_t} \sum_{i=1}^{N} (g_{\alpha,t}(\mathrm{x}_{i,t-1}) - \widehat{a}_i)^2, \qquad \sum_{i=1}^{N} g_{\alpha,t}(\mathrm{x}_{i,t-1}) g_{\beta,t}(\mathrm{x}_{i,t-1}) = 0$$

$\widehat{a}_i$ is the alpha from Fama-French-5.

- $g_{\alpha,t}(\mathrm{x}_{i,t-1})$ explains 20% variations in $\mathbb{E}(y_{it}|\mathrm{x}_{i,t-1})$ at each period.
- Fix $N = 500$ firms and $T = 200$ periods.

Simulation model

- Generate five characteristics: for each i, t,

$$x_{i,t,k} = \frac{1}{N+1}\text{rank}(\bar{x}_{i,t,k}), \quad \bar{x}_{i,t,k} = 0.98^k \bar{x}_{i,t,k-1} + \epsilon_{x,i,t,k},$$

- Set $g_{\alpha,t}(\mathrm{x}) = [\phi_1(\mathrm{x}), \dots, \phi_5(\mathrm{x})]\boldsymbol{\theta}_{\alpha,t}$, where:

$$\min_{\boldsymbol{\theta}_t} \sum_{i=1}^{N} (g_{\alpha,t}(\mathrm{x}_{i,t-1}) - \widehat{a}_i)^2, \qquad \sum_{i=1}^{N} g_{\alpha,t}(\mathrm{x}_{i,t-1}) g_{\beta,t}(\mathrm{x}_{i,t-1}) = 0$$

  $\widehat{a}_i$ is the alpha from Fama-French-5.

- $g_{\alpha,t}(\mathrm{x}_{i,t-1})$ explains 20% variations in $\mathbb{E}(y_{it}|\mathrm{x}_{i,t-1})$ at each period.

- Fix $N = 500$ firms and $T = 200$ periods.

In-sample and Out-of-sample RMSE

	alpha		risk	
	median	std $\times 10$	median	std $\times 10$
	In-sample			
DNN-varying	**0.880**	0.160	**0.589**	0.145
Linear-varying	1.334	0.687	0.657	0.175
DNN-mw	0.884	0.128	0.763	0.104
Linear-mw	1.420	0.823	0.925	0.209
	Out-of-sample			
DNN-varying	**0.893**	0.200	**0.518**	0.271
Linear-varying	0.972	0.444	0.547	0.286
DNN-mw	0.946	0.199	0.624	0.216
Linear-mw	1.047	0.539	0.717	0.385

★DNN-varying outperforms competing methods

Summary

★ New methods to understand neural network predictions in finance

★ Decompose in- and out-of-sample predictions:

•risk premium (1:5) •factor exposure •mispricing (5%-R^2, 2%-mRet.)

★ factor exposure has no pred power; riskP dominates pred

★ Important to model the time-varyingness.

★ First theoretical analysis of neural networks in finance

Summary

★ New methods to understand neural network predictions in finance

★ Decompose in- and out-of-sample predictions:
•risk premium (1:5) •factor exposure •mispricing (5%-R^2, 2%-mRet.)

★ factor exposure has no pred power; riskP dominates pred

★ Important to model the time-varyingness.

★ First theoretical analysis of neural networks in finance

The End

Thank

You

Technical Conditions

1. Cross sectional independence and serial weak dependences.
2. Smooth over time:

$$\mathbb{E}(y_{it}|\mathrm{x}_{i,t-1}) = m_i\left(\frac{t}{T}\right), \quad g_{\beta,t}(\mathrm{x}_{i,t-1}) = \mathrm{g}_i\left(\frac{t}{T}\right), \forall t = 1, \ldots, T,$$

   where the functions are continuously twice-differentiable.
3. For DNN: (i) functions $m_t^0, g_{\alpha,t}$ and $g_{\beta,t}$ belong to the Hölder ball:

$$\|f\|_{\mathcal{H},q,\gamma} = \sup_{\mathrm{a},\mathrm{b}} \frac{|f^{(q)}(\mathrm{a}) - f^{(q)}(\mathrm{b})|}{\|\mathrm{a}-\mathrm{b}\|^{\gamma}} < L.$$

   (ii) Dim of neuralnet satisfies: $p(\mathcal{M}_{J,L})\log^{3/2}(NT) = o(N)$.
4. Some moment bounds
5. $\mathbb{E}(\gamma_T|\mathcal{F}_T) = 0$, $\mathbb{E}(\mathrm{f}_{T+1}|\mathcal{F}_T) = \mathbb{E}\mathrm{f}_{T+1}$.

back to Thm1

More details about DNN

1. For DNN approximation to composition of smooth functions

$$f_0 = g_q \circ g_{q-1} \circ \ldots \circ g_1$$

   - (Schmidt-Hieber, 2020): with properly chosen width J and depth L,

$$\varphi_T = N^{-\max_{i \le q} \frac{\beta_i}{\beta_i + d_i}}, \qquad \beta_i = \text{smoothness of } g_i$$

   - d_i is the instrinsic dimension, and can be much smaller than $\dim(x)$.

e.g., for single index models, $d_i = 1$.

   - DNN adaptively achieves the fast approximation rate

2. For the complexity:

back to Thm1

   - $p(DNN)$ is the VC dimension of the graph $\{h(x) - y : h \in \mathcal{D}NN\}$.
   - Bartlett et al. (2019): For ReLu networks,

$$p(DNN) \le CJ^2 L^2 \log(JL).$$

References I

Avramov, D., S. Cheng, and L. Metzker (2021). Machine learning versus economic restrictions: Evidence from stock return predictability. *Available at SSRN 3450322*.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*, 135–171.

Bali, T., A. Goyal, D. Huang, F. Jiang, and Q. Wen (2021). Different strokes: Return predictability across stocks and bonds with machine learning and big data. *Swiss Finance Institute, Research Paper Series*, 20–110.

Bartlett, P. L., N. Harvey, C. Liaw, and A. Mehrabian (2019). Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *J. Mach. Learn. Res. 20*, 63–1.

Belkin, M., D. Hsu, and J. Xu (2020). Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science 2*(4), 1167–1180.

Bianchi, D., M. Büchner, and A. Tamoni (2021). Bond risk premiums with machine learning. *The Review of Financial Studies 34*(2), 1046–1089.

Bryzgalova, S., M. Pelger, and J. Zhu (2020). Forest through the trees: Building cross-sections of stock returns. *Available at SSRN 3493458*.

Chen, L., M. Pelger, and J. Zhu (2020). Deep learning in asset pricing. *Available at SSRN 3350138*.

Cong, L. W., K. Tang, J. Wang, and Y. Zhang (2021). Alphaportfolio: Direct construction through deep reinforcement learning and interpretable ai. *Available at SSRN 3554486*.

Connor, G. and R. A. Korajczyk (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics 15*(3), 373–394.

Connor, G., H. Matthias, and O. Linton (2012). Efficient semiparametric estimation of the fama-french model and extensions. *Econometrica 80*, 713–754.

DeMiguel, V., A. Martin-Utrera, F. J. Nogales, and R. Uppal (2020). A transaction-cost perspective on the multitude of firm characteristics. *The Review of Financial Studies 33*(5), 2180–2222.

Fan, J., Y. Liao, and W. Wang (2016). Projected principal component analysis in factor models. *Annals of Statistics 44*(1), 219–254.

Ferson, W. E. and C. R. Harvey (1999). Conditioning variables and the cross section of stock returns. *The Journal of Finance 54*(4), 1325–1360.

Freyberger, J., A. Neuhierl, and M. Weber (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies 33*(5), 2326–2377.

Gagliardini, P., E. Ossola, and O. Scaillet (2016). Time-varying risk premium in large cross-sectional equity data sets. *Econometrica 84*(3), 985–1046.

Ghysels, E. (1998). On stable factor structures in the pricing of risk: do time-varying betas help or hurt? *The Journal of Finance 53*, 549–573.

Giglio, S., Y. Liao, and D. Xiu (2021). Thousands of alpha tests. *The Review of Financial Studies 34*(7), 3456–3496.

Giglio, S. and D. Xiu (2021). Asset pricing with omitted factors. *Journal of Political Economy 129*(7), 000–000.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies 33*(5), 2223–2273.

Gu, S., B. T. Kelly, and D. Xiu (2019). Autoencoder asset pricing models.

Hastie, T., A. Montanari, S. Rosset, and R. J. Tibshirani (2019). Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*.

References III

Kelly, B. T., S. Pruitt, and Y. Su (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics 134*(3), 501–524.

Kelly, B. T., S. Pruitt, and Y. Su (2020). Instrumented principal component analysis. *Available at SSRN 2983919*.

Kim, S., R. A. Korajczyk, and A. Neuhierl (2021). Arbitrage portfolios. *The Review of Financial Studies 34*(6), 2813–2856.

Lettau, M. and S. Ludvigson (2001). Resurrecting the (c) capm: A cross-sectional test when risk premia are time-varying. *Journal of political economy 109*(6), 1238–1287.

Li, B. and A. G. Rossi (2020). Selecting mutual funds from the stocks they hold: A machine learning approach. *Available at SSRN 3737667*.

Mei, S. and A. Montanari (2019). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*.

Rossi, A. G. and S. P. Utkus (2020). Who benefits from robo-advising? evidence from machine learning. *Evidence from Machine Learning (March 10, 2020)*.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics 48*(4), 1875–1897.

Shanken, J. (1990). Intertemporal asset pricing: An empirical investigation. *Journal of Econometrics 45*(1-2), 99–120.

Stock, J. and M. Watson (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*, 1167–1179.

Stock, J. and M. Watson (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics 20*(2), 147–162.