Big Data and AI in Econometrics, Finance, and Statistics Conference Takeaways

LaTeX by Junfan Zhu
Stevanovich Center for Financial Mathematics
University of Chicago
October 3-5, 2024

October 7, 2024

Contents

1 Agenda

This is a notes from current leading academia research at the University of Chicago Stevanovich Center for Financial Mathematics's Annual Conference: Big Data and AI in Econometrics, Finance & Statistics, on 2024 Oct 3-5.

My interest of selected takeaways focuses more on Financial Mathematics and Econometrics than Statistics. Due to the complexity of the formulas, kindly contact me if you detect typos to correct.

Oct 3

- Session 1
  - Katja Smetanina, UChicago. Perceived shocks and impulse responses [15]
  - Zongming Ma, Yale. Multimodal data integration and cross-modal querying via orchestrated approximate message passing [9]
  - Ruimeng Hu, UC Santa Barbara. Deep reinforcement learning for games with controlled jump-diffusion dynamics [6]
- Session 2
  - Chao Gao, UChicago. Are adaptive robust confidence intervals possible? [4]
  - Fang Han, UWashington. Chattejee's rank correlation: what is new? [5]
  - Pragya Sur, Harvard. Generalization error of min-norm interpolators in transfer learning [16]
- Session 3
  - Bodhi Sen, Columbia. Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood. [14]
  - Wenxin Zhou, UIC. Nonparametric expected shortfall regression with tail-robustness [19]

Oct 4

- Session 4
  - Theodor Misiakiewicz, Yale. Deterministic equivalents and scaling laws for random feature regression [10]
  - Bryan Kelly, Yale. APT or 'AIPT'? The surprising dominance of large factor models [7]
  - Mladen Kolar, USC. Confidence sets for causal discovery [8]
- Session 5
  - Andrea Montanari, Stanford. Overparametrization in machine learning: insights from linear models [11]
  - Chenhao Tan, UChicago. Towards human-centered AI: predicting fatigue and generating hypothesis with LLMs [17]
- Session 6
  - Dacheng Xiu, UChicago. On the theory of autoencoders [18]
  - Denis Chetverikov, UCLA. Estimation of risk premia with many factors [2]

Oct 5

- Session 7
  - Whitney Newey, MIT. Automatic debiased machine learning via Riesz regressions [12]
  - Francis Diebold, UPenn. Machine learning and the yield curve: tree-based macroeconomic regime switching [3]
- Session 8
  - Yacine Ait-Sahalia, Princeton. Asset pricing in an economy with changing sentiment and price feedback [1]
  - Per Mykland/Lan Zhang, UChicago/UIC. Estimating the volatility of drift [13]

2 Machine Learning and the Yield Curve: Based Macroeconomic Regime Switching [3]

The Dynamic Nelson-Siegel (DNS) Model

Diebold and Li (2006):

$$y_t = \Lambda\mu + \Lambda F_t + \epsilon_t$$

$$F_t = AF_{t-1} + \eta_t$$

$$\begin{pmatrix}\epsilon_t \\ \eta_t\end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix}Q & 0 \\ 0 & H\end{pmatrix}\right)$$

$$\Lambda = \begin{pmatrix} 1 & \frac{1-e^{-\lambda\tau_1}}{\lambda\tau_1} & \frac{1-e^{-\lambda\tau_1}}{\lambda\tau_1} - e^{-\lambda\tau_1} \\ 1 & \frac{1-e^{-\lambda\tau_2}}{\lambda\tau_2} & \frac{1-e^{-\lambda\tau_2}}{\lambda\tau_2} - e^{-\lambda\tau_2} \\ \vdots & \vdots & \vdots \\ 1 & \frac{1-e^{-\lambda\tau_N}}{\lambda\tau_N} & \frac{1-e^{-\lambda\tau_N}}{\lambda\tau_N} - e^{-\lambda\tau_N} \end{pmatrix}$$

y_t is an N-vector of observed yields at various maturities.
$F_t = f_t - \mu$ is a 3-vector of latent factors, centered around their means.

Issues With DNS

- Assumes constant coefficients μ, A, etc.
- But in reality there may be abrupt shifts, or regime switching.
  e.g., Dai, Singleton and Yang (2007), Hevia et al. (2015)
- One can introduce Markov switching as in Hamilton (1989) and Kim (1994).
- But what do the latent regimes mean? Researchers tend to casually interpret them, ex post, as indicative of macroeconomic conditions.
  e.g., Hamilton (1989), Bansal and Zhou (2002).
- We will link yield-curve regimes clearly and directly, ex ante, to macroeconomic conditions. "Macro-guided yield curve regimes"

Macro Spanning

- We begin with a "yields-only" model (three yield factors), and then we move to a "yields-macro" model (three yield factors plus three macro variables). The yields-macro model lets us investigate macro spanning.
- Macro spanning means that all current and past macro relevant for future macro is contained in the current yield curve. Hence yields may Granger-cause macro, but macro should not Granger-cause yields.
- No consensus yet. Compare Bauer and Rudebusch (2017) to Joslin, Priebsch, and Singleton (2014) and Bekaert, Engstrom, and Ermolov (2021).
- We will provide a new perspective: Macro spanning may hold in some regimes, but be violated in others.

Yields-Macro DNS with Regime Switching

$$\begin{pmatrix} y_t \\ m_t \end{pmatrix} = \begin{pmatrix} \Lambda & 0 \\ 0 & I_3 \end{pmatrix} F_t + \begin{pmatrix} \Lambda & 0 \\ 0 & I_3 \end{pmatrix} \mu_{z_t} + \epsilon_t \tag{1}$$

$$F_{f_t} = A_{z_{t-1}} F_t + \eta_t \tag{2}$$

$$\begin{pmatrix} \epsilon_t \\ \eta_t \end{pmatrix} \sim N\left(0, \begin{pmatrix} Q & 0 \\ 0 & H_{z_t} \end{pmatrix}\right) \tag{3}$$

- State F_{f_t} is now a 6-vector:
  - 3 centered yield factors and 3 centered macro indicators
  - Inflation, capacity utilization, fed funds rate
  - Follows Diebold, Rudebusch, and Aruoba (2006)
- Facilitates investigation of macro spanning by allowing for yields-macro interaction
- Macro spanning, if any, may be regime-specific

Tree-Based Estimation

- Parameters with G regimes: $\Theta = (\lambda, \{A\}_{g=1}^G, \{\mu\}_{g=1}^G, \{H\}_{g=1}^G, Q)$
- Bayesian estimation conditional on exogenously-known regimes requires only Kalman filter/smoothers + MCMC.
- Standard priors:
  - Transition matrix A_g:
    - Gaussian for diagonal elements
    - Spike-and-slab for off-diagonal elements
  - Covariance matrices H_g and Q:
    - Inverse-Wishart for non-diagonal matrix H_g
    - Inverse Gammas for diagonal elements (variances) of diagonal matrix Q
  - Factor mean μ_g: Gaussian
  - Decay parameter λ: Uniform
- But regimes are not exogenously known. We specify $G \leq 3$, but we estimate regime locations along with the other DNS model parameters.
  - Conceptually: For each possible tree, evaluate the DNS model's marginal likelihood, and select the tree that produces the highest marginal likelihood.
  - In practice: Rather than evaluating all trees, proceed in a sequential greedy fashion.

Monthly Data, August 1971 - December 2022

Thirteen zero-coupon U.S. Treasury bond yields (Liu and Wu, 2021): Maturities of 3, 6, 9, 12, 24, 36, 48, 60, 72, 84, 96, 108, and 120 months.

Ten macroeconomic split candidates:

Variable	**Description**
DTB3	3-Month Treasury Bill Secondary Market Rate, Discount Basis
UNRATE	Unemployment Rate
INDPRO	Industrial Production: Total Index (percent change from a year ago)
CPI	Consumer Price Index for All Urban Consumers (percent change from a year ago)
M2	Billions of Dollars, Seasonally Adjusted
PAYEMS	All Employees, Total Non-farm (percent change from a year ago)
OILPRICE	Spot Oil Price: West Texas Intermediate (percent change from a year ago)
TERM_SPREAD	Term Spread
DEFAULT_SPREAD	Default Spread
VIX	CBOE Volatility Index

Tree Structure and Regime Interpretation

One split at each tree depth. Need to specify split candidates, ordering of split candidates, possible split points, stopping criterion, etc.

For example, we might select (and this foreshadows our empirical results):

- Three regimes, {R1, R2, R3}.
- Each regime has clear economic interpretation.
- R2 holds when DTB3 < 0.6 and UNRATE < 0.4.

Empirical Results for Three-Regime Yields-Only Model

Four equispaced possible split locations in [0,1]: 0.2, 0.4, 0.6, 0.8

- Three regimes (light green for R1, green for R2, and orange for R3):
- UNRATE ≥ 0.6 (R1, high unemployment, 244 months)
- UNRATE < 0.2 (R2, low unemployment, 150 months)
- $0.2 \leq$ UNRATE < 0.6 (R3, medium unemployment, 223 months)

Summary

- Bond yield curve dynamics may switch with the macroeconomy.
- We have provided and illustrated a modeling framework.
- We implemented both yields-only and yield-macro models.
- Macro spanning may be regime-specific, failing in some regimes but not others.

3 Sentiment-Based Asset Pricing [1]

Affiliations:

- Bendheim Center for Finance, Princeton University
- Research School of Economics, Australian National University
- Department of Mathematics and RiskLab, ETH Zürich
- Instituto Tecnológico Autónomo de México, Business School

This paper

- New continuous-time equilibrium model (CCAPM) with a stochastic sentiment process η_t.
  - Dynamics of η_t in the model are empirically validated; based on four empirical sentiment indices.
  - Sentiment incorporates price feedback: investor gets more optimistic when prices go up, and vice versa.
  - Sentiment can jump; more likely when it gets more out of line with fundamentals.
- Equilibrium outcome yields testable closed-form solutions of:
  - Asset price dynamics
  - Excess return and risk-free rate: sentiment affects asset prices
  - Conditional variance of returns; 3rd+4th moment

Asset prices and investor sentiment

Behavioral literature indicates that investor sentiment has long been known to affect asset prices. Many empirical and theoretical challenges remain, such as:

1. Investor sentiment is an elusive concept: how to measure it
2. Plausible to assume existence of feedback effects from prices into sentiment: in which direction does causality go, or in both directions
3. Theoretically challenging to account for sentiment swings and price feedback effects in equilibrium
4. Aim to capture empirically realistic boom and bust cycles in the stock market

Aim: This paper derives a tractable equilibrium framework that allows sentiment to drive investor beliefs and where past price changes feed back into investor sentiment.

Related Literature

- **Behavioral Finance:** First and second generation, including works by De Long et al. (1990), Lee et al. (1991), Hong and Stein (1999), Barberis et al. (1998).
- **Theory:** Equilibrium model with sentiments by Dumas et al. (2009), Barberis et al. (2015), Maenhout et al. (2021).
- **Empirical:** Design of sentiment indices, regressions, predictability by Baker and Wurgler (2006a), Yu and Yuan (2011), Greenwood and Shleifer (2014), Birru and Young (2022), van Binsbergen et al. (2024).
- **Macroeconomics:** DSGE model with a sentiment channel by Angeletos and La' (2013), Benhabib et al. (2015), López-Salido et al. (2017), Martin and Papadimitriou (2022).

Key Building Blocks of the Paper

- Two novel features:
  1. Investor is subject to sudden large changes (jumps) in sentiment. Investor is not aware that their beliefs are not grounded in reality. They only care about current sentiment and do not account for potential future changes.
  2. The sentiment is subject to price feedback effects. Asset prices affect the investor's sentiment: higher (lower) prices relative to past prices increase sentiment. This generates positive time-series momentum.
- These features allow us to:
  - Generate realistic boom and bust cycles due to feedback effects between sentiment and stock prices.
  - The equilibrium stock price process can exhibit rapid build-up and sudden crashes even though the fundamentals driving the stock price process (dividends) might not jump = a Minsky moment.

The Asset Market under the Objective Measure $\mathbb{P}$

Framework: Lucas-type endowment economy with one representative investor, who has to choose optimal consumption and can invest in two assets, standard except for jumps in dividends:

- risk-free asset, $\frac{dB_t}{B_t} = r_{f,t}\,dt, B_0 > 0$
- risky stock price, $\frac{dS_t}{S_t} = \mu_{S,t}\,dt + \sigma_{S,t}\,dW_t^S + dJ_t^S, S_0 > 0$

where $r_{f,t}$ denotes the risk-free rate, μ_S the drift of the stock price process, σ_S its volatility and J_t^S its jump component, all of these processes are endogenous and will be characterized in equilibrium.

The risky asset represents a claim against the aggregate dividend process, which follows

$$\frac{dD_t}{D_{t-}} = \mu_D\,dt + \sigma_D W_t^D + dJ_t^D$$

for a drift $\mu_D \in \mathbb{R}$, a volatility $\sigma_D \geq 0$, a Brownian motion W^D and a compensated compound Poisson process J^D of the form

$$J_t^D = \sum_{i=1}^{N_t^D} (e^{Y_i^D} - 1) - \lambda_D \phi_d t$$

where:

- N^D is an independent Poisson process with constant intensity $\lambda_D \geq 0$,
- i.i.d. random variables $Y_i^D, i \geq 1$,
- $W^D, N^D, Y_1^D, Y_2^D, ...$ are all independent.

Investor Sentiment with a Feedback Channel

The sentiment process $\eta_t \in [-1, 1]$ follows a jump-diffusion process of the form:

$$d\eta_t = k(m_t - \eta_t)dt + \sigma_\eta \sqrt{1-\eta_t^2} dW_t^\eta - (1+\eta_{t-})dJ_t^{\eta-} + (1-\eta_{t-})dJ_t^{\eta+},$$

where:

$$\text{negative jumps: } J_t^{\eta-} = \sum_{i=1}^{N_t^{\eta-}} \left(1 - e^{-Y_i^{\eta-}}\right) - \lambda_t^{\eta-}\phi_{\eta-}t, \quad \phi_{\eta-} = \mathbb{E}\left[1 - e^{-Y_i^{\eta-}}\right]$$

$$\text{positive jumps: } J_t^{\eta+} = \sum_{i=1}^{N_t^{\eta+}} \left(1 - e^{-Y_i^{\eta+}}\right) - \lambda_t^{\eta+}\phi_{\eta+t} \quad \phi_{\eta+} = \mathbb{E}\left[1 - e^{-Y_i^{\eta+}}\right]$$

$N^{\eta-}$ and $N^{\eta+}$ are Poisson processes with time-varying intensities $\lambda_t^{\eta-}$ (decreasing in η) and $\lambda_t^{\eta+}$ (increasing in η).

$$W_t^\eta = \rho W_t^D + \sqrt{1-\rho^2} W_t^i \text{ is a Brownian motion.}$$

Extreme optimism (pessimism) corresponds to the case when $\eta_t \to 1(-1)$, and neutral when $\eta_t = 0$. Asset price feedback effect captured by the stochastic mean reversion level m_t:

$$m_t = \tanh(qM_t), \quad M_t = \log\left(\frac{I^{-1}\int_{t-1}^t S_u du}{L^{-1}\int_{t-L}^t S_u du}\right)$$

where $q > 0$ and $L > 0$ are parameters.

Modeling a Moody Investor: Subjective Measure $\mathbb{P}^\eta$

The sentiment process η governs the investors' beliefs through $\mathbb{P}$. η parametrizes the change of measure, Radon-Nikodym derivative:

$$\frac{d\mathbb{P}_t^\eta}{d\mathbb{P}_t} = \exp\left(-\frac{1}{2}\int_t^T \theta^2\eta_t^2 \, \mathrm{d}s - \int_t^T \theta\eta_t \, \mathrm{d}W_s^D\right) \neq \exp\left(-\frac{1}{2}\int_0^t \theta^2\eta_s^2 \, \mathrm{d}s - \int_0^t \theta\eta_s \, \mathrm{d}W_s^D\right), \quad \theta \geq 0$$

Investor believes that the dividend process follows:

$$\frac{\mathrm{d}D_s}{D_{s-}} = (\mu_D + \sigma_D\theta\eta_t)\,\mathrm{d}s + \sigma_D \, \mathrm{d}\widetilde{W}_s^D + \mathrm{d}J_s^D$$

where:

- η_t: Introduces mood swings disconnected from the reality of the dividends process.
  - Under the investor's subjective beliefs, the drift rate of dividends is now given by $(\mu_D + \sigma_D\theta\eta_t)$.
  - Sentiment swings now affect the investor's beliefs about the growth rate of dividends through $\theta\eta_t$.
- θ: Controls the magnitude of the investor's mood swings.

Investor's preferences

The agent has recursive preferences

$$V_t = \mathbb{E}^{\mathbb{P}_t^\eta}\left[\int_t^\infty f(C_u, V_u)\,\mathrm{d}u \mid \mathcal{F}_s^D\right]$$

where $(\mathcal{F}_s^D)_{s\geq 0}$ is the filtration generated by the dividend process $(D_s)_{s\geq 0}$ and

$$f(C,V) = \frac{\beta(1-\gamma)}{1-\frac{1}{\psi}} V\left(\left(\frac{C}{((1-\gamma)V)^{\frac{1}{1-\gamma}}}\right)^{1-\frac{1}{\psi}} - 1\right)$$

with parameters

- $\beta > 0$: discount rate
- $\gamma > 0$: coefficient of relative risk aversion
- $\psi > 0$: elasticity of intertemporal substitution

We assume that the agent solves his/her optimization problem myopically

- Investor only takes into account the current value of his or her sentiment in forming his or her beliefs, and not the projected evolution of his/her sentiment

Investor's Time Inconsistency

- The representative agent is not aware that his/her beliefs are not aligned with the reality of the dividends process. He/she is not aware of the existence of sentiment, let alone of its dynamics.
- Given the structure of beliefs, the only source of randomness in the agent's view are the stochastic dividend rates $(D_s)_{s\geq t}$: the agent evaluates future prospects using the measure $\mathbb{P}_t^\eta$ but does so conditionally on D_t only (rather than conditionally on (D_t, η_t)).
- So at time t, the representative agent believes that the dividend rate will grow with drift $\mu_D + \sigma_D\theta\eta_t$ forever into the future.
- This makes the representative agent time-inconsistent since he/she is not taking into account that his/her beliefs might change in the future.
- But it would be somewhat strange to assume that the agent is somehow irrational enough to have beliefs driven by a sentiment process that distorts the reality of the dividend process, yet aware enough that his/her beliefs do not line up with the fundamentals and in which way they do, and then rational enough to attempt to hedge against changes in beliefs predicted by the dynamics of the sentiment process.

Partial Equilibrium: Optimal Portfolio & Consumption Choice

The dynamics of the investor's wealth is given by:

$$\frac{\mathrm{d}X_t}{X_{t-}} = \pi_t \frac{\mathrm{d}S_t + D_t\,\mathrm{d}t}{S_{t-}} + (1-\pi_t)\frac{\mathrm{d}B_t}{B_t} - \frac{C_t}{X_t}\,\mathrm{d}t$$

where π_t denotes the fraction of wealth invested in the stock market and C_t is the consumption rate. The associated HJB equation for the value function $v : (0,\infty) \to \mathbb{R}$ solving the HJB equation is:

$$\begin{aligned}&\max_{\pi,c}\left\{f(c,v(x)) + v'(x)\left[xr_t + \pi x\left(\mu_{D,t} - \lambda_D\phi_D + k_t - r_t\right) - c\right]\right.\\&\left.+\frac{1}{2}v''(x)\pi^2x^2\sigma_D^2 + \lambda_D\mathbb{E}\left[v\left(x + \pi x\left(e^{Y_1^D} - 1\right)\right) - v(x)\right]\right\} = 0\end{aligned}$$

with $\phi_D = \mathbb{E}[e^{Y^D} - 1]$. To solve this equation, we make the ansatz

$$v(x) = a_t\frac{x^{1-\gamma}}{1-\gamma} \quad \text{for coefficient } a_t > 0 \text{ depending on } \mu_{D,t} = \mu_D + \sigma_D h_t.$$

Proposition (Equilibrium asset prices and & risk free rate)

In an endowment economy populated by an investor with subjective sentiment-dependent preferences, the equilibrium asset price is given by

$$\begin{aligned}S_t &= \frac{D_t}{k_t}, \text{ where } k_t = A - B\eta_t > 0\\A &= \beta - \frac{\psi-1}{\psi}\left(\mu_D - \frac{\gamma\sigma_D^2}{2} - \lambda_D\zeta_D(\gamma)\right), B = \frac{\psi-1}{\psi}\sigma_D\theta\end{aligned}$$

The resulting equilibrium risk-free rate is

$$r_{f,t} = \underbrace{\beta + \frac{1}{\psi}\left(\mu_D - (1+\psi)\frac{\gamma\sigma_D^2}{2}\right)}_{\text{standard diffusive model}} + \underbrace{\frac{1}{\psi}\eta_t\theta_{\sigma_D}}_{\text{misjudgement}} \underbrace{-\lambda_D\left(\xi_D(\gamma) - \frac{\psi-1}{\psi}\zeta_D(\gamma)\right)}_{\text{dividend jump risk}}$$

Key role of inter-temporal substitution: We assume $\psi > 1$ because

- Case (I): $\psi = 1$, then $A = \beta$ and $B = 0 \to$ Price-dividend ratio becomes constant $S_t/D_t = 1/\beta$ (inconsistent with the data)

- Case (II): $\psi = 1/\gamma$, $\rightarrow$ For $\gamma > 1$, model makes counterfactual prediction that price-dividend ratio is lower when investor sentiment is optimistic $\eta \rightarrow 1$, i.e. $\frac{\partial(S_t/D_t)}{\partial \eta_t} = \frac{(\gamma-1)\theta\sigma_D}{\left(\beta+(1-\gamma)(\mu_D+\sigma_D\theta\eta_t)\right)^2} < 0, \theta > 0$
- Case (III): $\psi > 0$, For $\phi > 1$, higher sentiment leads to a higher price-dividend ratio which is consistent with empirical results.

Proposition (The Equity Risk Premium under sentiment swings)

The instantaneous excess return $\mu^E_{S,t} := \frac{1}{dt}\mathbb{E}_t\left[\frac{dS_t+D_t dt}{S_t}\right] - r_{f,t}$ is given by

$$\mu^E_{S,t} = \underbrace{\gamma\sigma_D^2 + \lambda_D\xi_D(\gamma)}_{\text{Standard model}} - \underbrace{\eta_t\theta\sigma_D}_{\text{misjudgment}} + \underbrace{\left(\frac{B}{A-B\eta_t}\right)^2\sigma_\eta^2(1-\eta_t^2)}_{P/D\text{-Sentiment risk premium}} + \underbrace{\frac{B\rho\sigma_D\sigma_\eta\sqrt{1-\eta_t^2}}{A-B\eta_t}}_{\text{Cov. risk premium}}$$

$$+ \frac{B}{A-B\eta_t}\{\underbrace{\kappa(m_t-\eta_t)}_{\text{Feedback Effect}} + \underbrace{B\left(\lambda_t^{\eta^-}\Psi_t^- + \lambda_t^{\eta^+}\Psi_t^+\right)}_{\text{Sentiment jump risk premium}}\}$$

where the negative and positive sentiment jump terms are given by

$$\Psi_t^- = \int_0^\infty \frac{(1+\eta_t)^2(1-e^{-y})^2}{A+B\left[1-(1+\eta_t)e^{-y}\right]}\nu_-^\eta(dy), \quad \psi_t^+ = \int_0^\infty \frac{(1-\eta_t)^2(1-e^{-y})^2}{A+B\left[(1-\eta_t)e^{-y}-1\right]}\nu_+^\eta(dy)$$

Discussion

- Fully rational investor implies $\eta_t = 0$, risk premium reduces to $\gamma\sigma_D^2 + \lambda_D\xi_D(\gamma)$
- Misjudgment term for an overly optimistic (pessimistic) investor is positive (negative) leading to a lower (higher) equity risk premium
- Feedback effect increases (decreases) equity risk premium provided that $m_t > \eta_t$, High past price appreciation, i.e. m_t increase can generate positive momentum.

Data

Asset Market Data

- Aggregate Stock Market: S&P 500 (logarithmic) returns including dividends obtained from the CRSP database.
- The risk-free rate is the one-month (constant maturity) Treasury bill rate, constructed via linear interpolation from daily effective fed funds rate from the Fred St. Louis Database (Ticker DFF) and the one-year maturity zero coupon yield obtained from the Federal Reserve.
- Frequency and time period: Daily, from January 1, 1962 until December 31, 2023.

Sentiment Data

- The Investor Sentiment Survey Index by the American Association of Individual Investors (AAII). Vote is then expressed in terms of whether their view is bullish, neutral or bearish about the stock market.
- We construct a bullish-bearish spread measure as a proxy of investor sentiment.
- Frequency and time period: Weekly frequency, from July 24th, 1987 to present day moving average of the difference between the count of positive and negative news about stocks.
- Frequency and time period: Daily starting in January 1st, 2000.

Empirical Validation of the Sentiment Process

Road map: We want to empirically show that:

- Feedback effects are important: drift
  - Prices affect sentiment through m_t and this feedback effect helps explain changes in sentiment η_t.
  - We run the following regression $\Delta\eta_{m,t} = \beta_0 + \beta_1 \left(K_{m,L_p} - \eta_{t,m}\right)$, with $K_{m,L_p} \in \{m_t, 0\}$.
  - Model-independent analysis: Estimate Vector Autoregressive (VAR) Model and perform Granger Causality tests.
- Diffusive volatility, i.e., $\sigma_d(\eta) = \sqrt{1-\eta^2}$, is hump-shaped in sentiment.
- Jump intensity in sentiment dependent on the level of sentiment.
  - Are upward (downward) jumps in sentiment more likely when sentiment is low (high)?

Validating the model for sentiment: Diffusive Volatility

- Model $\sigma_d(\eta_t) = \sqrt{1-\eta_t^2}$ implies humped-shaped diffusive volatility when plotted as a function of η_t
- Remark: Jump detection threshold $k\sigma_d(\eta_t)$, where $\sigma_d(\eta_t)$ does not contain any jumps.

Validating the model for sentiment: Jump Component

- Is the jump intensity a function of η_t?
- If sentiment becomes very optimistic (pessimistic), are downward (upward) jumps in sentiment more likely?
- Results suggest that when sentiment is high (low), the likelihood of observing a downward (upward) jump in sentiment is elevated.

Theory: Equilibrium risk-free rate and sentiment

Main conclusion:

- Positive association between the equilibrium risk-free rate and investor sentiment
- $\rightarrow$ This follows because sentiment affects the $r_{f,t}$ linearly through the term $\frac{1}{\psi}\eta_t\theta\sigma_D$
- $\rightarrow$ Result is in line with economic intuition: In equilibrium, as the investor becomes more optimistic, interest rates have to rise to prevent the investor from excessively investing in the risky asset

Empirics: Risk-free rate and AAII sentiment

Notes: The plot on the left shows the logaritm of the risk-free rate $r_{f,t}^{\prime vU,i,n}$ with $h = 1$ and the AAll index. Dashed lines represent linear regressions fits and the curve is the lowess fit using 25% of the data. Right plot shows forecast regressions for different h.

AAII: Dividend Yield and Sentiment

Recall: From our setup, the dividend yield is

$$k_t = D_t/S_t = A - B\eta_t$$
$$A = \beta - \frac{\psi - 1}{\psi}\left(\mu_D - \frac{\gamma\sigma_D^2}{2} - \lambda_D\zeta_D(\gamma)\right)$$
$$B = \frac{\psi - 1}{\psi}\sigma_D\theta$$

Testable model implications:

- Dividend yield falls when the investor becomes more optimistic $\eta_t \uparrow$ since $B > 0$ when $\psi > 1$.
- Constants A (intercept) and B (slope) can be inferred from a OLS regression of dividend yield regressed onto sentiment

Notes: The left plot shows the dividend yield $k_t^{NO,I,h}$ (right y-axis) and the AAll index (pentiment axis). The right plot shows a scatter plot of the dividend yield and the AAII together with a linear regressions fit (dashed line) and a lowess fit using 25% of the data (dotted curve).

Theory: Excess return and sentiment

Recall: The excess return of the model is a highly complex function of sentiment.

- Leading (misjudgement) term $-\eta_t\theta\sigma_D$ is negative whereas other terms are positive but not monotonic in sentiment
- $\rightarrow$ We use calibrated model parameters to infer dependence of excess return on sentiment

Notes:

- The red horizontal line indicates empirical annual excess return, which is 6.5%
- The jump intensity specification is $\lambda_{\eta-,t} = \bar{\lambda}_-\left(1+\eta_t\right)^\alpha, \lambda_{\eta+,t} = \bar{\lambda}_+\left(1-\eta_t\right)^\alpha, \alpha > 0$

Main conclusion:

- Negative relationship between excess return and sentiment driven by general equilibrium effect
- $\rightarrow$ Since this is an endowment economy with a sole representative agent, prices need to adjust such that the investor is willing to hold all the asset supply
- $\rightarrow$ For markets to clear properly, if the investor becomes very optimistic, the implied excess return must fall otherwise the agent would want to short sell the risk free asset to increase his/her holdings in the risky asset

Concluding Remarks

This paper:

- Develops a flexible yet tractable framework where the representative investor is subject to mood swings.
- Introduces a novel sentiment process which allows for price feedback effects (can generate boom and bust cycles without jumps in the stock price fundamentals).
- Empirically validates the model's features and shows that feedback effects as well as jumps in sentiment are important to capture stylized facts of asset markets.
- The model's predictions for the excess return, dividend yield, realized volatility, and the risk-free rate are largely supported by our empirical results.

Further research questions we want to explore:

- Relationship between investor sentiment and higher order moments.
- We have closed-form expressions for higher moments of excess return as a function of sentiment.
- Full estimation of the model to recover model parameters and analyze effect of sentiment changes on excess returns and the risk-free rate in more detail.

4 Volatility of Drift [13]

Per Mykland and Lan Zhang, UChicago, UIC

Object of Interest

Price drift typically refers to stock price movement. It is important to understand the drift behavior:

- As drift occurs for individual financial asset, relevant portfolio could deviate away from its original allocations over time, and thus veer off from the intended investment strategy of the fund.
- Understanding price drift is crucial to find Sharpe Ratios and alphas.

Drift is usually estimated using long time periods (years, decades, and up), with little use of high frequency data (HFD). We approach an in-between question: with the help of HFD, can we estimate the volatility of price drift? A high volatility of price drift can help predict a change in long term performance of a stock.

HD^2 Data

Ultra High Frequency (UHF) main feature:

- up to millisecond updates, almost continuous observation
- observation times can be irregular
- has microstructure noise

High Dimension (HD) main feature:

- non-synchronous
- sparsity

HD^2 data can also be found in internet streaming, neuroscience, geoscience, climate recordings, wind measurements, turbulence, fish, etc.

This talk focuses on UHF data.

Model

Consider the log price process $\{x_t\}$ follows semi-martingale,

$$dx_t = \theta_t dt + \sigma_t dW_t,$$

with drift θ_t, and volatility σ_t:

- Focus on the drift over an interval $\int_0^T \theta_t dt$.
- The analysis covers multi-period extensions with intervals $(T_0, T_1], (T_1, T_2], ...$
- Ill-posed problem. Need $T \to \infty$ in each interval (at least in most circumstances). We seek to estimate the volatility $[\theta, \theta]_T - [\theta, \theta]_0$, or possibly a finite sum that approximates this volatility, such as the Average Realized Volatility (ARV) of the drift θ:

$$\text{ARV} = \frac{1}{\mathcal{T}} \frac{1}{L} \sum_i \left(\theta_{T_{i+L}} - \theta_{T_i}\right)^2$$

Ingredients of Estimation: The QV, with internal lag L

Denote $X_{(s,u]} = X_u - X_s$, i.e., the log return over the interval (s, u]. Similarly, $\Theta_{(s,u]} = \int_s^u \theta_t dt$, the martingale difference $M_{(s,u]} = M_u - M_s$. We write increments of a continuous semimartingale X as

$$X_{(s,u]} = \Theta_{(s,u]} + M_{(s,u]},$$

where $\Theta_{(s,u]} = \int_s^u \theta_t dt$ and $M_{(s,u]} = \int_s^u \sigma_t dW_t$.

We now consider an estimator of the form:

$$QV_{B,k,L}(X) = \frac{1}{k} \sum_{i=k}^{B-k-L} (X_{(T_{i+L}, T_{i+L+k}]} - X_{(T_{i-k}, T_i]})^2.$$

Assume that the T_is are evenly spaced out over $[0, T]$, with $T_0 = 0$ and $T_B = T$. We can write $T_i - T_{i-1} = \Delta T$, $\delta = k\Delta T$, and $\lambda = L\Delta T$. $X_{(T_{i+L}, T_{i+L+k}]}$ and $X_{(T_{i-k}, T_i]}$ can be viewed as two log returns (over an interval of length δ) that are (k+L) ΔT apart in time. Typically: k is small and L is substantially larger (to capture drift).

Estimation: A tapered QV, and our Estimator

Tapered, but no internal lag:

$$QV^{\text{tapered}}_{B,k+L,0}(X) = \frac{1}{k} \left(\frac{1}{2} \sum_{i=k}^{k+L-1} + \sum_{i=k+L}^{B-(k+L)} + \frac{1}{2} \sum_{i=B-(k+L)+1}^{B-k} \right) \left(X_{(T_i, T_{i+k}]} - X_{(T_{i-k}, T_i]} \right)^2$$

Our estimator compares the two QVs:

$$\widehat{\text{ARV}} = \frac{1}{\mathcal{T}\lambda\delta} \left(QV_{B,k,L}(X) - QV^{\text{tapered}}_{B,k+L,0}(X) \right)$$

$\widehat{\text{ARV}}$ depends on k (short lag), L (long lag), and B.$\frac{\lambda}{\delta} = \frac{L}{k}$. Observations in $(0, \mathcal{T}]$ at equidistant times $0, \mathcal{T}/B, 2\mathcal{T}/B, \cdots, \mathcal{T}$ Assume that we observe X without noise. Because

- Observations may be slightly less frequent due to longer sampling period
- Most noise cancels due to subtraction in (6)
- Have to learn how to walk before learn how to run

Estimation: Decomposition

$$\begin{aligned}
&X_{\left(T_{i+L, T_{i+L+k}}\right]} - X_{(T_{i-k}, T_i]} \\
&= \Theta_{(T_{i+L}, T_{i+L+k}]} - \Theta_{(T_{i-k}, T_i]} + M_{(T_{i+L}, T_{i+L+k}]} - M_{(T_{i-k}, T_i]} \\
&= \underbrace{\left(\Theta'_{(T_{i+L}, T_{i+L+k}]} + M_{(T_{i+L}, T_{i+L+k}]} \right)}_{X'_{(T_{i+L}, T_{i+L+k}]}} + \underbrace{\left(\Theta''_{(T_{i-k}, T_i]} - M_{(T_{i-k}, T_i]} \right)}_{X''_{(T_{i-k}, T_i]}} + \delta \left(\theta_{T_{i+L}} - \theta_{T_i)} \right),
\end{aligned}$$

by using the spot-integral relation for semimartingales:

$$\Theta'_{\left(T_{i+L, T_{i+L+k}}\right]} = \int_{T_{i+L}}^{T_{i+L+k}} (T_{i+L+k} - t)\, d\theta_t$$

$$\Theta''_{(T_{i-k}, T_i]} = \int_{T_{i-k}}^{T_i} (t - T_{i-k})\, d\theta_t$$

$$\Theta_{(T_{i+L}, T_{i+L+k}]} = \Theta'_{(T_{i+L}, T_{i+L+k}]} + \theta_{T_{i+L}} \delta \text{ and } \Theta_{(T_{i-k}, T_i]} = -\Theta''_{(T_{i-k}, T_i]} + \theta_{T_i} \delta$$

Estimation: Construction

Taking the sum squares,

$$
\begin{aligned}
&QV_{B,k,L}(X) \\
&= \underbrace{\frac{1}{k}\sum_{i=k}^{B-k-L}\left(X'_{(T_{i+L},T_{i+L+k}]}\right)^2 + \frac{1}{k}\sum_{i=k}^{B-k-L}\left(X''_{(T_{i-k},T_i]}\right)^2}_{\text{main}_{1,L}=QV^{\text{tapered}}_{B,k+L,0}(X)+O_p(\delta^{1/2}\mathcal{T}^{1/2})} + \underbrace{\frac{\delta^2}{k}\sum_{i=k}^{B-k-L}\left(\theta_{T_{i+L}}-\theta_{T_i}\right)^2}_{\text{main}_{2,L}=(\delta\lambda\mathcal{T})\text{ARV}} \\
&\quad + \underbrace{\tfrac{2}{k}\textstyle\sum_{i=k}^{B-k-L} X'_{(T_{i+L},T_{i+L+k}]}X''_{(T_{i-k},T_i]}}_{\text{cross-term}_{1,L}=O_p(\delta^{1/2}\mathcal{T}^{1/2})} \\
&\quad + \underbrace{\tfrac{2}{k}\textstyle\sum_{i=k}^{B-k-L}\delta\left(\theta_{T_{i+L}}-\theta_{T_i}\right)\left(X'_{(T_{i+L},T_{i+L+k}]}+X''_{(T_{i-k},T_i]}\right)}_{\text{cross-term}_{2,L=O_p\left(\delta\lambda^{1/2}\mathcal{T}^{1/2}\right)}}
\end{aligned}
$$

Conditions for the Asymptotic Results

System

The equation for the system is given by:

$$dX_t = \theta_t dt + \sigma_t dW_t, \quad \text{where} \quad \theta_t \text{ and } \sigma_t^2 \text{ are also continuous Itô semimartingales.}$$

- All coefficients are locally bounded.
- $k \to \infty$
- $k < L$
- $\Delta T \to 0$ and $B \to \infty$
- $\frac{\delta}{T} = \frac{k}{B} \to 0$
- $\frac{\lambda}{T} = \frac{L}{B} \to 0$. But we do not need δ or $\lambda \to 0$
- $\mathcal{T} \to \infty$

Asymptotic Variance of $\widehat{ARV}$ – ARV Martingale

- $\widehat{ARV} - ARV$ = martingale + edge effects
- Asymptotic quadratic variation of martingale:

$$
\begin{aligned}
&(\mathcal{T}\lambda\delta)^{-2}\left(16\delta\int_0^{\mathcal{T}}\left(\langle M,M\rangle'_t\right)^2 dt + 8\delta^2\lambda\int_0^{\mathcal{T}}\left(\langle M,M\rangle'_t\langle\theta,\theta\rangle'_t + \left(\langle M,\theta\rangle'_t\right)^2\right)dt\right) \\
&= \mathcal{T}^{-2}\left(16\lambda^{-2}\delta^{-1}\int_0^{\mathcal{T}}\left(\langle M,M\rangle'_t\right)^2 dt + 8\lambda^{-1}\int_0^{\mathcal{T}}\left(\langle M,M\rangle'_t\langle\theta,\theta\rangle'_t + \left(\langle M,\theta\rangle'_t\right)^2\right).\right.
\end{aligned}
$$

- CLT with given asymptotic variance, stable convergence
- $\widehat{\text{ARV}} - ARV = O_p(T^{-1/2})$ for fixed δ, λ
- Recall $\langle M,M\rangle_t = \sigma_t^2$ is spot volatility of X, $\langle\theta,\theta\rangle'_t$ is spot volatility of θ, etc.
- Minimum asymptotic variance wish no further constraints: δ and λ should be as large as possible
- Something is missing. What is it?

The Dog(s) that Didn't Bark in the Night

So far, we have considered the estimation of:

$$\text{ARV} = \frac{1}{\mathcal{T}}\frac{1}{L}\sum_{i=k}^{B-k-L}\left(\theta_{T_{i+L}} - \theta_{T_i}\right)^2$$

Problems:

- Provides no "discipline" for L and k.
- What if we are instead interested in estimating the average integrated volatility?

$$IV = \frac{1}{\mathcal{T}}\left(\langle\theta,\theta\rangle_{T_{B-k-L}} - \langle\theta,\theta\rangle_{T_k}\right)$$

- "Edge I": Difference between IV and $\frac{1}{\mathcal{T}}\left(\langle\theta,\theta\rangle_{\mathcal{T}} - \langle\theta,\theta\rangle_0\right)$; either additional asymptotic variance term, or adjust summation limits to cover whole interval.
- "Edge II": Edge effect in $\widehat{ARV} - \text{ARV}$: bias and variance terms.

CLT for the Average Realized Volatility of Drift

- Assume also that θ is a martingale. Then

$$\frac{1}{\mathcal{T}}\left(\frac{1}{L}\sum_{i=k}^{B-k-L}\left(\theta_{T_{i+L}} - \theta_{T_i}\right)^2 - \left(\langle\theta,\theta\rangle_{T_{B-k-L}} - \langle\theta,\theta\rangle_{T_k}\right)\right) \sim \mathcal{T}^{-1}\lambda^{1/2}\eta_B Z_B$$

  where $\eta_B^2 = \frac{4}{3}\int_{T_k}^{T_{B-k-L}}\left(\langle\theta,\theta\rangle'_t\right)^2 dt$, and $\lambda = L(\Delta T)$. (Following Theorems 2-3 in ZMA (2005), assuming $B >> 2k + L$.
- Asymptotic variance form of above: $\mathcal{T}^{-2}\lambda\eta_B^2$
- Provides a common sense penalty for large λ, even when θ is a not a martingale
- "Uncertainty principle": If L is large, one can estimate volatility of drift, but cannot precisely locate it in time.

Full Asymptotic Expression in the stable CLT

- Asymptotic Bias:

$$\underbrace{\mathcal{T}^{-1}\left(\frac{1}{4}\langle\theta,\theta\rangle'_{T_B} - \frac{1}{4}\langle\theta,\theta\rangle'_0 + \langle M,\theta\rangle'_{T_B} + \langle M,\theta]'_0\right)}_{\text{from Edge II term, from } \widehat{\text{ARV}}-\text{ARV}}$$
$$\underbrace{-\mathcal{T}^{-1}\left(\delta\langle\theta,\theta\rangle'_0 + (\delta+\lambda)\langle\theta,\theta\rangle'_{\mathcal{T}}\right)}_{\text{Edge I term}}$$

- Asymptotic Variance

$$\underbrace{\mathcal{T}^{-2}\left(16\lambda^{-2}\delta^{-1}\int_0^{\mathcal{T}}\left(\langle M,M\rangle'_t\right)^2 dt + 8\lambda^{-1}\int_0^{\mathcal{T}}\left(\langle M,M\rangle'_t\langle\theta,\theta\rangle'_t + \left(\langle M,\theta\rangle'_t\right)^2\right)dt\right)}_{\text{martingale part of } \widehat{\text{ARV}}-\text{ARV}}$$
$$\underbrace{+\mathcal{T}^{-2}\lambda\frac{4}{3}\int_0^{\mathcal{T}}\left(\langle\theta,\theta\rangle'_t\right)^2}_{\text{ARV} - \text{IV} = \mathcal{T}^{-2}\lambda\eta_B^2} + \underbrace{\frac{1}{4}\mathcal{T}^{-2}\lambda^{-1}\left(\langle\sigma^2,\sigma^2\rangle'_0\right) + \langle\sigma^2,\sigma^2\rangle'_{\mathcal{T}_B}}_{\text{from Edge II}} = O_p\left(\mathcal{T}^{-1}\right)$$

The optimal δ and λ, ignoring Edge Terms

- Minimization of total asymptotic variance from non-edge terms): For given δ, unique minimum in λ :

$$\delta = \frac{1}{\lambda} \frac{32 \int_0^{\mathcal{T}} \left(\langle M, M \rangle_t'\right)^2 dt}{\left(\lambda^2 \eta_B^2 - 8 \int_0^{\mathcal{T}} \left(\langle M, M \rangle_t' \langle \theta, \theta \rangle_t' + \left(\langle M, \theta \rangle_t'\right)^2\right) dt\right)}$$

- Asymptotically, at the same time, $\delta \to 0$ and $\lambda \to \infty$
- To find an optimal δ, may need more precise description of the asymptotic variance. The above expressions are approximations based on assumptions $\delta/\mathcal{T} \to 0$, and $\lambda/\mathcal{T} \to 0$
- Optimization: to be continued...

Data

- Apple trade prices in 2020, from TAQ:
  - 60-second average data
  - 253 trading days
  - Covid impact in US stock market
  - 4-for-1 stock split
- Overnight price adjustment
- Stock split adjustment on August 31

Conclusion

- Attempt to extract price drift, in the form of its variation
- Incorporating HF technique (in-fill asymptotic) in drift analysis allows better estimation of drift variation
- Empirically, can detect drift change over months, even weeks in crisis time
- In higher dimension: Principal Component Analysis?
- Estimation of spot drift itself?

5 On the Theory of Deep Autoencoders [18]

Zhouyu Shen Dacheng Xiu
Chicago Booth and NBER

Autoencoders (AEs)

Autoencoders (AEs), originally proposed by LeCun (1987, PhD thesis), are specialized neural network models designed to replicate inputs at their outputs and are fundamental in unsupervised learning.

The canonical architecture of an AE includes two key components:

- An **encoder**, which compresses the input into a lower-dimensional representation known as features, codes, embedings, or factors.
- A **decoder**, which reconstructs the input from this compressed form.

Unsupervised Learning in Economics and Finance

- We are particularly drawn to AEs due to their close connection with linear factor models and their capability in conducting nonlinear dimension reduction.
- Factor Models in Economics and Finance: Stock and Watson (1999, JASA), Bai and Ng (2002, ECTA), Chamberlain and Rothschild (1983, ECTA), and Connor and Korajczyk (1986, JFE).
- The use of PCA, factor models and matrix completion is becoming increasingly widespread, forecasting, synthetic controls, missing value interpolation...
- It has long been known, e.g., Baldi and Hornik (1989, Neural Network), that a single-layer AE with linear activation is equivalent to PCA.

Motivating Questions

Our paper positions AEs as estimators for nonlinear factor models, and within this framework, we address several key questions to enhance the understanding of deep and nonlinear AEs:

- Can AEs capture the "commonalities" in the inputs, a procedure often referred to as denoising, and if so, what are the statistical error bounds?
- How do AEs' architecture parameters, such as depth, width, and the number of neurons, impact their statistical performance?
- Can AEs recover the hidden low-dimensional representations in a nonlinear factor model?

Related Literature

- Nonlinear Factor Models: Existing methods on nonlinear factor models when the underlying nonlinear functions are unknown still rely on the use of PCA.
- Theory of Neural Networks: Extensive literature spans from Barron (1993, TIT) to recent studies in 2023.

Related Literature

- Feature Learning
- Data Compression
- Noise Reduction
- Generative AI
- Other notable models: Contractive AEs, Convolutional AEs, Masked AEs, Sparse AEs...

Nonlinear Factor Model

We examine the nonlinear factor model introduced by Yalcin and Amemiya (2001, Statist. Sci.),

$$X_{it} = X_{it}^* + U_{it} = \phi(F_{it}) + U_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T.$$

- F_t^*, a K-dimensional vector, represents latent factors.
- $\phi_i^* : \mathbb{R}^K \to \mathbb{R}$, an unknown function, whose functional form can vary across i.
- U_{it} accounts for idiosyncratic noise.

This framework encompasses the classical linear factor model, where $\phi_i^*(F_t^*) = \Lambda_i^\top F_t^*$, with Λ denoting loading matrix.

Boundedness Assumption

- $\text{vec}(U)$ follows the distribution characterized by $\Sigma_U^{1/2}\text{vec}(Z)$, where $Z \in \mathbb{R}^{N\times T}$ consists of independent subGaussian random variables with subGaussian norm bounded by σ_z^2. Moreover, the matrix Σ_U, which is positive semi-definite, has its spectral norm bounded.
  - This assumption holds if $U = \Sigma_1^{1/2} Z \Sigma_2^{1/2}$, where Σ_1 and Σ_2 are positive semi-definite matrices with bounded spectral norms, as is assumed by Onatski (2005, REStat) and Ahn and Horenstein (2013, ECTA).
- A constant $B > 0$ exists such that $\mathbb{P}(\sup_{1\le t\le T} \|F_t^*\|_\infty \le B) = 1$.

Smoothness Assumption

$\varphi_i^\star$ lies in the Hölder ball $\mathcal{H}^\beta(\Omega, B)$ with Ω an open set containing $[-B, B]^K$.

The Hölder ball $\mathcal{H}^\beta(\Omega, B)$ is defined as

$$\left\{ f : \Omega \to \mathbb{R}, \max_{\alpha, |\alpha| \le \lfloor\beta\rfloor} \sup_{x\in\Omega} |D^\alpha f(x)| + \max_{\alpha : \|\alpha\|_1 = \lfloor\beta\rfloor} \sup_{\substack{x, x'\in\Omega \\ x\neq x'}} \frac{|D^\alpha f(x) - D^\alpha f(x')|}{\|x - x'\|^{\beta - \lfloor\beta\rfloor}} \le B \right\}$$

where $\lfloor\beta\rfloor$ represents the largest integer which is strictly smaller than β.

These assumptions are standard in the literature, ensuring that $\varphi_i^\star$ can be well approximated by neural networks.

Deep Neural Networks (DNNs)

The function f of a DNN with architecture parameters (d, w) can be expressed as:

$$f : \mathbb{R}^{n_0} \to \mathbb{R}^{n_{d+1}}, \quad \times \to f(x) = W_d \sigma_{v_d} W_{d-1} \sigma_{v_{d-1}} \cdots W_1 \sigma_{v_1} W_0 \times$$

- W_i represents the weight matrix and v_i the shift (bias) vector at layer i.
- d denotes depth of the DNN, whereas w is its width. n_0 and n_{d+1} represent the dimensions of the input and output.
- $\sigma_x : \mathbb{R}^r \to \mathbb{R}^r$ is the shifted ReLU activation function:

$$\sigma_x \begin{pmatrix} y_1 \\ \vdots \\ y_r \end{pmatrix} = \begin{pmatrix} \max(y_1 - x_1, 0) \\ \vdots \\ \max(y_r - x_r, 0) \end{pmatrix}$$

where $x = (x_1, \dots, x_r) \in \mathbb{R}^r$

Define function class of DNNs

$$\mathcal{F}_{n_0}^{n_{d+1}}(d, w, C, B) := \left\{ f \text{ of the form (1)} : \|f\|_\infty \le B, \max_{j=0,\dots,d} \|W_j\|_\infty, \max_{j=1,\dots,d} \|v_j\|_\infty \le C \right\}$$

Our AEs' Mathematical Formulation

We analyze a special class of AEs that have a disjoint output decoder:

- For any input $N \times 1$ vector X_t, the i-th output of this AE is given by

$$\varphi_i\left(\rho\left(X_t\right)\right), \quad i = 1, 2, \ldots, N$$

  where ρ and φ_i are DNNs.

- Formally, the AE class can be defined as follows:

$$\mathcal{F}_{\mathrm{AE}}^{K_1} := \left\{(\rho, \varphi_1, \ldots, \varphi_N) : \rho \in \mathcal{F}_{K_0}^{K_1}\left(d_1, w_1, T^{5\beta+5}, B\right), \varphi_i \in \mathcal{F}_{K_1}^{1}\left(d_2, w_2, T^{5\beta+5}, B\right)\right\}$$

- K_1 is the pre-selected number of neurons in the AE's bottleneck layer.

Comparing Disjoint-Output and Fully-Connected Decoders

AEs with disjoint output decoders can be considered a sparse-link variant of of the conventional fully connected AEs.

- This sparsity enhances network training efficiency by allowing a separate NN function, $\varphi_i(\cdot)$, to be estimated for each output X_i, thereby reducing the number of parameters that need to be learned.
- The number of neurons is reduced by a factor of N.

  Moreover, we demonstrate that
- Disjoint output decoders can effectively approximate $\varphi_i^\star(\cdot)$.
- Reducing the number of parameters helps to maintain the estimation error well under control uniformly across all is.

Training AEs

Training an AE yields a solution to the following nonlinear least squares problem:

$$(\widehat{\rho}, \widehat{\varphi}_1, \ldots, \widehat{\varphi}_N) = \underset{(W, v, \rho, \varphi_1, \ldots, \varphi_N) \in \mathcal{F}_{AE}^{K_1}}{\arg\min} \sum_{t=1}^{T} \sum_{i=1}^{N} \left(\varphi_i \circ \rho\left(X_t\right) - X_{it}\right)^2$$

As a result,

$$\widehat{X}_{it} = \widehat{\varphi}_i \circ \widehat{\rho}\left(X_t\right)$$

- In practice, to find a desirable solution, we adopt stochastic gradient descent with adaptive learning rates (e.g., RMSprop, Adam, ...)
- The key tuning parameter is K_1, which is closely tied to the architecture of the AE. A scree plot helps illuminate the required number of "linear" factors.

Error Decomposition

$$\sum_{t=1}^{T} \sum_{i=1}^{N} \left(\widehat{X}_{it} - \varphi_i^\star\left(F_t^\star\right)\right)^2 \lesssim \underbrace{\sum_{t=1}^{T} \sum_{i=1}^{N} \left(\varphi_i^\dagger\left(\rho^\dagger\left(X_t\right)\right) - \varphi_i^\star\left(F_t^\star\right)\right)^2}_{\text{Approximation Error}} + \underbrace{\sum_{t=1}^{T} \sum_{i=1}^{N} \left(\varphi_i^\dagger\left(\rho^\dagger\left(X_t\right)\right) - \widehat{X}_{it}\right)^2}_{\text{Estimation Error}}$$

- Approximation Error $\lesssim \mathrm{P}NT\left(T^{-\frac{2\beta}{2\beta+K}} + T^{-1} \inf_{\rho \in \mathcal{F}_0} \sum_{t=1}^{T} \left\|\rho\left(X_t\right) - F_t^\star\right\|^2\right) \log^4(T)$, where $K_1 \geq K$, and $d_2 \asymp \log(T), w_2 \asymp T^{\frac{K}{2(2\beta+K)}}$, and $\mathcal{F}_0 := \mathcal{F}_N^K\left(d_1, w_1, T^{5\beta+5}, B\right)$.

- Estimation Error $\lesssim \mathrm{P}\left(TK_1 + NT^{\frac{K}{2\beta+K}}\right)\log^4(T)$

  $NT^{\frac{K}{2\beta+K}}$ is effectively the total number of parameters in the decoder and TK_1 comes from the estimation error of the bottleneck layer.

AEs' Denoising Performance

Suppose that $K \leq K_1 \leq w_0, d_2 \asymp \log(T), w_2 \asymp T^{\frac{K}{2(2\beta+K)}}$. With probability at least $1 - C\exp(-cT)$, for $\min(N,T)$ sufficiently large,

$$\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\widehat{X}_{it} - \varphi_i^\star\left(F_t^\star\right)\right)^2 \leq \left(T^{-\frac{2\beta}{2\beta+K}} + N^{-1}K_1 + T^{-1}\inf_{\rho\in\mathcal{F}_0}\sum_{t=1}^{T}\left\|\rho\left(X_t\right) - F_t^\star\right\|^2\right)\log^4(T)$$

- The first term in the error, $T^{-\frac{2\beta}{2\beta+K}}$, is the minimax rate as if factors are known. For a nonparametric supervised learning task, this rate can be achieved by estimators other than NNs, see, for example, Peckman (1985, AoS), Newey (1997, JoE), among others.
- The second term comes from the estimation error of the bottleneck layer, which corroborates the results by Bai (2003, ECTA). Choosing more factors than necessary does not negatively affect the convergence rate, demonstrating the model's robustness against overestimation of factor numbers.

Overparameterized Encoder

In considering the third term, we observe that when the encoder's width and depth are overparameterized, the following holds:

$$T^{-1}\inf_{\rho\in\mathcal{F}_1}\sum_{t=1}^{T}\left\|\rho\left(X_t\right) - F_t^\star\right\|^2 = 0$$

- In this case, the encoder $\hat{\rho}$ effectively overfits the data, achieving optimal in-sample performance.
- However, when a new data point X_{T+1} is processed through the encoder, $\hat{\varphi}\left(X_{T+1}\right)$ lacks information about the factor $F_{T+1}^\star$ due to overfitting, which results in poor out-of-sample performance.

Factor Pervasiveness

With an additional assumption on factor pervasiveness, we are able to ensure the out-of-sample performance with a possibly sparse encoder.

- There exists a matrix $W^\star \in \mathbb{R}^{K\times N}$ satisfying $\left\|W^\star\right\|_\infty \leq L^{-1}B$ and $\left\|W^\star\right\|_0 \asymp L$ such that for some fixed constant $c > 0$, the following holds:

$$c\|x-y\| \leq \left\|W^\star\varphi^\star(x) - W^\star\varphi^\star(y)\right\|, \quad \text{for any } x, y \in [-B,B]^K$$

  Intuitively, this assumption guarantees that there exist L variables containing sufficient information for estimating the factors.

  In the context of a linear factor model, where $\varphi^\star\left(F_t\right) = \Lambda F_t^\star$, setting $W^\star = \left(\Lambda^\top\Lambda\right)^{-1}\Lambda^\top$ ensures that $W^\star\varphi^\star(x) = x$, thus satisfying inequality (3).

Properly Parameterized Encoder

Under the aforementioned assumptions, if $K \leq K_1 \leq \min\left(w_1, w_2\right), d_1 \asymp d_2 \asymp \log(T)$, and $w_1 \asymp w_2 \asymp T^{\frac{K}{2(2\beta+K)}}$. Additionally, we assume total number of weights in the encoder is asymptotically bounded by $L + T^{\frac{K}{2\beta+K}}\log T$ and $\log T = \circ(L)$, then with probability at least $1 - C\exp(-cT) - C\exp(-cL)$, as $\min(N,T)$ becomes large enough, we have:

$$\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\hat{X}_{it}-\varphi_i^\star\left(F_t^\star\right)\right)^2 \lesssim \left(N^{-1}K_1+T^{-\frac{2\beta}{2\beta+K}}+L^{-1}\right)\log^4(T),$$

where $\hat{X}_{it} := \hat{\varphi}_i\left(\hat{\rho}\left(X_t\right)\right), c$ and C are constants independent of N, T.
Moreover, when the data is i.i.d., for a new data X_{T+1}, we have

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(\hat{\varphi}_i\circ\hat{\rho}\left(X_{T+1}\right)-\varphi_i^\star\left(F_{T+1}^\star\right)\right)^2 \lesssim \left(N^{-1}K_1+T^{-\frac{2\beta}{2\beta+K}}+L^{-1}+T^{-1}L\right)\log^4(T)$$

As long as $\min\left(N,T,L,L^{-1}T\right)\to\infty$, the out-of-sample error converges to zero.

Feature(Factor) Learning

Factors are identifiable up to invertible transformations. With any injective function $\mu:\mathbb{R}^K\to\mathbb{R}^K$, the DGP can be rewritten as:

$$X_{it}=\varphi_i^\star\circ\mu^{-1}\circ\mu\left(F_t^\star\right)+U_{it}$$

so that $\mu\left(F_t^\star\right)$ can, equivalently, serve as factors.

We show that, for $\widehat{F}_t=\hat{\rho}\left(X_t\right)$, there exists a function $\mu:\mathbb{R}^{K_1}\to\mathbb{R}^K$, composed of ρ and $\hat{\varphi}$, such that with probability at least $1-C\exp(-cT)$, it holds that

$$\frac{1}{T}\sum_{t=1}^{T}\left\|\mu\left(\widehat{F}_t\right)-F_t^\star\right\|_2^2 \le C\left(N^{-1}K_1+T^{-\frac{2\beta}{2\beta+K}}+T^{-1}\inf_{\rho\in\mathcal{F}_0}\sum_{t=1}^{T}\left\|\rho\left(X_t\right)-F_t^\star\right\|^2\right)\log^4(T)$$

Simulated DGPs

We conduct Monte Carlo simulations to assess the comparative performance of AEs and PCA based on the following DGPs:

Linear: $\varphi_i^\star\left(F_t^\star\right)=C\lambda_i^\top F_t^\star$

Nonlinear:

- Model 1: $\varphi_i^\star\left(F_t^\star\right)=C\exp\left(\lambda_i^\top F_t^\star\right)$
- Model 2: $\varphi_i^\star\left(F_t^\star\right)=C\exp\left(-\left\|\lambda_i-F_t^\star\right\|^2\right)$
- Model 3: $\varphi_i^\star\left(F_t^\star\right)=C_1\left(\lambda_{1i}^\top F_t^\star\right)+C_2\left(\lambda_{2i}^\top F_t^\star\right)^2+C_3\left(\lambda_{3i}^\top F_t^\star\right)^3$

Here we set $K=5$, i.e., $F_t^\star\in\mathbb{R}^5$.

- Element-wise, $\lambda_i,\lambda_{1i},\lambda_{2i},\lambda_{3i}\overset{\text{i.i.d.}}{\sim}\mathbb{U}(-1,1), F_t^\star\overset{i.i.d.}{\sim}\mathbb{U}(-2,2), U_t\sim\mathcal{N}(0,1)$.
- Calibrate C,C_1,C_2,C_3 such that the variance of $\varphi_i^\star\left(F_t^\star\right)$ is 1 , and for model 3 , the three individual terms have equal variances.

PCA as Benchmark

PCA is traditionally associated with linear factor models, but recent research indicates its effectiveness on nonlinear factor models.

- Udell and Townsend (2019, SIAM J. Math. Data Sci.) established that large matrices with small spectral norms can be approximated by low-rank matrices.

  Let $X\in\mathbb{R}^{m\times n}$ with $m\ge n$ and $0<\varepsilon<1$. Then, with $r=\lceil 72\log(2n+1)/\varepsilon^2\rceil$.

$$\inf_{\operatorname{rank}(Y)\le r}\|X-Y\|_\infty\le\varepsilon\|X\|_2$$

- In nonlinear factor models with $\varphi_i^\star(F_t^\star) = \varpi(\lambda_i, F_t^\star), \lambda_i, F_t^\star \in \mathbb{R}^K$, Griebel and Harbrecht (2014, IMA J. Numer. Anal.) and Xu (2017, ICML) proved that the matrix $X_{it} := \omega(\lambda_i, F_t^\star)$ allows for a low-rank representation:

  For any $\delta > 0$, with $r \asymp \delta^{-K}$,

$$\inf_{\text{rank}(Y)\leq r} \|X - Y\|_\infty \lesssim \delta^\beta$$

Simulation Details

We conduct 100 simulations and report the average training loss, $N^{-1}T^{-1}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\widehat{X}_{it} - X_{it}\right)^2$, and denoising loss, $N^{-1}T^{-1}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\widehat{X}_{it} - \varphi_i^\star(F_t^\star)\right)^2$, where

- $(N, T) = (50, 500), (200, 500), (200, 50)$
- For PCA, we vary the number of factors from 1 to 20 in increments of two.
- For AE, we consider an architecture with a single hidden layer in the encoder with 20 neurons and K_1 varying from 1 to 20.
  - AE1: Single hidden layer in the decoder with two neurons.
  - AE2: Single hidden layer in the decoder with four neurons.
  - AE3: Single hidden layer in the decoder with eight neurons.
  - AE4: A fully connected decoder based on the model AE3.

Nowcasting GDP Growth

- Dataset: From Giannone et al. (2008, J. Monet. Econ.), focusing on predicting quarterly real GDP growth using 189 US macroeconomic indicators from January 1982 to December 2004.
- We consider a simplified exercise by using only end-of-quarter data and view the GDP to nowcast as missing values.
- Expanding window method is used. First, train AE and PCA on data from Q1 1982 to Q1 1995 and treat GDP growth in Q1 1995 as missing; output $\hat{X}_{NT}$ represents the nowcasting result for this quarter. We then adding one quarter data into the training data and repeat this exercise until Q4 2004.
- Nowcasting Results: Out-of Sample Evaluation. Mean and std of squared error for AE, PCA and Historical Mean's nowcasting results.

Nonlinear Asset Pricing Model

- According to arbitrage pricing theory, asset returns follow a linear factor model:

$$R_{it} = \lambda_i F_t^\star + U_{it}$$

- Borri et al. (2024, arXiv) proposed that excess returns can be modeled as:

$$R_{it} = \varphi(\Lambda_i^\top F_t^\star) + U_{it}$$

  where $\Lambda_i, F_t^\star \in \mathbb{R}^K$ with $K = 1$, and they fit a polynomial function for φ.

- In contrast, we employ AEs to estimate this model. This is achieved by constraining the parameters after the first layer of the decoder to be identical across all outputs and removing the bias in the first layer, ensuring that the output takes the form $\varphi(\Lambda_i^\top F_t^\star)$.

  By using AEs, we can handle cases where $K > 1$ and estimate more complex models beyond polynomial functions.

Cross-Sectional Asset Pricing

- Dataset: Long-short portfolios from Chen and Zimmermann (2022, CFR), covering the period from January 1970 to December 2022 (N = 140, $T = 636$).
- Using the estimator $\widehat{R}_{it}$ obtained from both PCA and AEs, we conduct a cross-sectional regression by regressing the time-series average of realized excess returns on the time-series average of estimated returns: $T^{-1}\Sigma_{t=1}^{T} R_{it} \sim T^{-1}\Sigma_{t=1}^{T} \widehat{R}_{it}$.
- Under the null hypothesis $\mathbb{E}[R_{it}] = \mathbb{E}[\varphi(\Lambda_i^\top F_t^\star)]$, the regression R^2 should approach one, and the regression intercept should be close to zero, indicating no mispricing.

Denoising Measurement Error

- Dataset: Based on the seminal paper by Autor et al. (2013, AER). Dataset at the commuting zone (CZ) level with 722 CZs , each characterized by 30 covariates.
- We manually add Gaussian noise manually into the data X with noise-to-signal ratios of $0.5, 1$, and 2 , applying AE and PCA and calculating the MSE between the original data X and the denoised output $\widehat{X}$.

Conclusion

- AEs hold promise for nonlinear dimension reduction.
- We provide non-asymptotic analysis to AEs' denoising errors and factor learning errors, which achieves the optimal rate for nonparametric regression under mild conditions.
- Through simulations and empirical illustrations, AEs provide superior performance compared with PCA under nonlinear factor models.

6 Arbitrage Pricing Theory or AI Pricing Theory? The Surprising Dominance of Large Factor Models [7]

Antoine Didisheim U. Melbourne Shikun (Barry) Ke Bryan Kelly Semyon Malamud Yale 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 clashes with massive parameterizations adopted by modern ML algorithms

- Leading edge GPT-4 language model uses >1 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-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
- Small models may be good for some things, but not for optimizing out-of-sample performance

The "Virtue of Complexity" in Asset Pricing

Building the "Case" for Financial ML

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

"Virtue of Complexity in Return Prediction" (Kelly, Malamud, Zhou, J. of Finance 2024)

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

This Paper: ML in Cross-sectional Asset Pricing

- Main theoretical result: SDF 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: History

SDF representable as managed portfolios: $M_{t+1}^{\star} = 1 - \sum_{i=1}^{n} w(X_t)' R_{i,t+1}$, s.t. $E_t\left[M_{t+1}^{\star} R_{i,t+1}\right] = 0 \forall i$

- Cross-sectional asset pricing is about $w_t = w(X_t)$
- Fundamental challenge in cross-sectional asset pricing: w must be estimated
  This is a high-dimensional (complex) problem
- Standard approach: Restrict w's functional form
  - E.g., Fama-French: $w_{i,t} = b_0 + b_1$ Size $_{i,t} + b_2$ Value $_{i,t}$ (Brandt et al. 2007 generalize)
  - Reduces parameters, implies factor model: $M_{t+1} = 1 - b_0 MKT - b_1 SMB - b_2 HML$
  - "Shrinking the cross-section" Kozak et al. (2020) - use a few PCs of anomaly factors
  - The role of theory

Complexity in the Cross Section: Machine Learning Perspective

SDF representable as $M_{t+1}^{\star} = 1 - \sum_{i=1}^{n} w(X_t)' R_{i,t+1}$, s.t. $E_t\left[M_{t+1}^{\star} R_{i,t+1}\right] = 0 \forall i$
Rather than restricting $w(X_t)$ ...

- ...expand parameterization, saturate with conditioning information
- E.g. approximation via neural network: $w\left(X_{i,t}\right) \approx \lambda' S_{i,t}$, where $P \times 1$ vector $S_{i,t}$ is known nonlinear function of original predictors $X_{i,t}$
- Implies that empirical SDF is a high-dimensional factor model with factors F_{t+1}

$$M_{t+1}^{\star} \approx M_{t+1} = 1 - \lambda' \underbrace{S_t' R_{t+1}}_{=F_{t+1} \in \mathbb{R}^{p \times 1}} = 1 - \lambda' F_{t+1}$$

Theory Environment

- n assets with returns R_{t+1}
- Empirical SDF $M_{t+1} = 1 - \lambda' S_t' R_{t+1}$
  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

$$\frac{\text{Max Sharpe Ratio}}{\min_{\lambda} E\left[\left(1 - \lambda' S_t' R_{t+1}\right)^2\right] + z\lambda'\lambda} \quad \text{or} \quad \frac{\text{Min Pricing Error (HJ-distance)}}{\min_{\lambda} E[MF]' E\left[FF'\right]^{-1} E[MF] + z\lambda'\lambda}$$

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

Main Theorem

In the limit as $N, P, T \to \infty, P/T \to c, P/P^* \to q$, the expected out-of-sample moments of the ridge SDF portfolio satisfy

- i. $\lim E\left[\hat{R}^M_{T+1}(z;q;c)\right] = \mathcal{E}\left(Z^*(z;q;c);q\right)$ where

$$Z^*(z;q;c) = z(1+\xi(z;q;c)) \in (z, z+c) \text{ and } \xi(z;q;c) = \frac{c(1-m(-z;c;q)z)}{1-c(1-m(-z;c;q)z)}$$

- ii. $\lim \text{Var}\left[\hat{R}^M_{T+1}(z;q;c)\right] = \mathcal{V}\left(Z^*(z;q;c);q\right) + G(z;q;c)\mathcal{R}\left(Z^*(z;q;c);q\right)$ where $G(z;q;c) = (z\xi(z;q;c))' \in (0, cz^{-2}]$, $\mathcal{R}(z;q) \equiv (1-\mathcal{E}(z;q))^2 + \mathcal{V}(z;q)$
- iii. $\lim \frac{\text{Var}\left[\hat{R}^M_{T+1}(z;q;c)\right]}{E\left[\hat{R}^M_{T+1}(z;q;c)\right]^2} = (1+G(z;q;c))\frac{1}{\mathcal{S}^2(Z^*;q)} + G(z;q;c)\left(\frac{1-\mathcal{E}(Z^*;q)}{\mathcal{E}(Z^*;q)}\right)^2$, where $\mathcal{S}^2(z;q) = \mathcal{E}^2(z;q)/\mathcal{V}(z;q)$
- iv. $\lim E\left[\mathcal{D}^{HJ}_{OS}(z;q;P;T) - \bar{F}'_{OS}B^+_{OS}\bar{F}_{OS}\right] = (1+G(z;q))\mathcal{R}\left(Z^*(z;q;c);q\right)$.

Empirical Analysis

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

Random Fourier Features

- Empirical model: $M_{t+1} = 1 - \lambda' S'_t R_{t+1}$
- 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

$$S_{\ell,i,t} = \left[\sin\left(\gamma'_\ell X_{i,t}\right), \cos\left(\gamma'_\ell X_{i,t}\right)\right], \quad \gamma_\ell \sim \text{iidN}(0, \gamma I)$$

  $S_{\ell,i,t}$: 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

Training and Testing

- We estimate out-of-sample SDF with:
  i. Thirty-year rolling training window ($T = 360$)
  ii. Various shrinkage levels, $\log_{10}(z) = -12, \dots, 3$
  iii. Various complexity levels $P = 10^2, \dots, 10^6$
- 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

7 Nonparametric Expected Shortfall Regression with Tail-robustness [19]

Given $\tau \in (0,1)$, the standard QR estimator is

$$\hat{\beta}(\tau) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \left(Y_i - X_i^T \beta\right)$$

where $\rho_\tau(u) = u\{\tau - \mathbb{1}(u < 0)\}$

- Robustness against heavy-tailed response-median regression
- Capture heterogeneity in the set of important predictors at different quantile levels of the response distribution

Quantile vs. Expected Shortfall

- Value-at-Risk (VaR, a quantile of the loss distribution) is the most widely (mis-)used risk measure in the financial industry, not a coherent risk measure. Risk can be reduced by diversification, which violates sub-additivity.
- Coherent measures: expectile (EVaR), expected shortfall (ES, CVaR)
- Applications: risk modeling, portfolio optimization using CVaR
  Let $q_\tau(Z) = F_Z^{-1}(\tau)$ be the τ-quantile of Z The expected shortfall at level τ is

$$\frac{1}{\tau} \int_0^\tau q_u(Z) \mathrm{d}u$$

or

$$\mathbb{E}\left\{Z \mid Z \leq q_\tau(Z)\right\}$$

Challenges in Estimation

- Quantiles are easier to estimate
- Sample quantiles are robust to outliers and heavy-tailedness
  Elicitability: A statistical functional is said to be "elicitable" if there exists a loss function such that it is the solution to minimizing the expected loss
  - Mean: $\mu = \operatorname{argmin}_{u \in \mathbb{R}} \mathbb{E}(Y-u)^2$
  - Quantile: $q_\tau = \operatorname{argmin}_{u \in \mathbb{R}} \mathbb{E}\rho_\tau(Y-u)$
  - Expectile : $e_\tau = \operatorname{argmin}_{u \in \mathbb{R}} \mathbb{E}\left|\tau - 1_{\{Y \leq u\}}\right|(Y-u)^2$
- ES by itself is NOT elicitable Quantile and ES are jointly elicitable

Estimation via a Joint Loss Function

Fissler & Ziegel (2016) proposed a class of loss functions

$$\begin{aligned}\rho(q,e;y) =& \{\tau - 1(y \leq q)\}\{g_1(y) - g_1(q)\} \\ & - g_2(e)\left\{\frac{1}{\tau}(y-q)1(y \leq q) + q - e\right\} - G_2(e)\end{aligned}$$

where $g_2 = G_2'$, g_1 is increasing, G_2 is increasing & convex.
(q_τ, e_τ) is a global minimum of $(q,e) \mapsto \mathbb{E}\rho(q,e;Y)$ on the domain $A_0 = \{(x_1,x_2) \in \mathbb{R}^2 : x_1 \geq x_2\}$
(q_τ, e_τ) is the unique minimum if G_2 is strictly increasing & convex.

Joint Quantile/ES Regression

Dimitriadis & Bayer (2019) and Patton et al. (2019) studied a joint regression framework for the quantile and ES:

$$q_\tau(Y \mid X) = X^T\beta^* \quad e_\tau(Y \mid X) = X^T\theta^*$$

where $\beta^*, \theta^* \in \mathbb{R}^p$ are unknown vectors of regression coefficients.
Joint M-estimator: $\underset{\beta,\theta\in\Theta}{\text{minimize}} L_n(\beta,\theta) := \frac{1}{n}\sum_{i=1}^n \rho\left(X_i^T\beta, X_i^T\theta; Y_i\right)$
where $\Theta \subseteq \mathbb{R}^p$ is a compact subset.
Limitations:

- $(\beta,\theta) \mapsto L_n(\beta,\theta)$ is non-convex and non-differentiable
- DB (2019) used the Nelder-Mead simplex algorithm: a heuristic search method without convergence guarantees
- R package 'esreg' is not scalable for large-scale data

Two-step Procedures

Define surrogate response variables

$$Z_i(\beta) = \tau^{-1} \underbrace{\left(Y_i - X_i^T\beta\right) 1_{\{Y_i \le X_i^T\beta\}}}_{\text{negative part of QR residual}} + X_i^T\beta,$$

satisfying

$$\mathbb{E}\left\{Z_i\left(\beta^*\right) \mid X_i\right\} = X_i^T\theta^*$$

Using generated surrogate response $\hat{Z}_i := Z_i(\hat{\beta})$, Barendse (2020) proposed the two-step least squares estimator (2S-LSE)

$$\hat{\theta}^{\text{ls}} = \text{argmin}_\theta \sum_{i=1}^n \left(\hat{Z}_i - X_i^T\theta\right)^2$$

and established consistency and asymptotic normality when p is fixed.

Pros and Cons

- Pro: The two-step ES estimator constructed from surrogate response variables is "robust" against first-stage QR estimation error.
- Drawback: If Y_i is heavy-tailed, the distribution of $\psi\left(\beta^*, \theta^*; X_i\right)$ is also heavy-tailed and, at the same time, skewed. The LSE performs poorly in the presence of asymmetry and heavy-tailedness.
- Further connections: The use of surrogate responses is closely related to the influence function approach from robust statistics (Chetverikov, Liu & Tsyvinski, 2022).

Nonparametric Qt-ES Regression

$$q_\tau(Y \mid X) = f_0(X) \quad e_\tau(Y \mid X) = g_0(X)$$

Nonlinear/nonparametric
Nadaraya-Watson & local polynomial regression
Series methods Regression trees and ensemble methods
Reproducing kernel Hilbert space (RKHS) regression
Deep neural network regression

RKHS

Fix a compact space $\mathcal{X}$, such as $[0,1]^p$

Positive Semidefinite Kernel

$K(\cdot,\cdot): X \times X \to \mathbb{R}$ is symmetric and PSD : for any $n \geq 1$ and $\{x_i\}_{i=1}^n \subseteq \mathcal{X}, \left(K_{ij} = K\left(x_i, x_j\right)\right)_{1 \leq i,j \leq n}$ is positive semidefinite.

Reproducing Kernel Hilbert Space

RKHS $\mathcal{H} \subseteq L_2(\mathcal{X})$ associated to a PSD kernel K with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the unique Hilbert space satisfying

- For any $x \in \mathcal{X}$, the function $K(\cdot, x): \mathcal{X} \to \mathbb{R}$ belongs to $\mathcal{H}$
- Kernel reproducing property:

$$\langle f, K(\cdot, x) \rangle_{\mathcal{H}} = f(x), \quad \forall f \in \mathcal{H}$$

Kernel Trick

Fitting linear models depends on inner product $\langle x, x' \rangle$ in $\mathbb{R}^d$

Mercer's theorem

There exist eigenvalues $\mu_j \geq 0$ and eigenfunctions ϕ_j-an orthonormal basis of $L_2(\mathcal{X}, \mathbb{P})$ - of the integral operator

$$T_K(f)(x) = \int_{\mathcal{X}} K\left(x, x'\right) f\left(x'\right) \mathrm{d}\mathbb{P}\left(x'\right)$$

such that $T_K\left(\phi_j\right) = \mu_j \phi_j$ and $K\left(x, x'\right) = \sum_{j=1}^{\infty} \mu_j \phi_j(x) \phi_j\left(x'\right)$

Feature map

$\Phi: \mathcal{X} \to \ell_2(\mathbb{N})$ embeds d-variate vectors into a subset of $\ell_2(\mathbb{N})$:

$$\Phi(x) = (\sqrt{\mu_1}\phi_1(x), \sqrt{\mu_2}\phi_2(x), ...)$$

$$\langle \Phi(x), \Phi\left(x'\right) \rangle_{\ell_2(\mathbb{N})} = \sum_{j=1}^{\infty} \mu_j \phi_j(x) \phi_j\left(x'\right) = K\left(x, x'\right)$$

Kernel Ridge Regression

Given $\{(Y_i, X_i)\}_{i=1}^n$ with $X_i \in \mathcal{X}, Y_i \in \mathbb{R}$,

$$\hat{f} \in \operatorname{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(Y_i - f\left(X_i\right)\right) + \lambda \|f\|_{\mathcal{H}}^2$$

The regularization parameter λ controls bias/variance trade-off:
$\lambda \uparrow \Rightarrow$ encourages $\hat{f}$ to have smaller $\|\hat{f}\|_{\mathcal{P}}$, or smoother
$\lambda \downarrow \Rightarrow$ $\hat{f}$ fits the data better (less bias) at the cost of smoothness

Kernel Ridge Quantile Regression

In the first step, estimate the conditional quantile function f_0 using kernel ridge regression:

$$\hat{f} = \hat{f}_n(\lambda_q) = \text{argmin}_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho_\tau (Y_i - f(X_i)) + \lambda_q \|f\|_{\mathcal{H}}^2 \right\},$$

where λ_q is a regularization parameter

Equivalent form

By the representer theorem (Kimeldorf & Wahba, 1971), the KRR estimator can be expressed as $\hat{f}(\cdot) = \sum_{j=1}^{n} \hat{\alpha}_j K(\cdot, X_j)$

then we get

$$\min_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \left(Y_i - \sum_{j=1}^{n} \alpha_j K(X_i, X_j) \right) + \lambda_q \alpha^T \mathbf{K} \alpha \right.$$

where $\mathbf{K} = \left(K(X_i, X_j)\right)_{1 \le i,j \le n} \in \mathbb{R}^{n \times n}$ is the kernel matrix.

Kernel Ridge ES Regression

Construct surrogate response variables $\hat{Z}_i = Z_i(\hat{f})$ by plugging-in the Qt-KKR estimate:

$$Z_i(f) = \tau^{-1} \min \{Y_i - f(X_i), 0\} + f(X_i)$$

Then, a two-step ES-KRR estimator $\hat{g}$ is defined as

$$\hat{g} = \hat{g}_n\left(\lambda_e; \hat{f}\right) \in \text{argmin}_{g \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left\{ Z_i(\hat{f}) - g(X_i) \right\}^2 + \lambda_e \|g\|_{\mathcal{H}}^2 \right\}$$

we get

$$\min_{\mathbf{a} \in \mathbb{R}^n} \left[\frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{Z}_i - \sum_{j=1}^{n} \alpha_j K(X_i, X_j) \right\}^2 + \lambda_e \alpha^T \mathbf{K} \alpha \right]$$

Statistical Theory: Yu et al. (2024a)

Pointwise Inference with Bootstrap

From training data $\mathcal{D}_n = \{(Y_i, X_i)\}_{i=1}^n$ compute KRR estimators

$$\hat{f} = \hat{f}_n(\lambda_q), \quad \hat{g} = \hat{g}_n(\lambda_e)$$

Generate i.i.d. random weights $W_1, W_2, \ldots, W_n$, independent of $\mathcal{D}_n$, satisfying

$$\mathbb{E}(W_i) = 1, \quad \text{Var}(W_i) = 1$$

The bootstrap ES-KRR estimator $\hat{g}^b = \hat{g}_n^b(\lambda_e)$ is defined as

$$\hat{g}^b \in \text{argmin}_{g \in \mathcal{C}} \left[\frac{1}{n} \sum_{i=1}^{n} W_i \left\{ Z_i(\hat{f}) - g(X_i) \right\}^2 + \lambda_e \|g\|_{\mathcal{H}}^2 \right]$$

- $\hat{g}^b$ can be viewed as a bootstrap estimate of $\hat{g}$
- Construct bootstrap confidence intervals using $\{\hat{g}_b^b\}_{b=1}^B$

Computational Costs

Fast construction of confidence intervals

△ Computational complexity of fitting one KRR (exactly) is $\mathcal{O}(n^3)$, and thus increases to $\mathcal{O}(Bn^3)$ for the bootstrap procedures

Approximate algorithms include

- Divide-and-conquer approach (Zhang et al., 2013)
- Nyström method (Williams and Seeger, 2000)
- Randomized sketches (Yang et al., 2017)
- Multi-layer kernel machine (Dai & Zhong, 2024)

These algorithms require tuning parameters in addition to λ. The computational cost remains high when B is large.

Fast Bootstrap Inference

Motivated by the Bahadur representation presented later, we bypass solving the quadratic program by approximating the distribution of $\hat{g}(x_0) - g_0(x_0)$ with the linear statistic

$$\mathfrak{B}^b(x_0) := \frac{1}{n}\sum_{i=1}^{n}(W_i - 1)\left\{Z_i(\hat{f}) - \hat{g}(X_i)\right\}\left(\hat{T} + \lambda_e I\right)^{-1} K_{X_i}(x_0)$$

where $K_{X_i}(\cdot) = K(X_i, \cdot)$, $\hat{T} = (1/n)\sum_{i=1}^{n} K_{X_i} \otimes K_{X_i}$, and

$$\left(K_{X_i} \otimes K_{X_i}\right)(h) := \left\langle K_{X_i}, h\right\rangle_{\mathcal{P}} K_{X_i} = h(X_i) K_{X_i}$$

In practice, it suffices to calculate

$$\mathbf{v}(x_0) = (\mathbf{K}/n + \lambda_e I)^{-1}\mathbf{k}(x_0) \in \mathbb{R}^n$$

where $\mathbf{k}(x_0) = (K(X_1, x_0), \ldots, K(X_n, x_0))^T \in \mathbb{R}^n$. Then

$$\mathfrak{B}^b(x_0) = \frac{1}{n}\sum_{i=1}^{n}(W_i - 1)\left\{Z_i(\hat{f}) - \hat{g}(X_i)\right\} v_i(x_0)$$

8 Perceived Shocks and Impulse Responses [15]

Raffaella Giacomini UCL Jason Lu IMF Katja Smetanina University of Chicago Booth School of Business

Idea

- Data requirement: a panel of expectation revisions for one variable across different forecast horizons and over time
- Econometric method: fit a time-varying factor model and recover latent factors (shocks) and loadings (impulse responses) at every point in time
- Relies on "weak" assumptions:

  Only uses expectation data so no need to make assumptions about agents' information sets, models, rationality

  Factor structure compatible with several existing theories of expectation formation

Perceived IRFs

- Novel empirical measure revealing beliefs about shock persistence/shape of IRF
- Of direct interest to central banks (e.g., anchoring of long-run inflation expectations - e.g., Carvalho, Eusepi, Moench, Preston, 2023)
- Could help document new stylized facts (e.g. for theories where beliefs about persistence of shocks drive aggregate fluctuations, e.g. Blanchard, l'Huillier and Lorenzoni, 2013)
- Goes beyond the dicothomy permanent/transitory shocks considered in several literatures (e.g., Stock and Watson, 2007 for inflation, classic literature on income shocks and consumption), which may be too restrictive in some cases (as we will see for inflation)

Perceived Shocks

Related to large literature on high-frequency measures of narrative shocks using market-based measures of expectations. Factor extraction across different measures already in Gurkaynak, Sack and Swanson (2005).

Differences:

- Leverages forecast revisions of one variable across different horizons instead of market-based measures (known to be contaminated by, e.g., risk premia)
- Formalizes econometric procedure and lets the data tell us how many shocks agents react to (speaks to literature on understanding number of shocks, e.g. Angeletos, Collard and Dellas, 2020)
- Does not impose parameter stability (analysis is "local in time")
- Caveat: as in existing literature, identifying specific structural shocks would require additional assumptions (e.g., restrictions on loadings)

Data

Need a panel data of forecast revisions X_{ht} for a term structure of horizons $h = 1, \dots, H$ and times $t = 1, \dots, T$. Revisions constructed as follows:

- t is the frequency at which the forecasts are produced (say monthly)
- At each t consider a target variable for a set of future horizons $Y_{h,t}, h = 1, \dots, H$
  - Can have mixed frequency, i.e. monthly forecasts of quarterly inflation

  - Can have unequally spaced horizons
  - Can mix different types of forecasts, as long as they embed different horizons (e.g., for inflation: nowcast, 1-quarter ahead forecast, 5 -year 5 -year rate)

- Denote forecast made at time t for this future variable by $\widehat{Y}_{h|t}$
- Construct the forecast revisions as difference in forecasts of the same variable made in two consecutive months: $X_{ht} = \widehat{Y}_{h|t} - \widehat{Y}_{h|t-1}$

Model

Idea is to model the term structure of forecast revisions as a factor model, allowing for time-varying loadings:

$$X_{ht} = \lambda_{ht}' F_t + e_{ht}$$

where F_t is a vector of $r < N$ latent factors

- e_{ht} is noise
- F_t are the perceived shocks
- λ_{ht} are the perceived impulse responses to each shock at each horizon h and at each point in time t
- More on the specific assumptions later....

Normalizations

Factors and loadings are only identified up to scale, so need to impose normalization restrictions

- If the focus of the analysis is the IRF, normalize the factors → loadings represent IRFs to a unit-standard-deviation shock
- Depending on the application, can also normalize the shocks to have, e.g., unit impact
- If the focus is recovering the shocks, normalize the loadings so we can interpret the scale of the shocks

Econometric Challenges

- Small H ($\approx 6 - 8$)
- Heteroskedastic noise → makes PCA estimator invalid in finite H (Bai and Wang, 2016)
- We thus want a PCA method that works in finite H and that allows for heteroskedastic noise
- Large-T needed if we want time-varying loadings, otherwise assume constant loadings
- Under parametric assumptions, one could alternatively use ML/Bayesian methods for small H (heteroskedasticity also a problem)
- Here we want to keep things nonparametric, both in using PCA and in modeling time variation

Solution

- We consider an approach for PCA estimation in finite samples under heteroskedastic noise from the matrix completion statistics literature (Zhang, Cai and Wu, 2022)
- Extend approach to time-varying loadings
- Cost: assume noise is independent across h (whereas in large-H, PCA allows for weak dependence)
- Clunky name: time-varying HeteroPCA

Why the factor structure?

- We now show that the factor structure for expectation revisions is compatible with different theories of expectation formation:

  Rational expectations

  Information rigidities (noisy information)

- In all following cases, easy to see how time-varying parameters give rise to time-varying loadings

Rational expectations and Vector Moving Average model

- Suppose agents use a correctly specified structural VMA models (e.g., Plagborg-Moller, 2019) to forecast one of the variables in the system, Y_t :

$$Y_t = \Theta(L)\varepsilon_t$$

  where ε_t is a vector of structural shocks and $\Theta(L)$ the lag polynomial

- The forecast revision between times $t-1$ and t is

$$X_{ht} = \underset{\tilde{\lambda}'_{ht}}{\theta'_h} \underset{\tilde{F}_t}{\varepsilon_t}$$

- Implies a factor structure with no noise, where perceived IRF = true IRF and perceived shocks = structural shocks

- Note: ability to recover multiple shocks from forecast revisions of only one variable (if $n.$ of horizons $> n.$ of shocks)

- If agents forecast different variables in the survey, can use method to understand if agents react to same shocks across different variables

- Under assumption that agents use same model to forecast multiple variables, can exploit this third dimension to obtain more precise estimates of shocks

Rational expectations and factor model

- Common to forecast using factor models (e.g., Stock and Watson, 2002)

- E.g., affine term structure or Nelson Siegel models for interest rates (typically assuming three dynamic factors β_t) implies that for one maturity we have

$$Y_t = \gamma'\beta_t + v_t$$
$$\beta_t = \Phi\beta_{t-1} + \varepsilon_t$$

- The forecast revision between times $t-1$ and t is

$$X_{ht} = \underbrace{\gamma'\phi^h}_{\lambda'_{ht}} \underset{\tilde{F}_t}{\varepsilon_t}$$

- Implies a factor structure with no noise, where perceived IRF = true IRF and perceived shocks = surprises from the state equation

Information rigidities

- E.g., noisy information model of Coibion and Gorodnichenko (2015), where target variable is an AR(1)

$$Y_t = \rho Y_{t-1} + \varepsilon_t$$

  with ε_t Gaussian white noise

- Agent observes a noisy signal

$$Z_t = Y_t + v_t$$

  with v_t Gaussian white noise independent of ε_t

- Agent forecasts using the Kalman filter:

$$\widehat{Y}_{t+h|t} = \rho^h \widehat{Y}_{t|t}$$
$$\widehat{Y}_{t|t} = GZ_t + (1-G)\widehat{Y}_{t|t-1}$$

  where the Kalman gain G captures the degree of information rigidity

- Forecast revision is

$$X_{ht} = \widehat{Y}_{t+h|t} - \widehat{Y}_{t+h|t-1} = \underbrace{\rho^h}_{\lambda_{ht}} \underbrace{\left(\widehat{Y}_{t|t} - \widehat{Y}_{t|t-1}\right)}_{F_t}$$

- The (single) factor is $F_t = \widehat{Y}_{t|t} - \widehat{Y}_{t|t-1} = G\left(Z_t - \widehat{Y}_{t|t-1}\right)$, the surprise from the Kalman filter updating equation

- Implies factor structure with no noise, where perceived IRF = true IRF and perceived shock = filtered shock

Noise

- Noise can be due to various reasons:
  - When considering consensus forecasts (as we do), changing composition in survey participation across months and horizons introduces noise in revisions
  - Rounding of forecasts
  - Horizon-specific biases/adjustments to model-based forecasts
  - Mixing different types of forecasts or different surveys for different horizons
- Arguably independent across horizons and heteroskedastic

Finite-sample method with const loadings: HeteroPCA

Zhang, Cai and Wu (2022) proposed a method, HeteroPCA, for estimating factor models under heteroskedasticity in finite H and T. They assume constant loadings

- Problem: heteroskedastic noise means that eigenvectors of sample covariance of X_t can be very different from F_t. Issue is with the diagonal (because of independent noise assumption)
- Solution: iteratively impute the diagonal of the sample covariance by the diagonal of its low-rank approximation to improve accuracy
- In what sense does this work? They characterize the minimax upper and lower bounds of the estimation error (using perturbation analysis) and prove optimal rate of convergence + good finite sample performance

Time-varying HeteroPCA

- We extend this finite-sample procedure to allow for time-varying loadings
- We model time variation in loadings as deterministic functions of time, which we then estimate nonparametrically:

$$\lambda_{ht} = \lambda_h(t/T), \quad t = 1, \ldots, T$$

where, for each $h, \lambda_h(\cdot) : [0,1] \to \mathbb{R}$ is an unknown piece-wise smooth function of the rescaled time t/T

- Assumption is common in nonparametric estimation (e.g. Robinson, 1989, Cai, 2007) and is necessary to achieve consistency
- X_{ht} is thus locally stationary in the sense of Dahlhaus et al. (2019)
- Su and Wang (2017), among others, allow for time-varying loadings of this form, but require large H

Technical assumption n.2

Assumption ("Incoherence" $\approx$ strong factors) Let: $\mathbb{O}_{H,r} := \left\{U_t \in \mathbb{R}^{H\times r} : U_t' U_t = I_r\right\}$ be the set of matrices with orthonormal columns; $\widetilde{e}_i$ be the i-th standard basis with i-th coordinate equal to 1 and other coordinates 0; $\lambda_r(\Lambda_t)$ be the smallest nonzero eigenvalue of $\mathrm{Cov}(\Lambda_t' F_t)$; $\sigma_r(\Lambda_t)$ be the r-th diagonal entry of Λ_t; $T \geq Cr, \lfloor Th \rfloor \wedge H \geq C\left(\sigma_{t,\,\mathrm{sum}}^2 / \sigma_r(\Lambda_t)\right)$, where $\sigma_{t,\,\mathrm{sum}}^2 = \sum_{h=1}^{H} \sigma_{ht}^2$. Let $\|\Lambda_t\| / \lambda_r(\Lambda_t) \leq C_0$ for some constant $C_0 > 0$. For any $\cup_t \in \mathbb{O}_{H,r}$, there exists a constant c_{I} such that for $t = 1, \ldots, T$:

$$I(U_t) := \frac{H}{r} \max_{1\leq h\leq H} \left\|\widetilde{e}_h' U_t\right\|_2^2 \leq c_l H/r$$

Standard in the matrix completion literature (also in microeconometrics, e.g., Agarwal and Singh, 2024). Similar to assuming strong factors

Practical implication: rules out a large proportion of/close to zero loadings. Caution if forecasts don't vary much/infrequent updating + small bandwidth

Overview of application: Perceived shocks and IRFs of Inflation

- We apply the method to extract historical perceived shocks and IRF of inflation
- We consider aggregated consensus expectations
- Summary of findings:

  (1) Agents react to one shock, highly correlated with inflation surprises

  (2) Secular decrease in the perceived persistence of the shock

  (3) Recent movements in long-run expectations due to large perceived shocks, not "deanchoring"

Data

- We want CPI inflation expectations for a term structure of forecast horizons that are produced monthly.
- We merge two data sources: Blue Chip (BC) and Cleveland Fed (CF):
  - Short- and medium horizons: BC forecast quarterly inflation for current and next year (max. n. of usable horizons: nowcast +4 quarters of next year)
  - BC long-run inflation expectations infrequent $\to$ use monthly expectations from CF (can be viewed as interpolating BC surveys with high-frequency data)

    - Mixing of data sources to construct a term structure of expectations (not revisions) also in Aruoba (2014), to answer different question
    - Key advantage: timing of the CF and BC forecasts essentially aligned

- Our term structure thus has H = 7 points: nowcast, 1- to 4-quarters ahead from BC and 2-year 3-year and 5-year 5-year from CF
- Main findings (one shock + secular decrease in persistence of the shock) robust to dropping CF

Results

Time-varying impulse response

- Clear evidence of time variation in loadings
- IRF at a particular point in time is a vertical slice from the figure. Shape of IRF also changes over time

IRF at 3 particular points in time

The unit-standard deviation normalization has too many moving parts to compare IRFs across time (standard deviation changes over time)

IRF under unit-impact normalization

- Consider different normalization of IRFs: unit-impact
- Perceived persistence of shock has decreased over time, including recently

Deanchoring or large shocks?

Volker disinflation (smaller shocks + deanchoring) vs. recent high-inflation episode (large shocks + anchoring)

9 Deep Reinforcement Learning for Games with Controlled Jump-diffusion Dynamics [6]

Ruimeng Hu, UCSB

Motivation

- Financial markets exhibit unpredictable shocks and volatility
- Traditional models like Brownian motions fail to capture sudden market jumps
- Lévy models are capable of capturing jumps of varied sizes and frequencies

Control and Games under Lévy Models

Contribution

- We develop efficient ML algorithms for solving control and games
- We analyze portfolio games in jump diffusion markets and derive semi-explicit solutions

Start with Generic Stochastic Control with Jumps

- State process X_t follows a controlled Itô-Lévy process:

$$\mathrm{d}X_t = b\left(X_{t-}, u_t\right)\mathrm{d}t + \sigma\left(X_{t-}, u_t\right)\mathrm{d}W_t + \int_{\mathbb{R}^d} G\left(X_{t-}, z, u_t\right)\tilde{N}(\ \mathrm{d}t,\ \mathrm{d}z)$$

- Components:
  - u_t : control process
  - W_t : standard Brownian motion
  - N : Poisson random measure with the Lévy measure ν
  - $\tilde{N}(\ \mathrm{d}t,\ \mathrm{d}z) = N(\ \mathrm{d}t,\ \mathrm{d}z) - \nu(\mathrm{d}z)$: compensated Poisson random measure

Dynamics Programming and Value Function

Optimal Value Function:

$$v(t,x) = \sup_{u\in U} J^u(t,x), \quad J^u(t,x) = \mathbb{E}\left[\int_t^T f\left(s, X_s, u_s\right)\mathrm{d}s + g\left(X_T\right) \mid X_t = x\right]$$

Solves the partial-integro differential equation (PIDE):

$$\frac{\partial v}{\partial t}(t,x) + \sup_{u\in U}\left[\mathcal{L}^u v(t,x) + {}^v f(t,x,u)\right] = 0, \quad v(T,x) = g(x)$$

$\mathcal{L}^u$ is the generator of X_t

$$\begin{aligned}\mathcal{L}^u v(t,x) =& b(x,u)\cdot\nabla v(t,x) + \frac{1}{2}\mathrm{Tr}\left[\sigma(x,u)\sigma^T(x,u)\mathrm{H}(v(t,x))\right] \\ &+ \int_{\mathbb{R}^d}\left(v(t,x+G(x,z,u)) - v(t,x) - G(x,z,u)\cdot\nabla v(t,x)\right)\nu(\mathrm{d}z)\end{aligned}$$

Key Challenges

- Non-locality of PIDE: The presence of jumps introduces partial-integro differential equations (PIDE), the term $\int_{\mathbb{R}^d}$ is hard to evaluate numerically
- High-Dimensionality: The dimensionality of the control space and state variables may be high
- Full Nonlinearity: The presence of u in G makes no explicit expression of u^*

Current State of Research:

- Rich literature on ML solvers for problems under Brownian noise: Deep Galerkin/PINN, Deep BSDE, DBDP,...
- Very recent works have attempted to solve PIDEs or BSDEs: Gnoatto-Patacca-Picarelli (2022), Castro (2022), Backer-Jentzen-Kuckuck-Pellissier (2023), Alasseur-Bensaid-Dumitrescu-Warin (2024)
- Primarily address linear or semilinear problems, not suitable for fully nonlinear control problems

Deep Reinforcement Learning Approach

Actor-critic type framework to solve both the optimal value function $v(t,x)$ and the optimal control $u^*(t,x)$ simultaneously:

- Value function (critic): policy evaluation
- Control (actor): policy improvement
- Parameterize both by neural nets

Advantage:

- Accurate solution for the control policy
- Good for fully nonlinear problems (no explicit u^* required)
- Easy update via SGD

Policy Evaluation

Objective: Compute the value function $J^u(t,x)$ for a given policy $u(t,x)$

Temporal Difference Learning

- Incremental learning procedure driven by the error between the current value estimate and the actual observed reward plus the future value
- Updates occur after every state: $X_{t_1} \to X_{t_2} \cdots X_{t_n} \to X_{t_{n+1}} \cdots$

Recall Bellman equation:

$$J^u\left(t_n, X_{t_n}\right) = \mathbb{E}\left[R_{t_n t_{n+1}} + J^u\left(t_{n+1}, X_{t_{n+1}}\right) \mid X_{t_n}\right]$$

$$R_{t_n t_{n+1}} = \int_{t_n}^{t_{n+1}} f\left(t, X_t, u_t\right) \mathrm{d}t$$

TD Update Rule: parameterize J^u by a Neural Net $\mathcal{N}_v$

$$\text{Loss } = \left(R_{t_n t_{n+1}} + \mathcal{N}_v\left(t_{n+1}, X_{t_{n+1}}\right) - \mathcal{N}_v\left(t_n, X_{t_n}\right)\right)^2$$

Enhanced Update Rule

Improvement of TD Update

- Our update rule for $\mathcal{N}_v$ includes additional martingale terms
- Defined alternative reward $\widetilde{R}_{t_n t_{n+1}}$:

$$\widetilde{R}_{t_n t_{n+1}} = \int_{t_n}^{t_{n+1}} f\left(t, X_t, u\left(t, X_t\right)\right) \mathrm{d}t - \int_{t_n}^{t_{n+1}} \left(\sigma\left(X_t, u\left(t, X_t\right)\right)^T \nabla \mathcal{N}_v\left(t, X_t\right)\right)^T \mathrm{d}W_t$$
$$- \int_{t_n}^{t_{n+1}} \int_{R^d} \left[\mathcal{N}_v\left(t, X_t + G\left(X_t, z_v u\left(t, X_t\right)\right)\right) - \mathcal{N}_v\left(t, X_t, u\left(t, X_t\right)\right)\right] \widetilde{N}(\mathrm{d}t, \mathrm{d}z)$$

Theoretical Advantage

- If $\mathcal{N}_v = J^u$, then $\widetilde{R} + \mathcal{N}_v\left(t_{n+1}, X_{t_{n+1}}\right) - \mathcal{N}_v\left(t_n, X_{t_n}\right) = 0, \quad \mathbb{P}$-a.s. (almost surely)
- Reduced variance due to $\mathbb{P}$-a.s. condition, improving stability and performance.

Additional Challenge: $\tilde{N}(\mathrm{d}t, \mathrm{d}z)$ term hard to evaluate $\rightarrow \mathcal{N}_{\text{non}}(t, x) \rightarrow$ consistency issue between $\mathcal{N}_{\text{non}}$ and $\mathcal{N}_v$

Policy Improvement

Objective: Improve the current policy u to achieve higher $J^u(t, x)$

A natural approach: parameterize u by $\mathcal{N}_\pi$, and improve $\mathcal{N}_\pi$ by GD of

$$J^u(t, x) = \mathbb{E}^{t,x}\left[\int_0^T f\left(s, X_s, u\left(s, X_s\right)\right) \mathrm{d}s + g\left(X_T\right)\right]$$

Adjusted Value Function: add martingale terms

$$\widetilde{J}^u(t, x) = \mathbb{E}\left[\int_t^T f\left(s, X_s, u_s\right) \mathrm{d}s - \int_t^T \left(\sigma\left(X_s, u\left(s, X_s\right)\right)^T \nabla \mathcal{N}_v\left(s, X_s\right)\right)^T \mathrm{d}W_s\right.$$
$$\left. - \int_t^T \int_{\mathbb{R}^d} \left[\mathcal{N}_v\left(s, X_s + G\left(X_s, z, u\left(s, X_s\right)\right)\right) - \mathcal{N}_v\left(s, X_s, u\left(s, X_s\right)\right)\right] \widetilde{N}(\mathrm{d}s, \mathrm{d}z) + g\left(X_T\right)\right]$$

Key Observations:

- $\widetilde{J}^u(t, x) = J^u(t, x)$ theoretically for any control u
- Numerical experiments show $\widetilde{J}$ provides slightly better accuracy but increases computational cost

Simplified Algorithm for Poisson Jumps

State Process:

$$\mathrm{d}X_t = b\left(X_{t-}, u_t\right) \mathrm{d}t + \sigma\left(X_{t-}, u_t\right) \mathrm{d}W_t + \sum_{k=1}^m z_k\left(X_{t-}, u_t\right) \mathrm{d}M_t^k$$

where $M_t^k = N_t^k - \lambda_k t$ is the compensated Poisson process.
Key Features:

- Special case: $\nu(\mathrm{d}z)$ is a discrete measure
- Better interpretability and traceability, with applications in portfolio optimization and option pricing

- $\int_{\mathbb{R}^d} \cdots \nu(\,\mathrm{d}z)$ reduces to a finite sum, no need to have $\mathcal{N}_{\text{non}}$

HJB Equation Generator:

$$\begin{aligned}\mathcal{L}^u v(t,x) =& b(x,u)\cdot\nabla v(t,x) + \frac{1}{2}\operatorname{Tr}\left[\sigma(x,u)\sigma^T(x,u)\mathrm{H}(v(t,x))\right] \\ &+ \sum_{k=1}^{m}\lambda_k\left(v\left(t,x+z_k(x,u)\right) - v(t,x) - z_k(x,u)\cdot\nabla v(t,x)\right)\end{aligned}$$

Merton's problem

Consider investing between a bond S_t^0 and a stock S_t^1 :

$$\begin{aligned}\mathrm{d}S_t^0 &= rS_t^0\,\mathrm{d}t \\ \mathrm{d}S_t^1 &= S_{t-}^1\left(\mu\,\mathrm{d}t + \sigma\mathrm{d}B_t + z\,\mathrm{d}M_t\right)\end{aligned}$$

The wealth process is

$$\mathrm{d}X_t = \left(r + u_t(\mu - r)\right)X_{t-}\mathrm{d}t + \sigma u_t X_{t-}\mathrm{d}B_t + z u_t X_{t-}\mathrm{d}M_t$$

and the agent aims to maximize $v(t,x) = \sup_u \mathbb{E}\left[g\left(X_T\right) \mid X_t = x\right]$, utility func. g

Stochastic LQR problem

The state process X_t satisfies

$$\mathrm{d}X_t = u_t\,\mathrm{d}t + \sigma\mathrm{d}B_t + z\,\mathrm{d}M_t$$

where $M_t = \left(M_t^1, \cdots, M_t^d\right), M_t^i = N_t^i - \lambda_i t$ and N_t^i denotes a Poisson process with intensity λ_i. The agent aims to solve

$$v(t,x) = \inf_u \mathbb{E}^{t,x}\left[\int_t^T \left(q\left\|u_s\right\|_r^2 + b\left\|X_s\right\|^2\right)\mathrm{d}s + a\left\|X_T\right\|^2\right]$$

Stochastic Games with Jumps

State Process Dynamics: $X_t = \left(X_t^1, X_t^2, \ldots, X_t^n\right)^T$ satisfy:

$$\mathrm{d}X_t = b\left(X_{t-}, \pi_t\right)\mathrm{d}t + \sigma\left(X_{t-}, \pi_t\right)\mathrm{d}W_t + \int_{\mathbb{R}^d} G\left(X_{t-}, z, \pi_t\right)\tilde{N}(\,\mathrm{d}t,\ \mathrm{d}z)$$

where $\pi_t = \left(\pi_t^1, \ldots, \pi_t^n\right)$ is the control vector, and $b(\cdot), \sigma(\cdot), G(\cdot)$ describe the drift, diffusion, and jump terms, respectively

Game Objective: Agent i aims to maximize the expected utility:

$$v^i(t,x) = \sup_{\pi^i} J_i^{\pi}(t,x) = \sup_{\pi^i}\mathbb{E}^{t,x_s}\left[\int_t^T f_i\left(s, X_s, \pi_s\right)\mathrm{d}s + g_i\left(X_T\right)\right]$$

Definition (Nash Equilibrium)

A tuple of strategies $(\pi_1^*, \cdots, \pi_n^*)$ is a Nash equilibrium if, under the initial condition $X_0 = x_0$, for any agent i and any strategy π_i, we have

$$J_i^{(\pi_1^*, \cdots, \pi_{i-1}^*, \pi_i^*, \pi_{i+1}^*, \cdots, \pi_n^*)}(0,x) \geq J_i^{(\pi_1^*, \cdots, \pi_{i-1}^*, \pi_i, \pi_{i+1}^*, \cdots, \pi_n^*)}(0,x)$$

$\rightarrow v^i$ satisfy a coupled Hamilton-Jacobi-Bellman (HJB) System

Multi-agent portfolio game

- The stock price and agent i 's wealth process:

$$\mathrm{d}S_t^i = S_{t-}^i \left(\mu_i \,\mathrm{d}t + \nu_i \,\mathrm{d}W_t^i + \sigma_i \,\mathrm{d}B_t + \alpha_i \,\mathrm{d}M_t^i + \beta_i \,\mathrm{d}M_t^0\right), \quad 1 \le i \le n$$
$$\mathrm{d}X_t^i = \frac{\pi_t^i}{S_{t-}^i} \,\mathrm{d}S_t^i = \pi_t^i \left(\mu_i \,\mathrm{d}t + \nu_i \,\mathrm{d}W_t^i + \sigma_i \,\mathrm{d}B_t + \alpha_i \,\mathrm{d}M_t^i + \beta_i \,\mathrm{d}M_t^0\right)$$

- Agent i aims to maximize

$$J_i\left(\pi^1, \cdots, \pi^n\right) = \mathbb{E}\left[U_i\left(X_T^1, \cdots, X_T^n\right)\right]$$

For the exp case, agent i 's utility function U_i depends on the arithmetic average of others':

$$U_i\left(x_1, x_2, \cdots, x_n\right) = -e^{-\frac{1}{\delta_i}\left(x_i - \theta_i \frac{1}{n}\sum_{k=1}^n x_k\right)}$$

For power and log cases, U_i takes into account the geometric average:

$$U_i\left(x_1, x_2, \cdots, x_n\right) = \frac{1}{p_i}\left(\frac{x_i}{\left(\prod_{k=1}^n x_k\right)^{\frac{\theta_i}{n}}}\right)^{p_i}, \log \frac{x_i}{\left(\prod_{k=1}^n x_k\right)^{\frac{\theta_i}{n}}}$$

Theorem (Uniqueness of NE, Lu-H.-Yang-Zhu, 2024)

Under proper conditions, the portfolio game has a unique constant Nash equilibrium, given by the solution of a system of algebric equations.

Sketched proof:

1. Proposed ansatz → the HJB system constant solutions described by the solution of algebric equation → provide a optimal control (verification theorem).
2. The strict concavity of the optimization problem → no other constant solution to the HJB equation.
3. Algebric equations has a unique solution → unique constant NE.

10 References

References

[1] Yacine Ait-Sahalia. Asset pricing in an economy with changing sentiment and price feedback. *Princeton*.

[2] Denis Chetverikov. Estimation of risk premia with many factors. *UCLA*.

[3] Francis X. Diebold. Machine learning and the yield curve: tree-based macroeconomic regime switching. *UPenn*.

[4] Chao Gao. Are adaptive robust confidence intervals possible? *UChicago*.

[5] Fang Han. Chattejee's rank correlation: what is new? *UWashington, Seattle*.

[6] Ruimeng Hu. Deep reinforcement learning for games with controlled jump-diffusion dynamics. *UC Santa Barbara*.

[7] Bryan Kelly. Arbitrage pricing theory or '*AI* pricing theory'? the surprising dominance of large factor models. *Yale*.

[8] Mladen Kolar. Confidence sets for causal discovery. *USC*.

[9] Zongming Ma. Multimodal data integration and cross-modal querying via orchestrated approximate message passing. *Yale*.

[10] Theodor Misiakiewicz. Deterministic equivalents and scaling laws for random feature regression. *Yale*.

[11] Andrea Montanari. Deterministic equivalents and scaling laws for random feature regression. *Stanford*.

[12] Whitney Newey. Automatic debiased machine learning via riesz regressions. *MIT*.

[13] Lan Zhang Per Mykland. Estimating the volatility of drift. *UChicago, UIC*.

[14] Bodhi Sen. Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood. *Columbia*.

[15] Katja Smetanina. Perceived shocks and impulse responses. *UChicago*.

[16] Pragya Sur. Generalization error of min-norm interpolators in transfer learning. *Harvard*.

[17] Chenhao Tan. Towards human-centered ai: predicting fatigue and generating hypothesis with *LLM*s. *UChicago*.

[18] Dacheng Xiu. On the theory of autoencoders. *UChicago*.

[19] Wenxin Zhou. Nonparametric expected shortfall regression with tail-robustness. *UIC*.