Narrative Asset Pricing: Interpretable Systematic Risk Factors from News Text

Review of Financial Studies, 2023

Leland Bybee, Bryan Kelly, Yinan Su

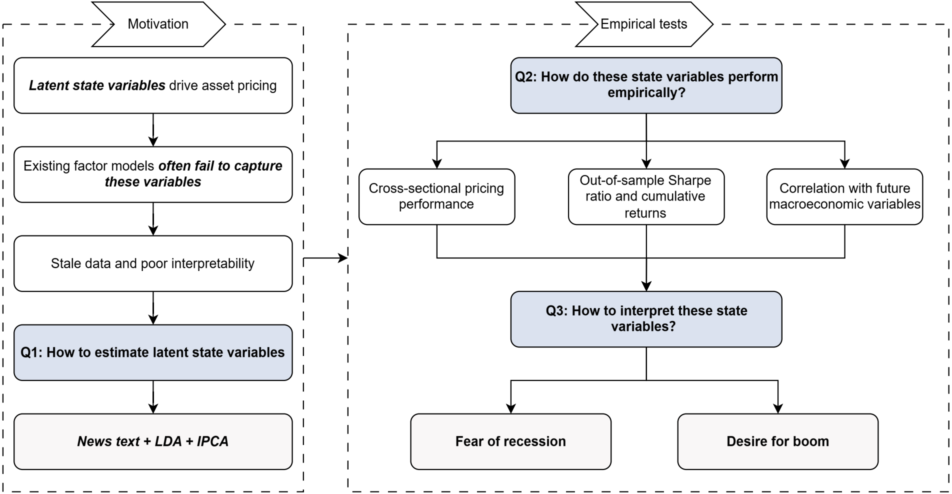
Presented by Kang Guo

November 30, 2025

Overview

1. Introduction
2. Design
3. Result
4. Idea

Framework


Motivation
Latent state variables drive asset pricing
Existing factor models *often fail to capture these variables*
Stale data and poor interpretability
Q1: How to estimate latent state variables
News text + LDA + IPCA
Empirical tests
Q2: How do these state variables perform empirically?
Cross-sectional pricing performance
Out-of-sample Sharpe ratio and cumulative returns
Correlation with future macroeconomic variables
Q3: How to interpret these state variables?
Fear of recession
Desire for boom

Question

- Q1: How to estimate the latent state variables in the ICAPM?
- Q2: How do these state variables perform in cross-sectional pricing and forecasts future macroeconomic variables?
- Q3: How to interpret these state variables?

Motivation

- In asset pricing, it is unclear which **fundamental risks** investors care about
- Merton's (1973) ICAPM: risk is tied to **latent macro state variables** that influence **investors' wealth and future investment opportunities**, such as interest rates
- ICAPM state variables are hard to measure with existing methods:
  1. Visible macroeconomic variables
     - Industrial production, investment, and inflation ⇒ Stale
  2. Characteristic-sorted portfolios
     - FF3, FF5, and **IPCA** (Kelly et al., 2019,JFE) ⇒ Poor interpretability

- **This paper: News text + LDA + IPCA framework**
  - News ⇒ timely & forward-looking pricing information
  - LDA ⇒ maps text onto interpretable topic distributions
  - IPCA ⇒ maps topic distributions to tradable investment portfolios

Marginal contribution

- Estimate the ICAPM factor pricing model
  - Prior studies
    - Using visible macroeconomic variables (Bali and Engle, 2010, JME; Rossi and Timmermann, 2015, RFS) and characteristic-sorted portfolios (Fama and French, 1996, JF; Hou et al., 2015, RFS; Kelly et al., JFE)
  - This paper
    - Using news text data and finding that the news-based ICAPM model performs better in cross-sectional pricing and interpretability than existing factor models

Hypothesis

- H1: The narrative factor model exhibits lower cross-sectional pricing errors than existing factor models
  - More closely related to latent macroeconomic state variables
- H2: The pricing kernel (optimal linear combination of state variables) is procyclical
  - The pricing kernel depends on stocks' covariances with the news-based topics, and most individual stocks are procyclical
- H3: The pricing kernel mainly captures recession fears and boom desires

Q1: Estimating the ICAPM factor model using news text

- Step 1: Obtain the change in the news topic distribution (z_τ) vector (L $\times$ 1) using LDA

$$z_\tau = \theta_\tau - \frac{1}{5}\sum_{j=1}^{5}\theta_{\tau-j} \tag{1}$$

- θ_τ: **the weighted average topic distribution of all articles on day** τ

- Step 2: For each stock i and month t, calculate covariances between $r_{i,\tau}$ and z_τ from daily data (1 $\times$ L):

$$\widehat{cov}_{i,t} = cov(r_{i,\tau}, z_\tau^T) \tag{2}$$

Q1: Estimating the ICAPM factor model using news text

- Step 3: Append a constant to the covariances to form a set of $(L+1)$ instruments $c_{i,t} = [1, \widehat{cov}_{i,t}]$, assuming an IPCA factor model (kelly et al., 2019, JFE):

$$r_{i,t+1} = c_{i,t}\,\Gamma\, f_{t+1} + e_{i,t+1}, \beta_{i,t} = c_{i,t}\,\Gamma \tag{3}$$

- $\Gamma \in \mathbb{R}^{(L+1)\times K}$ is a matrix with rows indexed from 0 to L $(\Gamma = [\Gamma_0; \Gamma_1; \ldots; \Gamma_L])$
- f_t are portfolio returns constructed on latent state variables

- Step4: Estimate the f_t and Γ with Sparse IPCA:
  - Input: $r_{i,t+1}$ and $c_{i,t}$
  - Output: f_t and Γ

Q1: Narrative interpretation of the latent state factors

- Infer the values of the latent state variables from f_t and Γ

$$x_\tau = (A^\top A)^{-1} A^\top z_\tau = I_{z\to x}^\top z_\tau, \tag{4}$$

$$A = \widetilde{\Gamma}(\widetilde{\Gamma}^\top \widetilde{\Gamma})^{-1} \Sigma_{\mathrm{ff}}^{-1} \tag{5}$$

$$\widetilde{\Gamma} = [\Gamma_1; \ldots; \Gamma_L] \tag{6}$$

- $L \times K$ **matrix** $I_{z\to x} = A(A^\top A)^{-1}$ **summarizes how changes in the** L **new topic distributions affect the** K **states**
- Finally, we calculate the pricing kernel

$$x_\tau^{\mathrm{MVE}} = b^{\mathrm{MVE}} x_\tau = b^{\mathrm{MVE}} (A^\top A)^{-1} A^\top z_\tau \ = \ I_{z\to \mathrm{MVE}}^\top z_\tau \tag{7}$$

- where the $L \times 1$ **"impact vector"** $I_{z\to\mathrm{MVE}}$ **summarizes the impact of each narrative on the pricing kernel**

Data

- Stock return data are from CRSP for firms listed on NYSE, AMEX, and NASDAQ
- Daily news from WSJ
- Full sample period: 198501–201612
- Out-of-sample period: 200101-201612

Q2: Cross-sectional pricing performance

- The narrative factor model outperforms traditional factor models in explaining anomalies

A. 78 anomaly portfolios as test assets

Factors	avg $\lvert\hat{\alpha}_a\rvert$	avg $\lvert t(\hat{\alpha}_a)\rvert$	$\frac{\#\lvert t(\hat{\alpha}_a)\rvert>1.96}{\#\text{test assets}}$	GRS
Mkt	1.11	2.64	0.54	8.56
FF3	0.97	2.65	0.54	8.50
FF5	1.18	3.20	0.68	7.48
FFC6	1.27	3.43	0.74	7.41
NF1	1.29	3.03	0.73	8.78
NF2	0.97	2.72	0.54	7.98
NF3	0.84	2.62	0.54	7.84
NF4	0.92	2.81	0.55	7.53
NF5	0.91	2.78	0.60	7.43
NF6	0.96	2.89	0.63	7.38

B. 25 size/bm double sorts as test assets

Factors	avg $\lvert\hat{\alpha}_a\rvert$	avg $\lvert t(\hat{\alpha}_a)\rvert$	$\frac{\#\lvert t(\hat{\alpha}_a)\rvert>1.96}{\#\text{test assets}}$	GRS
Mkt	0.42	3.03	0.84	11.61
FF3	0.32	3.89	0.84	15.12
FF5	0.27	3.31	0.72	13.06
FFC6	0.28	3.40	0.76	12.68
NF1	0.35	2.33	0.64	6.56
NF2	0.16	1.14	0.16	5.29
NF3	0.25	1.95	0.56	5.72
NF4	0.23	1.75	0.44	5.58
NF5	0.23	1.78	0.48	5.48
NF6	0.25	1.91	0.48	5.56

Q2: Cross-sectional pricing performance

- The narrative factor model provides incremental pricing information

Sharpe ratios of MVE portfolios combining FFC6 and NF

	K (number of NF added)						
Specification	0	1	2	3	4	5	6
NF		0.48	1.00	1.10	1.26	1.32	1.31
NF + Mkt	0.25	0.48	1.00	1.08	1.26	1.32	1.31
NF + FF3	0.36	0.41	0.60	0.68	1.02	1.17	0.98
NF + FF5	0.90	0.89	0.94	0.92	1.01	1.01	1.08
NF + FFC6	0.67	0.65	0.76	0.80	0.87	0.92	1.19

Q2: Out-of-sample MVE Sharpe ratio

- The out-of-sample factor returns for month t +1 is

$$f_{t+1}^{OOS} = \left(\sum_i \hat{\beta}_{i,t} \hat{\beta}_{i,t}^{\top} + 2 I_K \right)^{-1} \left(\sum_i \hat{\beta}_{i,t} r_{i,t+1} \right) \quad (8)$$

- Buy high beta stocks and short low beta stocks

- The linear weighting of out-of-sample factor returns

$$f_{t+1}^{MVE,OOS} = \hat{\mu}_f^{\top} \hat{\Sigma}_{ff}^{-1} f_{t+1}^{OOS} \quad (9)$$

Q2: Out-of-sample MVE Sharpe ratio

- Narrative factor models achieve higher sharpe ratios
- Sparse IPCA perform better than traditional IPCA

A. Narrative factor model

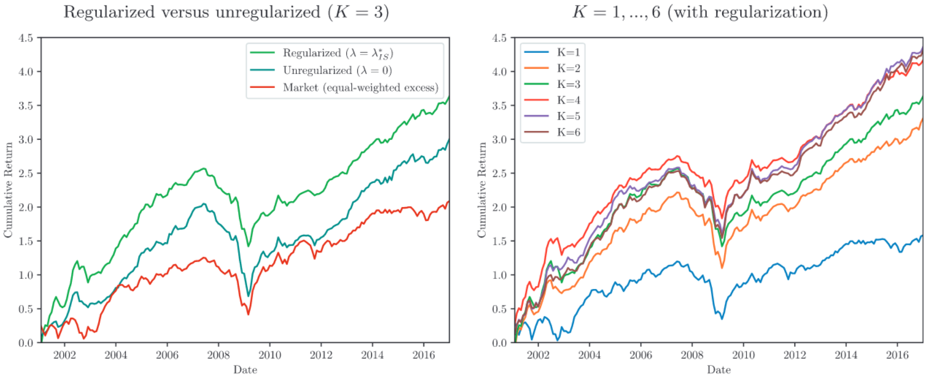
		K					
Tuning	Statistics	1	2	3	4	5	6
$\lambda=\lambda_{\mathbb{S}}^{*}$	Sharpe ratio	0.48	1.00	1.10	1.26	1.32	1.31
	# narratives	2.9	4.9	12.1	39.1	43.4	61.8
$\lambda=0$	Sharpe ratio	0.44	0.66	0.73	0.73	0.79	0.91
	# narratives	All 180 narratives, no selection					

B. Benchmark factors

	Mkt	SMB	HML	RMW	CMA	UMD
Sharpe ratio	0.25	0.13	0.36	0.82	0.90	0.67

Q2: Out-of-sample MVE cumulative returns

- Sparse IPCA perform better than traditional IPCA in time series
- Investment performance of the narrative MVE portfolio is not concentrated in a particular period or driven by a particular event



Q2: x^{MVE} and future macroeconomic variables

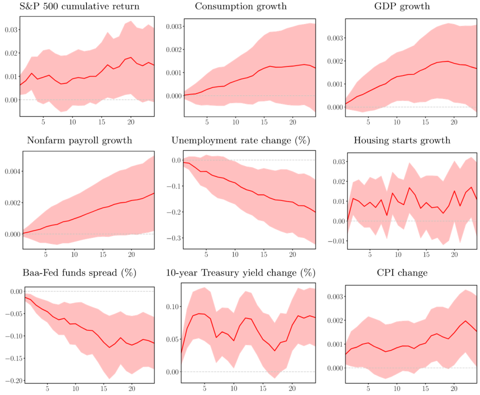
- For each forecast target, authors predict the cumulative changes over different horizons (h)

$$\sum_{s=1}^{h} \psi_{t+s} = b_h \left(\frac{x_t^{\text{MVE}}}{\text{std}(x_t^{\text{MVE}})} \right) + \varepsilon_t^{(h)} \tag{10}$$

- ψ_t denotes the one-month change in a macroeconomic variable, such as $\psi_t =$ GDP growtht , and the summation takes the cumulative change in the future horizon of h months
- Authors standardize x_t^{MVE} so that the coefficient b_h can be interpreted as the effect per one-standard-deviation change in the state variable

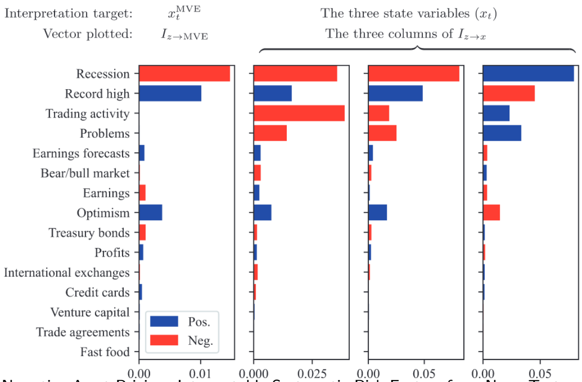
Q2: x^{MVE} and further macroeconomic variables

- The pricing kernel (x^{MVE}) is procyclical



Q3: Interpretation results: Impact of article topics on the MVE state variable

- The 'Recession' narrative has the most negative impact on the pricing kernel
- The narratives 'Record High' and 'Optimism' have the largest positive impact on x^{MVE}



Q3: Interpretation results: Impact of term topics on the MVE state variable

- The term clouds most reflect keywords from the 'Recession' and the 'Record high' narratives

Negative

Positive

Q3: Narrative retrieval

- The impact of a single article m's topic distribution on market return

$$I_{z\to \mathrm{Mkt}}^{\top} z(m) \tag{11}$$

Recession	
1993-05-07	Auto Registrations Continued to Slump In Europe Last Month
2001-04-25	Consumer Confidence Slides on Fears of Layoffs
2009-02-19	U.S. News: Housing Starts Hit Lowest Level In Half-Century
2011-08-02	World News: Manufacturing Slowdown Adds to Gloom on Economy
2016-07-08	World News: U.K. Consumer Sentiment Takes Dive
Record high	
1989-07-05	Japan Vehicle Sales Rise
1994-07-01	Purchasing Managers In U.K. Survey Report Rise for June Orders
1995-02-27	Hiring Outlook For Second Quarter Appears Vigorous
2006-01-12	Wall Street Bonuses Hit a Record in 2005
2016-07-20	U.S. News: Home Building Continues Recovery as Demand Rises
Trading activity	
1993-12-30	Industrials Rise A Bit to Record; Bonds Decline
1994-10-20	Profit News Helps Boost Stock Prices — Indexes Gain Ground Despite Weakness Of Bonds and Dollar
1996-06-21	Nasdaq Sinks Amid Sell-Off Of Tech Stocks
1997-12-09	Blue Chips Fall As Dollar's Rise Causes Concern
1998-04-21	Drug Stocks Resume Gains; Blue Chips Fall

Extension

1. Cross-sectional pricing
   - Use stocks' covariances with changes in news-topic distributions as inputs to machine learning methods
2. The improvement in narrative topics
   - Incorporate the dynamic evolution of sentiment or semantic structure
   - Focus on domain-specific risks, such as green (or climate-related) risk
3. Extract economic narratives from news photos
4. Extend the framework to the cryptocurrency and fixed income markets