Discussions on

Missing Financial Data

Svetlana Bryzgalova Sven Lerner Martin Lettau Markus Pelger

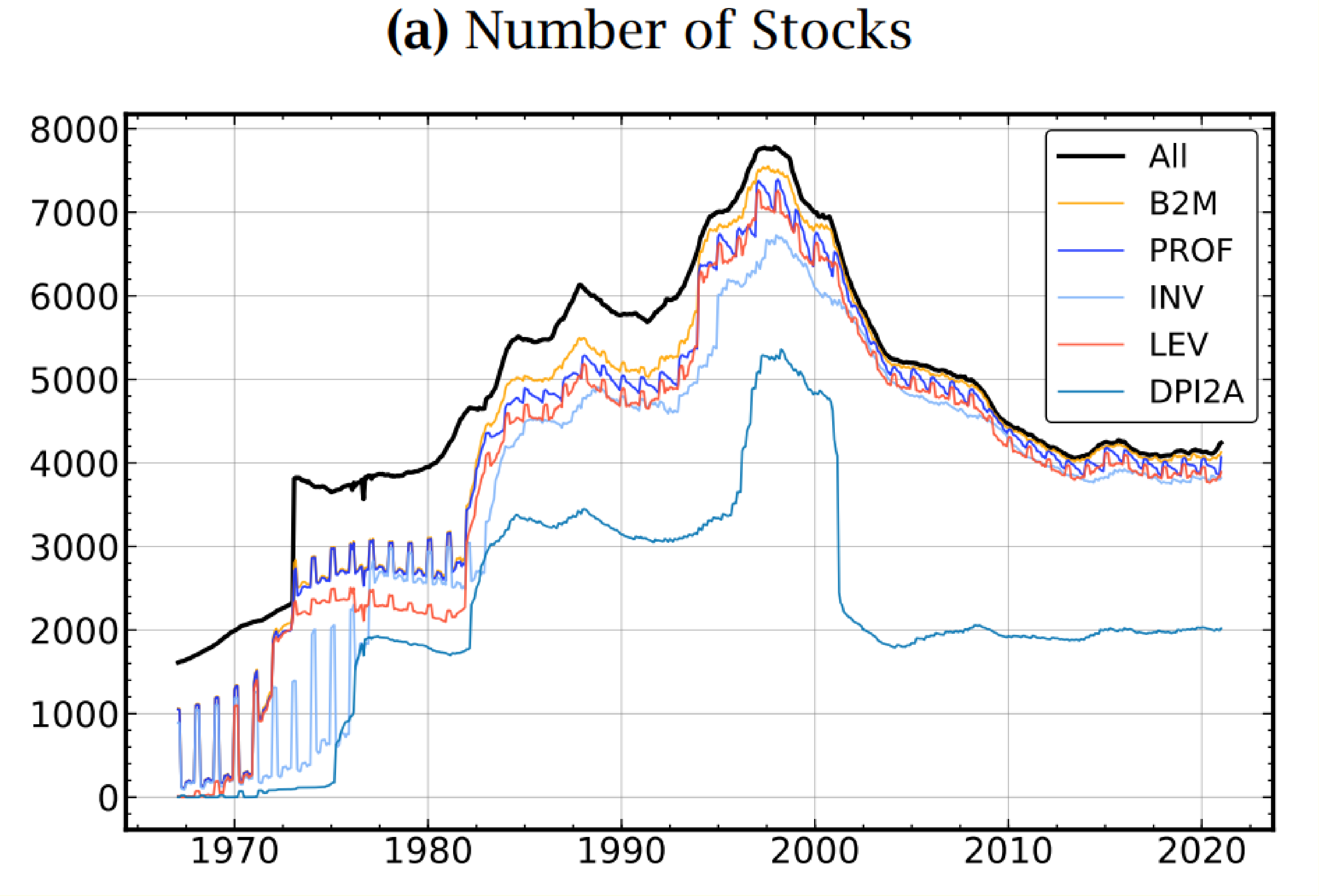
Comments by Guofu Zhou

AI & BIG DATA IN FINANCE RESEARCH FORUM

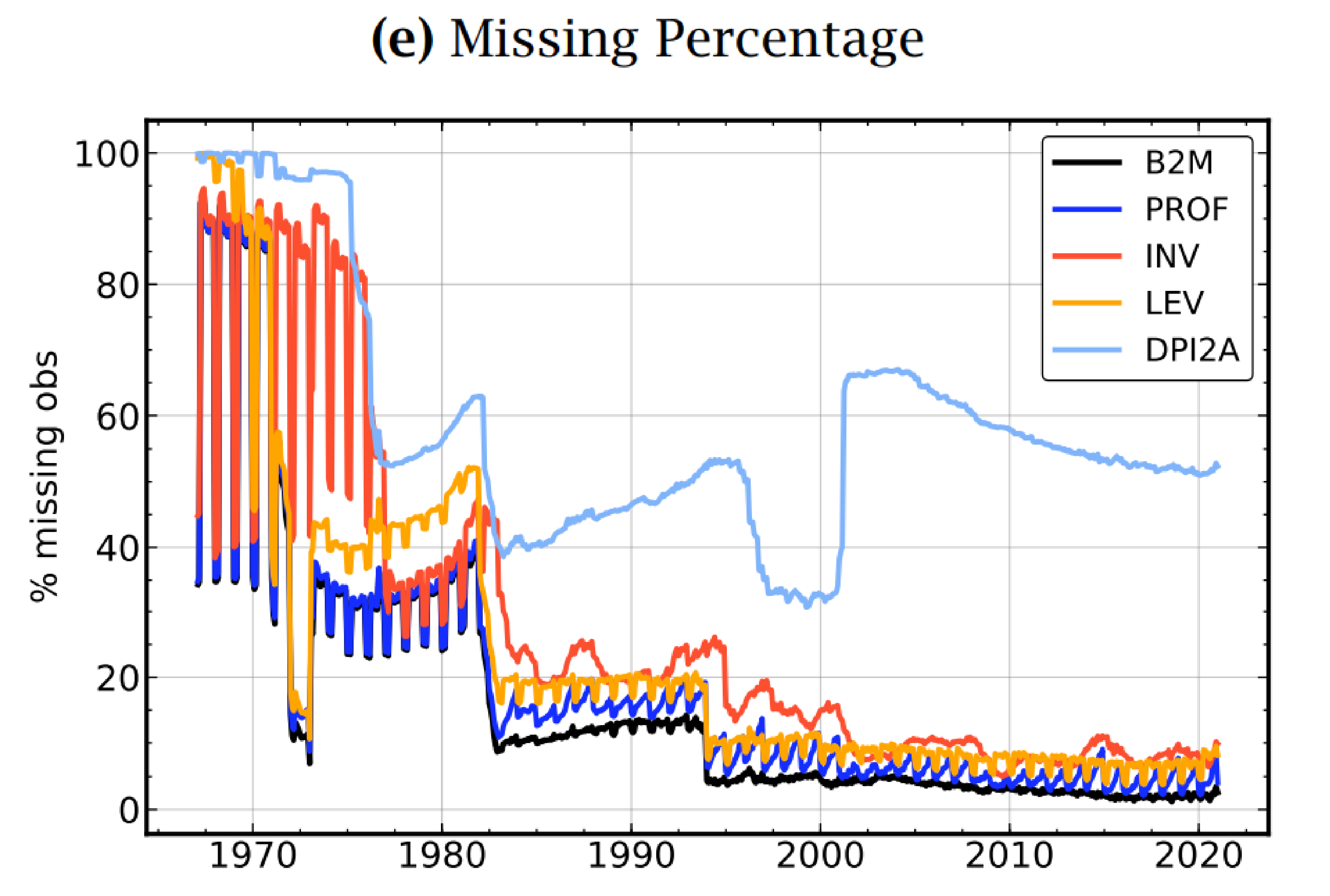
ABFR Webinar, April 28, 2022

Is there really a problem?

Problem of Missing Values: A



Problem of Missing Values: B



Why Is It Important ?

- Thousands of Studies reply on firm characteristics
  - Fama and French 3-factor models, …
  - Hundreds of anomalies
  - Empirical corporate; Accounting
- Potential issues: under- or over-estimate
- What about machine learning?
  - Panel approach: selection bias
  - Exception: Han, et al, 2022, "Expected Stock Returns … E-LASSO …"
    - univariate XS: replacing missing forecasts by others; competitive to NN.
  - Literature correction: Rapach, Strauss and Zhou (2013, JF) is perhaps the first academic study (published in a top finance journal) that applies LASSO, Enet in finance.
- Pathbreaking !
  - a generic approach dealing with missing data
  - Another paper, Freyberger, et al, 2021, "Missing Data in asset pricing"
  - Both offer unique insights

How to solve the problem?

The Idea

At time t, Let $C_{i,l}^t$ be firm i's characteristic l.

Assume a factor model for the $N_t \times L$ matrix:

$$C_{i,l}^t = F_i^t \Lambda_l^{t\top} + e_{i,l}^t$$

PCA with all data: $$\tilde{\Sigma}_t^{XS} = \frac{1}{L}\sum_{l=1}^{L} C_l^t C_l^{t\top}, \qquad N_t \times N_t$$

PCA with missing: $$\hat{\Sigma}_t^{XS} = \frac{1}{|Q_{i,j}^t|}\sum_{l \in Q_{i,j}^t} C_l^t C_l^{t\top},$$

summing over observed data.

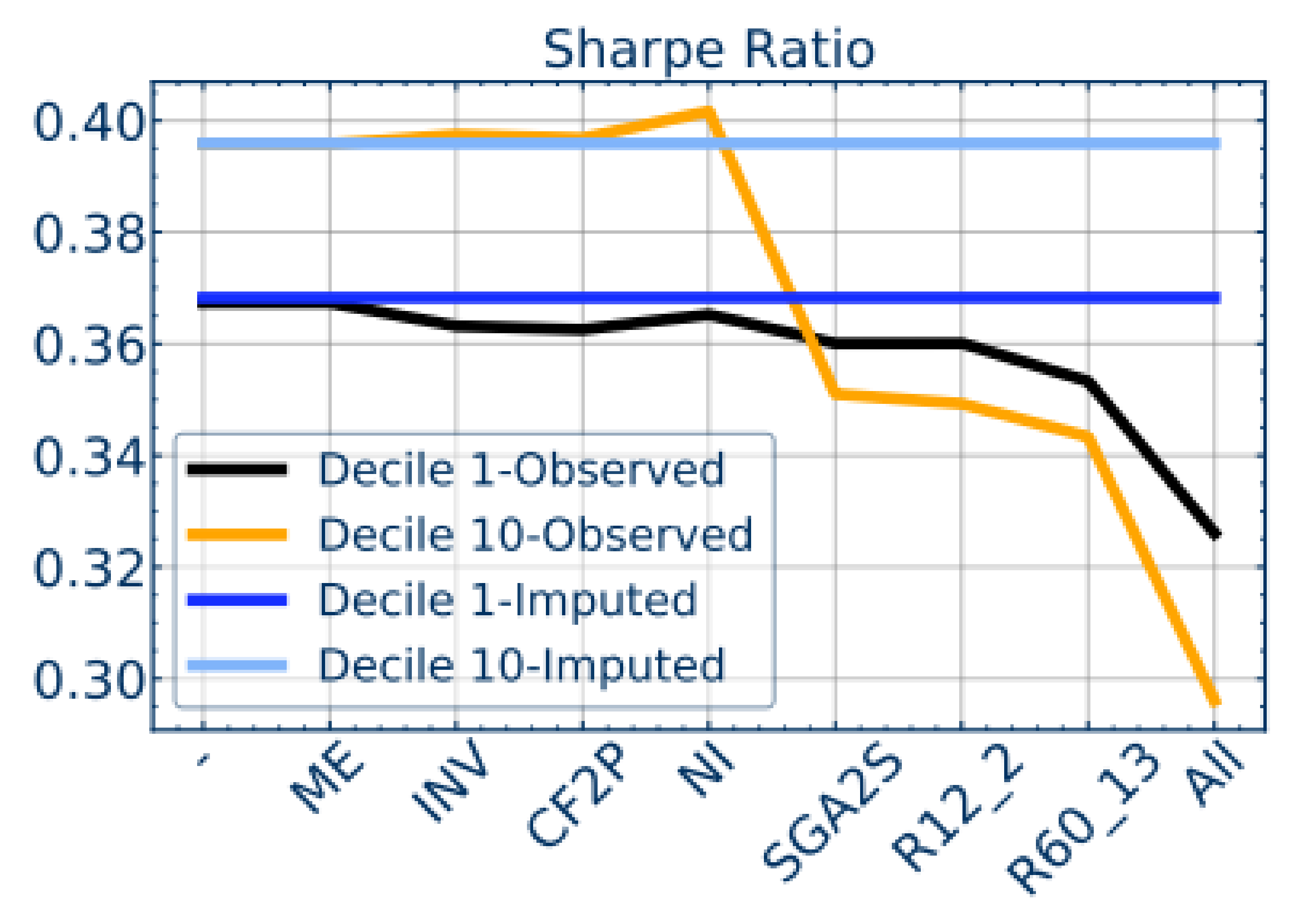
Major Results: 1

Table 3: Imputation Error for Different Imputation Methods

	In-Sample			OOS MAR			OOS Block		
Method	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly
global BF-XS	**0.11**	**0.10**	**0.13**	**0.15**	**0.15**	**0.14**	**0.17**	**0.16**	**0.19**
global F-XS	0.10	0.07	0.14	0.16	0.17	0.16	0.18	0.17	0.20
global B-XS	0.15	0.15	0.14	0.16	0.16	0.15	0.19	0.18	0.20
global XS	0.19	0.18	0.21	0.23	0.22	0.24	0.22	0.21	0.24
global B	0.16	0.17	0.15	0.17	0.17	0.15	0.21	0.20	0.22
local B-XS	**0.15**	**0.16**	**0.14**	**0.16**	**0.17**	**0.15**	**0.19**	**0.19**	**0.20**
local XS	0.21	0.20	0.22	0.23	0.22	0.24	0.23	0.22	0.24
prev	0.18	0.18	0.18	0.19	0.19	0.19	0.23	0.21	0.25
local B	0.16	0.17	0.15	0.17	0.17	0.15	0.21	0.20	0.22
XS-median	0.29	0.29	0.29	0.29	0.29	0.29	0.28	0.28	0.29
ind-median	0.29	0.29	0.29	0.29	0.29	0.29	0.28	0.28	0.29

Major Results: 2

Figure 15: Univariate Sorts With and Without Missing Values



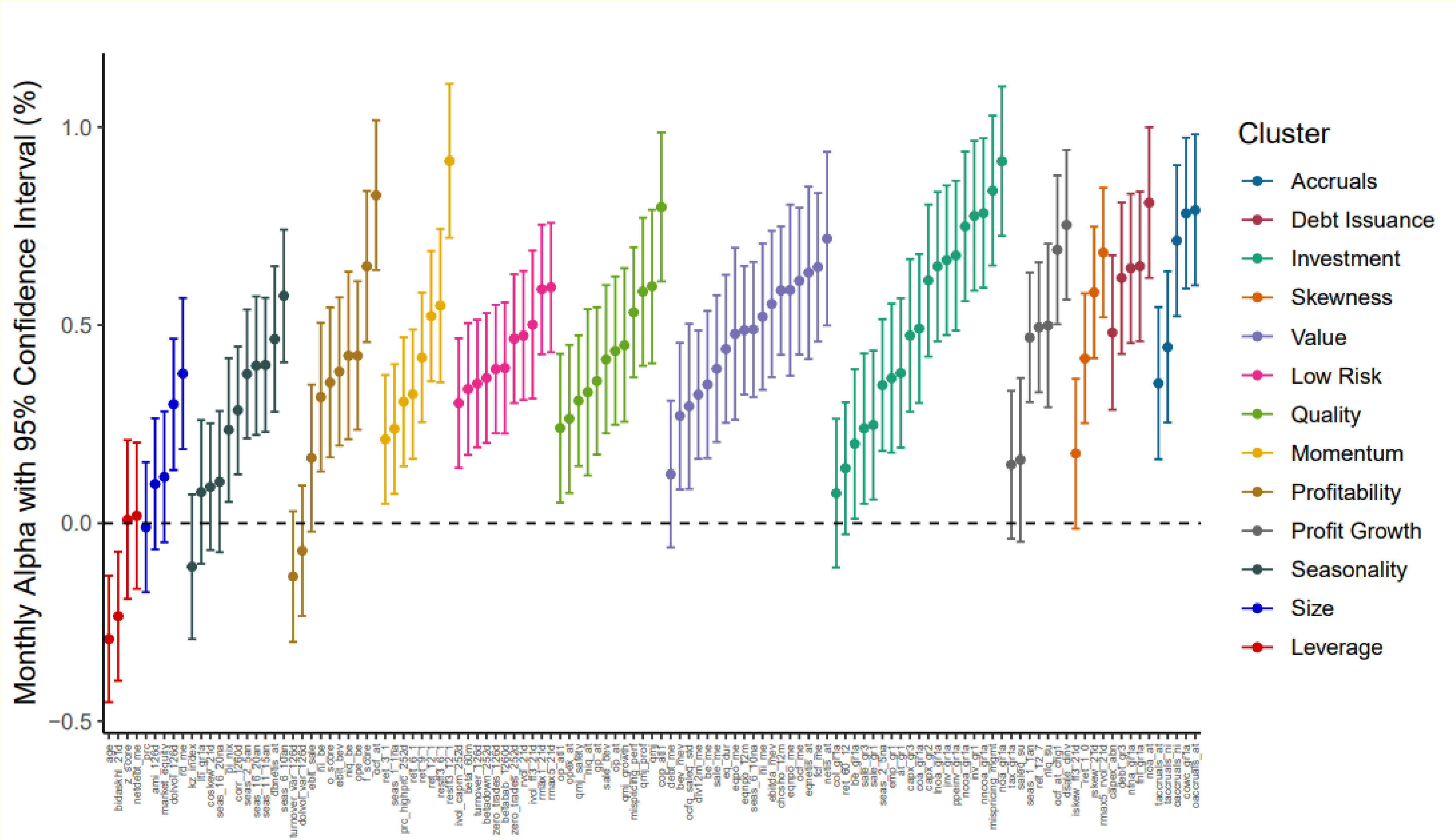
Book-to-Mkt: sequentially to have more characteristics available.

Model Assumptions?

- Doing PCA on characteristics
  - more discussions on the types of missing?
    - missing at random
    - missing completely at random
  - simple:
    - given firm, it can miss 50% of the data?
    - Given characteristic, 75% of firm should not miss?
- The impact of different missing
  - Errors in identifying K?
  - Errors in estimating the eigenvalues/vectors?

Interpretation of PCs?

- Given, say, 10 PCs:
  - What are the economic interpretations ?
    - data-driven grouping
  -



Source: Jensen, Kelly, and Pedersen (2021, Is There a Replication Crisis in Finance?).

Interpretation of PCs?

- Given, say, 10 PCs:
  - What are the economic interpretation?
    - data-driven grouping
  - Which one is the most important? The least?
- If hard to explain, sparse PCA?
  - Pelger and Xiong (2021)
  - Rapach and Zhou (2021)
    - Get interpretable macro factors
      - Each PC is a combination of a few highly related macros
    - The factors are competitive to Fama-French factors!

s-PCA

PCA on:

$$\hat{\Sigma}_t^{XS} = \frac{1}{|Q_{i,j}^t|} \sum_{l \in Q_{i,j}^t} C_l^t C_l^{t\top},$$

PCA+ LASSO

$$\max_{\boldsymbol{u}_1, \boldsymbol{v}_1} \boldsymbol{u}_1' \boldsymbol{X} \boldsymbol{v}_1 \text{ subject to } \|\boldsymbol{u}_1\|_2^2 = 1,\ \|\boldsymbol{v}_1\|_2^2 = 1,\ \|\boldsymbol{v}_1\|_1 \leq c,$$

Q: 1st PC about <u>trading friction</u> characteristics?

Source: Witten, D. M., R. Tibshirani, and T. Hastie (2009), A Penalized Matrix Decomposition, with Applications to Sparse Principal Components and Canonical Correlation Analysis. Biostatistics 10, 515–534.

WTH exploit the biconvexity to develop an efficient iterative algorithm.

Asset Pricing Implications

- Given K characteristic factors,
  - do they contain all the info of the characteristic to price all stocks?
  - only K categories of anomalies?
  - is the largest risk of characteristics carry the most risk premium?
  - which imputation method does the best in explaining the expected returns?
  - the Sharpe ratio?

Timely Forecasts?

- In cross-section forecasting, one often lags the characteristics a few months.
- No longer necessary!
  - if missing only a small amount, why “throw out the baby with the bath water”?
    - ☞ more timely info should be more valuable.
  - if in doubt, impute them. How will this affect the results?

A Paper to cite ?

- Liu, Tang and Zhou (2022, JFE, forth)
  - “Recovering the FOMC Risk Premium”
- Anything in common ?
  - missing data
  - options with expiration right after the FOMC
  - early years unavailable options
    - matrix completion via implied volatility surface
- Why cite ?
  - missing data problem too in option pricing
  - an alternative solution to a different problem
  - that paper does cite this one (in the last minute; good to inform readers on general approaches dealing with missing data).

Overall

- Thought provoking paper !
- Impressive results !
- Wide applications !
  - Bonds, FXs, mutual funds, etc.
  - Corporate, Accounting