Missing Financial Data

Svetlana Bryzgalova, Sven Lerner, Martin Lettau, Markus Pelger.

Working paper,2022

解读者：Tu Xueyong

2023.03.22

1. Introduction-- Motivation

- Phenomenon： missing data in firm fundamentals
- The issue of missing data is widespread yet little-researched
  - Many Compustat variables are sparsely populated
- Several potential effects for asset pricing
  - it reduces the number of stocks in portfolios
  - the set of stocks in portfolios may vary by characteristic
  - factor premia might be affected if not random

1. Introduction-- Objectives

- Provide a comprehensive analysis of missing data
- Estimate an econometric model for imputing missing values
- Analyze how missingness affects returns of portfolios sorted on characteristics

1. Introduction-- Stylized facts

- Missing financial data is very prevalent
- Worse whenever one requires multiple characteristics
- Data is not missing at random
- Returns on their own depend on whether a firm has missing fundamentals
    - Missing → Lower return

1. Introduction-- Contributions

- Provide a comprehensive analysis of missing data
- Provide a novel approach to the imputation of missing firm fundamentals
- Our imputation method strongly dominates leading conventional approaches

1. Introduction-- Related Literature

- Most widely used approaches of imputation:
  - cross-sectional median imputation
  - fully observed data
- Missing data in panels:
  - Xiong and Pelger (2019): cross-sectional factor model
  - Bai and Ng (2021), Cahan et al. (2021), and Jin et al. (2021) : alternative latent factor with different assumptions on the missing pattern
- Causal inference in a panel:
  - Athey et al. (2021) and Xiong and Pelger (2019)
  - The unobserved counterfactual outcomes can be modeled as missing values
- Direct implications for the multidimensional challenge
  - Chen et al. (2019), Gu et al.(2020)

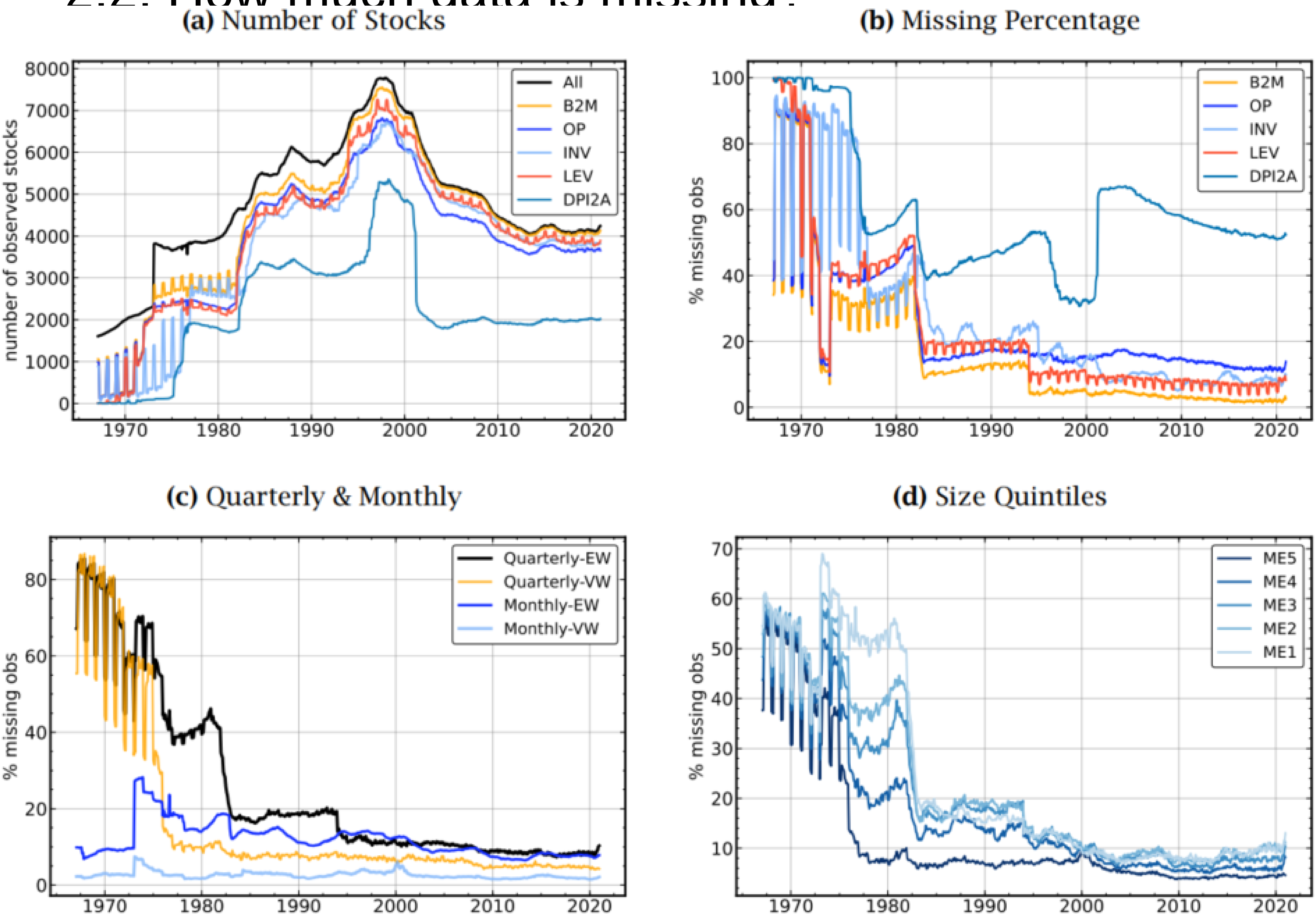
1. Introduction-- Related Literature

- Addresses the problem of missing financial:
  - Freyberger et al. (2021) : general GMM estimation
  - Xiong and Pelger (2022) : causal inference in finance
  - Blanchet et al. (2022) analyze the trade-off between look-ahead-bias and variance in an imputation

2. Missing values

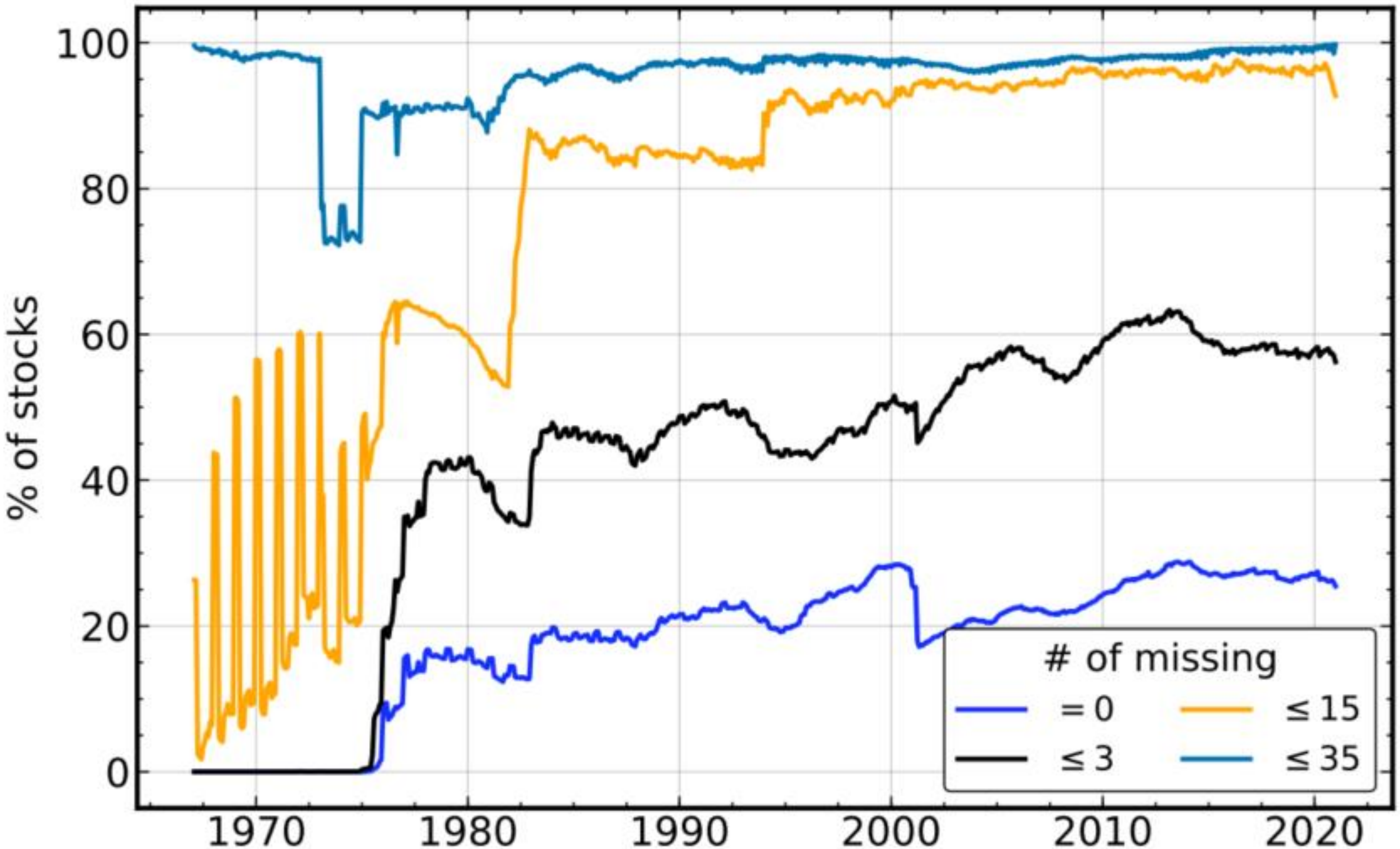
- 2.1. Data
  - CRSP/Compustat universe
  - January 1967 to December 2020
  - 45 characteristics
  - Converted into centered rank quantiles and scaled to be in the $[-0.5, 0.5]$ interval
  - Updated monthly or quarterly→ mixed-frequency

2.2. How much data is missing?



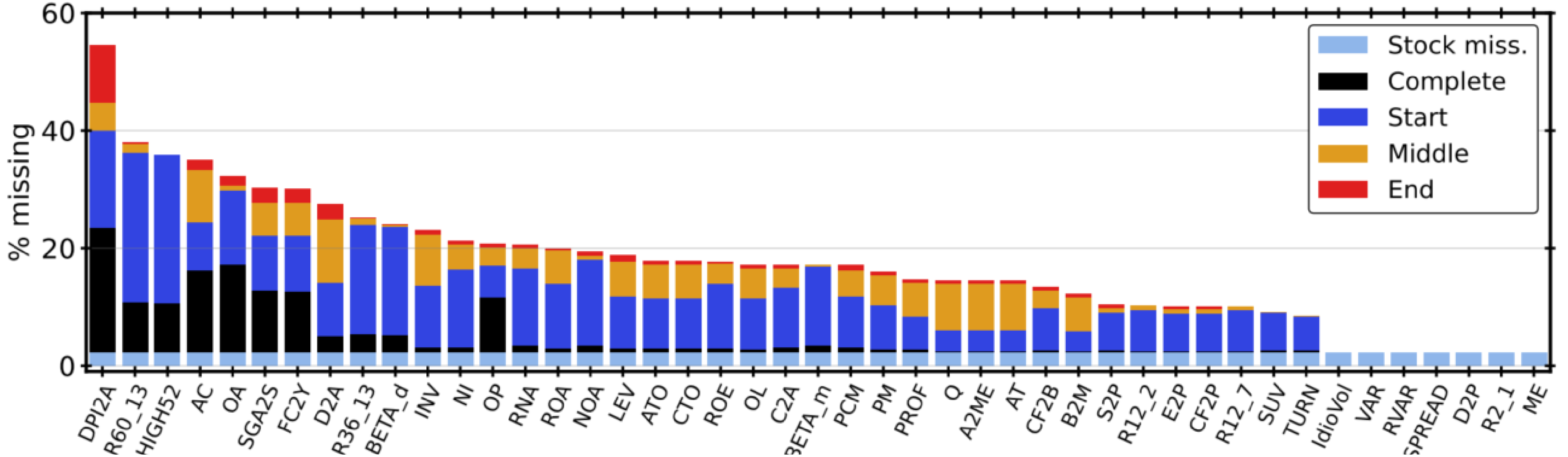
2.2. How much data is missing?

(e) Multiple Chars.

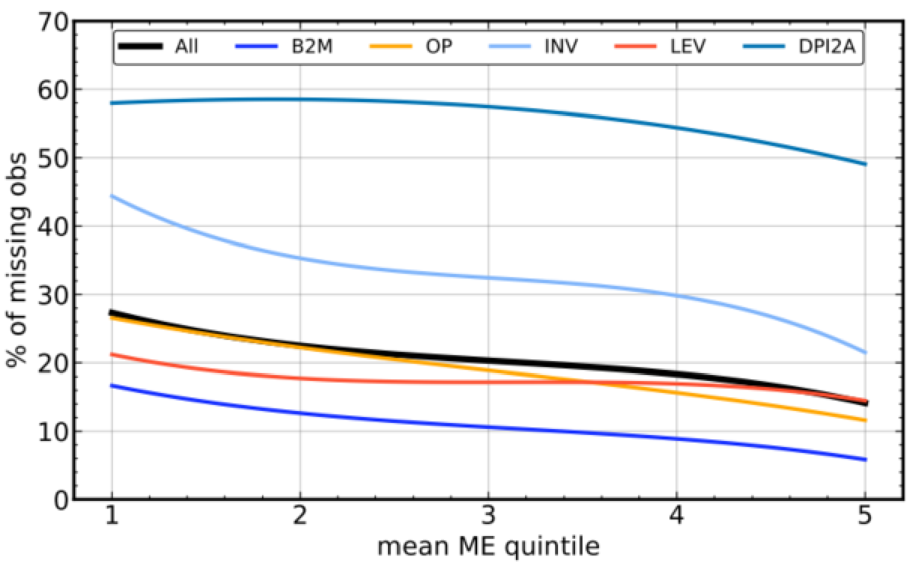


2.3. What is the structure of missingness?

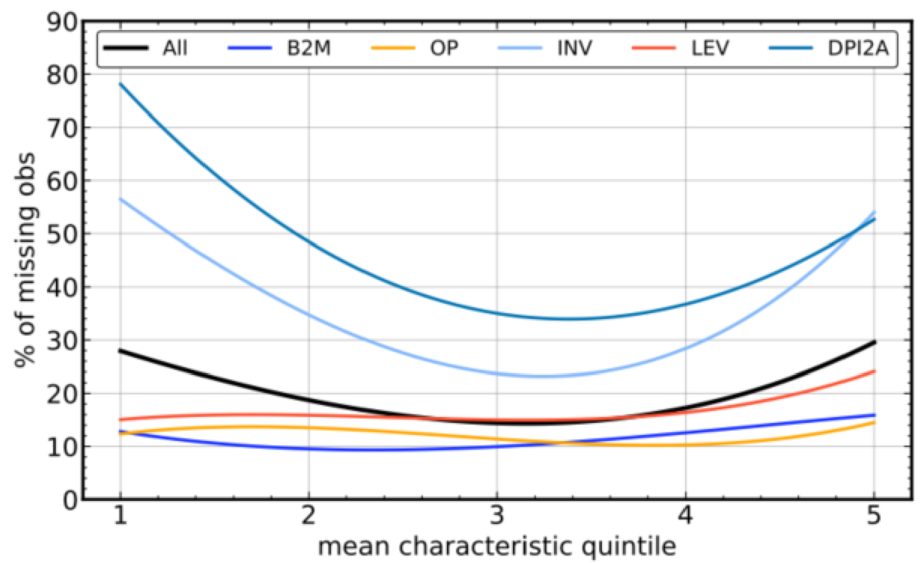
Figure 3: Missing Observations by Characteristic



(a) Mean ME Quintiles



(b) Mean Char. Quintiles



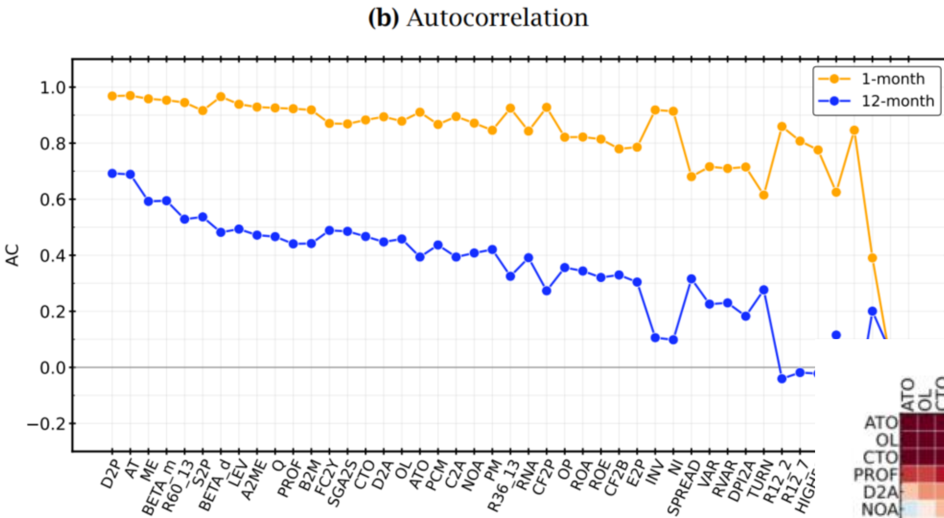
2.3. What is the structure of missingness?

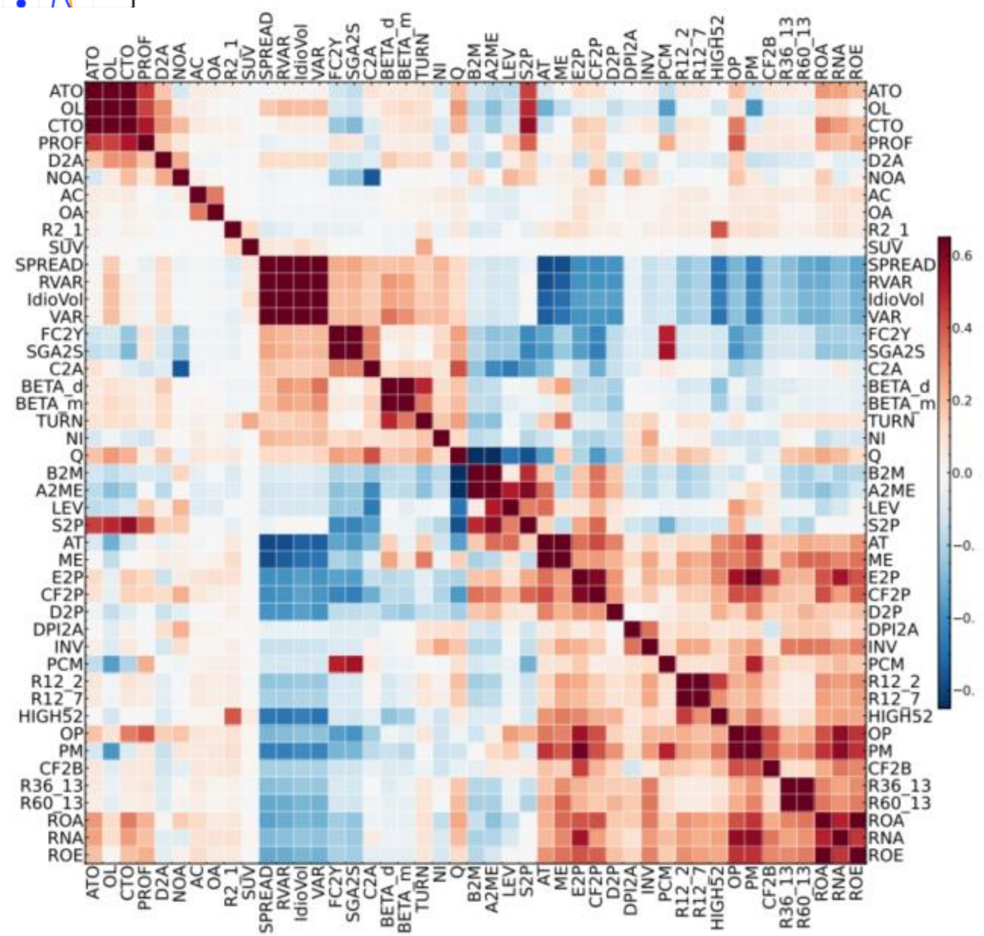
Table 2: Logistic regressions explaining missingess

D2P	IdioVol	ME	R2_1	SPREAD	TURN	VAR	FE	Last Val	Missing Gap	train AUC	test AUC
Missing at the beginning											
1.76*** [230.70]	-0.33*** [-22.57]	-1.30*** [-158.80]	0.07*** [11.87]	0.66*** [68.62]	0.53*** [91.24]	0.62*** [41.44]	F	F	F	0.50	0.51
1.85*** [176.56]	-0.28*** [-16.30]	-0.60*** [-63.71]	-0.07*** [-11.05]	0.63*** [55.74]	0.70*** [104.68]	0.43*** [24.44]	F	F	0.06*** [439.24]	0.64	0.65
							T	F	F	0.72	0.76
							T	F	0.02*** [186.00]	0.72	0.75
0.52*** [41.63]	-0.03*** [-1.33]	-0.64*** [-55.48]	0.10*** [13.42]	-0.06*** [-4.42]	-0.18*** [-21.58]	-0.31*** [-14.41]	T	F	0.02*** [174.11]	**0.74**	**0.77**

- Missingness is heterogeneous
- Characteristics are cross-sectionally correlated
- Missingness is correlated over time.

2.4. Characteristics Dependency





- persistent
- cross-sectionally correlated

3. Model

- Two fundamental challenges
  - Take advantage of **all available information**

    (imputing the cross-sectional median would incur an omitted variable bias)
  - The model for characteristics, that is estimated on the observed data, needs to be **valid on the unobserved data** as well

- Basic symbols
  - Our data set of month/stock/characteristic observations forms a three-dimensional vector space

$$C_{i,t,l} \qquad \text{with } i = 1, ..., N_t, t = 1, ..., T \text{ and } l = 1, ..., L.$$

  - The $N_t \times L$ matrix of characteristics at time t

$$C^t_{i,l} \qquad \text{with } i = 1, ..., N_t \text{ and } l = 1, ..., L$$

3.1. Cross-Sectional Information

- We start by estimating a low-dimensional cross-sectional factor model by PCA for each month t

$$C_{i,l}^t = F_i^t \Lambda_l^{t\top} + e_{i,l}^t \qquad \text{with } i = 1, ..., N_t \text{ and } l = 1, ..., L.$$

a K factor model $F^t \in \mathbb{R}^{N_t \times K}$ and $\Lambda^t \in \mathbb{R}^{L \times K}$

- Without missing values: apply a simple PCA to $C^t C^{t\top}$ to get K eigenvectors $F^t \in \mathbb{R}^{N_t \times K}$

$$\frac{1}{L}\sum_{l=1}^{L} C_l^t C_l^{t\top}.$$

- With missing values: estimate F^t as the eigenvectors of the K largest eigenvalues of (Xiong and Pelger,2019)

$$\tilde{\Sigma}_{i,j}^{\mathrm{XS},t} = \frac{1}{|Q_{i,j}^t|}\sum_{l \in Q_{i,j}^t} C_{i,l}^t C_{j,l}^t,$$

where $Q_{i,j}^t$ is the set of all characteristics which are observed for the two stocks i and j at time t

3.1. Cross-Sectional Information

- We start by estimating a low-dimensional cross-sectional factor model by PCA for each month t

$$C_{i,l}^t = F_i^t \Lambda_l^{t\top} + e_{i,l}^t \qquad \text{with } i = 1, ..., N_t \text{ and } l = 1, ..., L.$$

- The characteristic loadings

$$\hat{\Lambda}_l^t = \left(\sum_{i=1}^{N_t} W_{i,l}^t \hat{F}_i^t \hat{F}_i^{t\top} \right)^{-1} \left(\sum_{i=1}^{N_t} W_{i,l}^t \hat{F}_i^t C_{i,l}^t \right),$$

- where $W_{i,l}^t = 1$ if characteristic l is observed for stock i at time t and $W_{i,l}^t = 0$ otherwise.
- The "loadings" Λ are close to constant over time

$$C_{i,l}^t = F_i^t \Lambda_l^\top + e_{i,l}^t \qquad \text{with } i = 1, ..., N_t \text{ and } l = 1, ..., L.$$

- A pooled regression

$$\hat{\Lambda}_l = \left(\sum_{t=1}^{T} \left(\sum_{i=1}^{N_t} W_{i,l}^t F_i^t F_i^{t\top} \right) \right)^{-1} \left(\sum_{t=1}^{T} \left(\sum_{i=1}^{N_t} W_{i,l}^t F_i^t C_{i,l}^t \right) \right)$$

3.2. Time-Series Information

- Combine the XS (cross-sectional) information with TS (time-series) information
- Backward cross-sectional model (B-XS) :

$$\hat{C}_{i,t}^{l,\text{B-XS}} = \beta^{l,\text{B-XS}\top} \begin{pmatrix} C_{i,t-1}^{l} & \hat{F}_{i,1}^{t} & \cdots & \hat{F}_{i,K}^{t} \end{pmatrix}$$

- Backward-forward-cross-sectional model (BF-XS):

$$\hat{C}_{i,t}^{l,\text{BF-XS}} = \beta^{l,\text{BF-XS}\top} \begin{pmatrix} C_{i,t-1}^{l} & C_{i,t+1}^{l} & \hat{F}_{i,1}^{t} & \cdots & \hat{F}_{i,K}^{t} \end{pmatrix}$$

- For a given set of cross-sectional and time-series information in the vector $X_i^{l,t}$
  - the local regression $\hat{\beta}^{l,t} = \left(\sum_{i=1}^{N_t} W_{i,l}^{t} X_i^{l,t} X_i^{l,t\top} \right)^{-1} \left(\sum_{i=1}^{N_t} W_{i,l}^{t} X_i^{l,t} C_{i,t}^{l} \right)$
  - the global regression $\hat{\beta}^{l} = \left(\sum_{t=1}^{T} \left(\sum_{i=1}^{N_t} W_{i,l}^{t} X_i^{l,t} X_i^{l,t\top} \right) \right)^{-1} \left(\sum_{t=1}^{T} \left(\sum_{i=1}^{N_t} W_{i,l}^{t} X_i^{l,t} C_{i,t}^{l} \right) \right)$

3.2. Time-Series Information

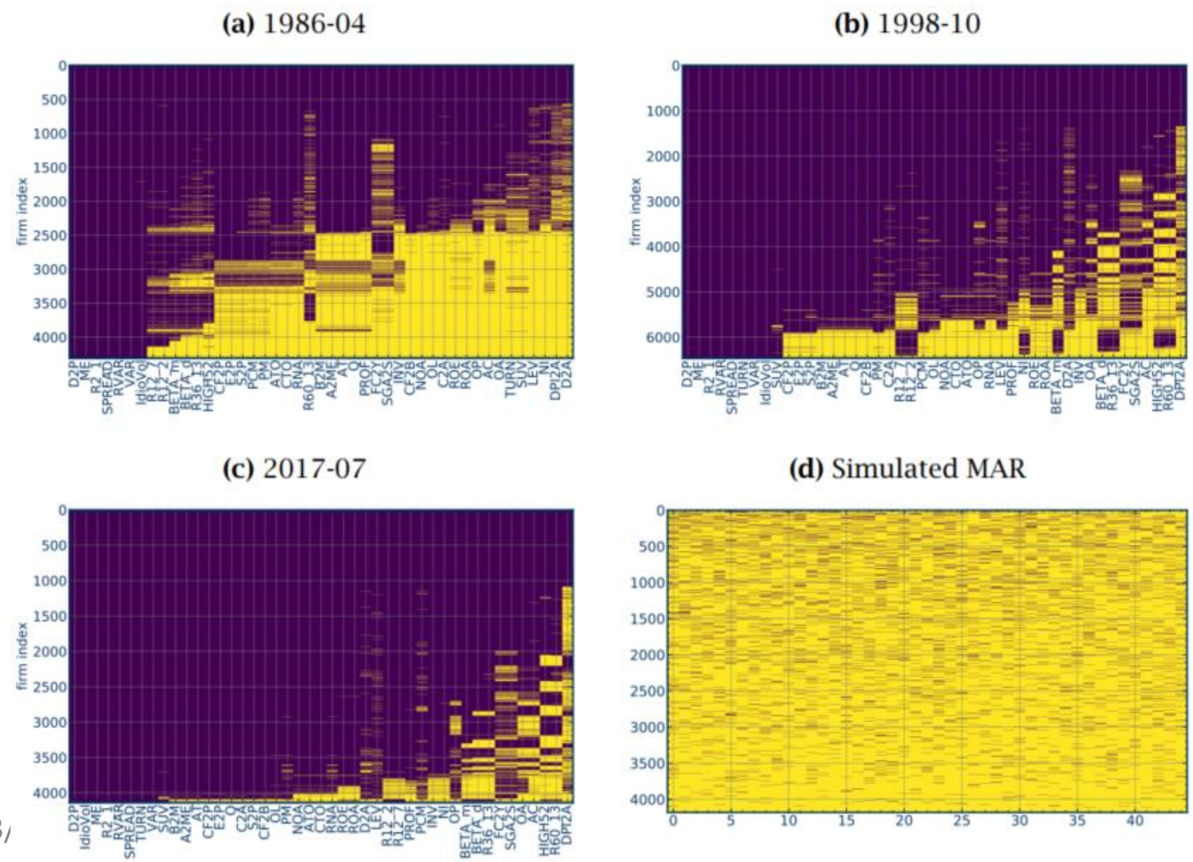
- Different Imputation Methods

Method	Estimation
Backward-Forward-XS (BF-XS)	$\hat{C}_{i,t}^{\text{BF-XS}} = (\hat{\beta}^{\text{BF-XS}})^\top \begin{pmatrix} C_{i,t-1}^l & C_{i,t+1}^l & \hat{F}_{i,1}^l & \cdots & \hat{F}_{i,K}^l \end{pmatrix}$
Backward-XS (B-XS)	$\hat{C}_{i,t}^{\text{B-XS}} = (\hat{\beta}^{\text{B-XS}})^\top \begin{pmatrix} C_{i,t-1}^l & \hat{F}_{i,1}^l & \cdots & \hat{F}_{i,K}^l \end{pmatrix}$
Forward-XS (F-XS)	$\hat{C}_{i,t}^{\text{F-XS}} = (\hat{\beta}^{\text{F-XS}})^\top \begin{pmatrix} C_{i,t+1}^l & \hat{F}_{i,1}^l & \cdots & \hat{F}_{i,K}^l \end{pmatrix}$
Cross-sectional (XS)	$\hat{C}_{i,t}^{\text{XS}} = (\hat{\beta}^{\text{XS}})^\top \begin{pmatrix} \hat{F}_{i,1}^l & \cdots & \hat{F}_{i,K}^l \end{pmatrix}$
Time-series (B)	$\hat{C}_{i,t}^{\text{B}} = (\hat{\beta}^{\text{B}})^\top \begin{pmatrix} C_{i,t-1}^l \end{pmatrix}$
Previous value (PV)	$\hat{C}_{i,t}^{\text{PV}} = C_{i,t-1}^l$
Cross-sectional median	$\hat{C}_{i,t}^{\text{median}} = 0$

3.3. Distribution of Missingness

- Characteristics are not missing at random
- A machine learning application with random masking on the training data, could lead to a bias in imputed values

Figure 7: Joint Distribution of Missing Patterns



- 3.4. Look-ahead bias
  - Backward (B-XS) model
  - Backward-Forward (BF-XS) model ✓

- 3.5. Rank normalization vs. raw characteristics
  - deal with the outliers
  - achieve stationarity

- 3.6. Evaluation metrics

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{L}\sum_{l=1}^{L}\frac{1}{N_t}\sum_{i=1}^{N_t}\left(C_{i,t,l} - \hat{C}_{i,t,l}\right)^2}.$$

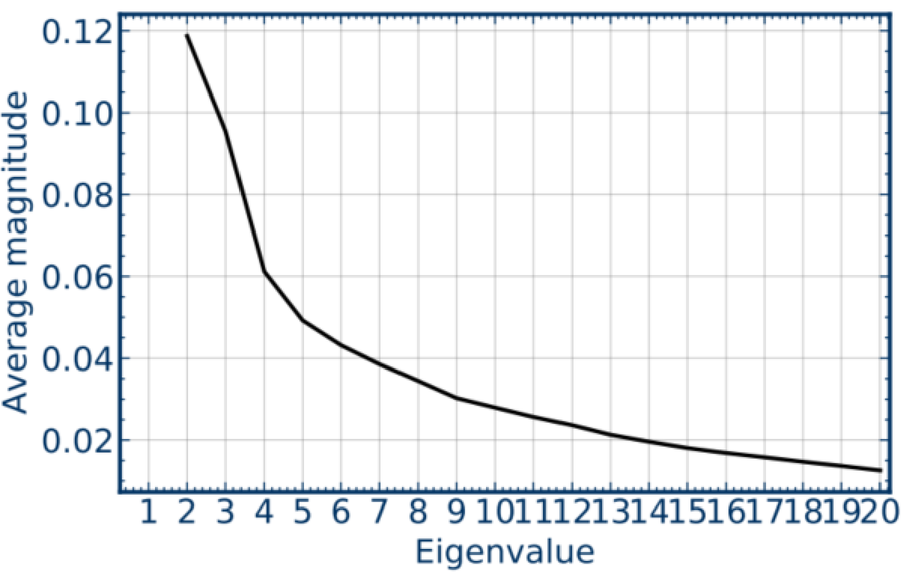
•

4. Factor Structure in Characteristics

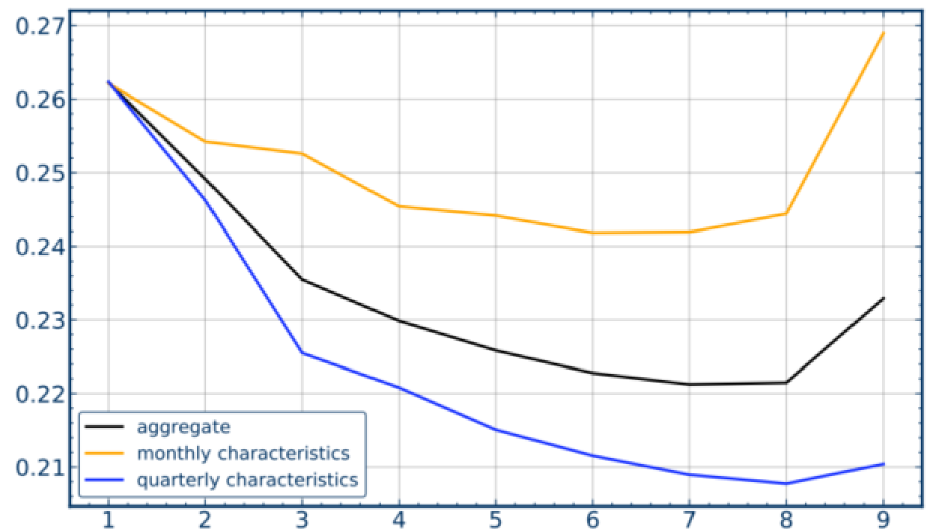
➢ 4.1. Number of factors

- strong evidence for a factor structure
- select the number of factors by minimizing the out-of-sample RMSE→ six factors

(a) Eigenvalues of $\Sigma^{XS,t}$



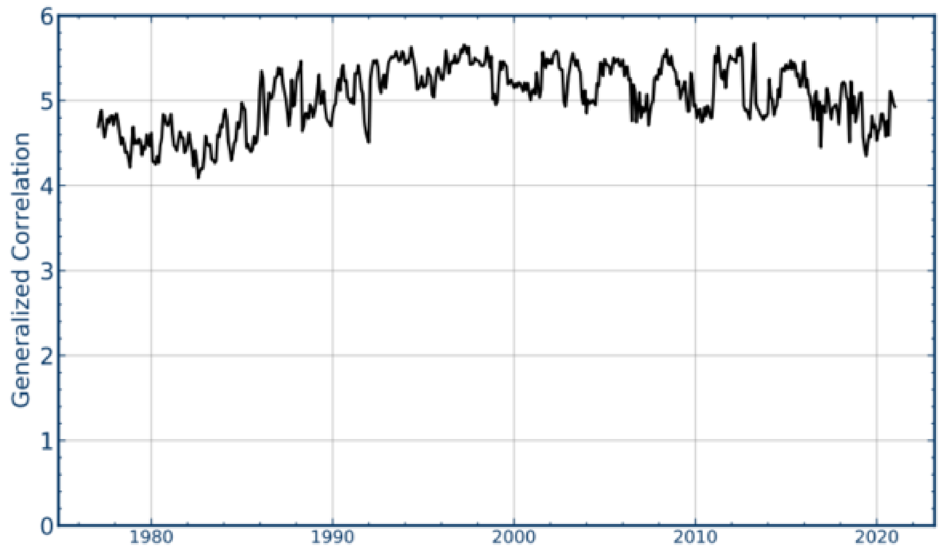
(b) Out-of-Sample RMSE



4. Factor Structure in Characteristics

- 4.2. Local vs. global factors
  - The loading structure of the cross-sectional factor model is **relatively stable** over time.

Figure 9: Generalized Correlation of Global and Local Factor Weights



$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{L}\sum_{l=1}^{L}\frac{1}{N_t}\sum_{i=1}^{N_t}\left(C_{i,t,l} - \hat{C}_{i,t,l}\right)^2}.$$

4. Factor Structure in Characteristics

- 4.3. Structure of factors
  - factors have a meaningful economic interpretation
  - linked to characteristic categories: e.g. value characteristics

- 4.4. Rank normalization vs. raw characteristics
  - the rank quantile space is appropriate for the latent factor model and provides better results

5. Imputation

- 5.1. Aggregate comparison between methods

Table 4: 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

- Mask 10% of the data either missing at random or missing in time-series blocks for 12 consecutive months

5. Imputation

➢ 5.2. Imputation results for different types of missingness

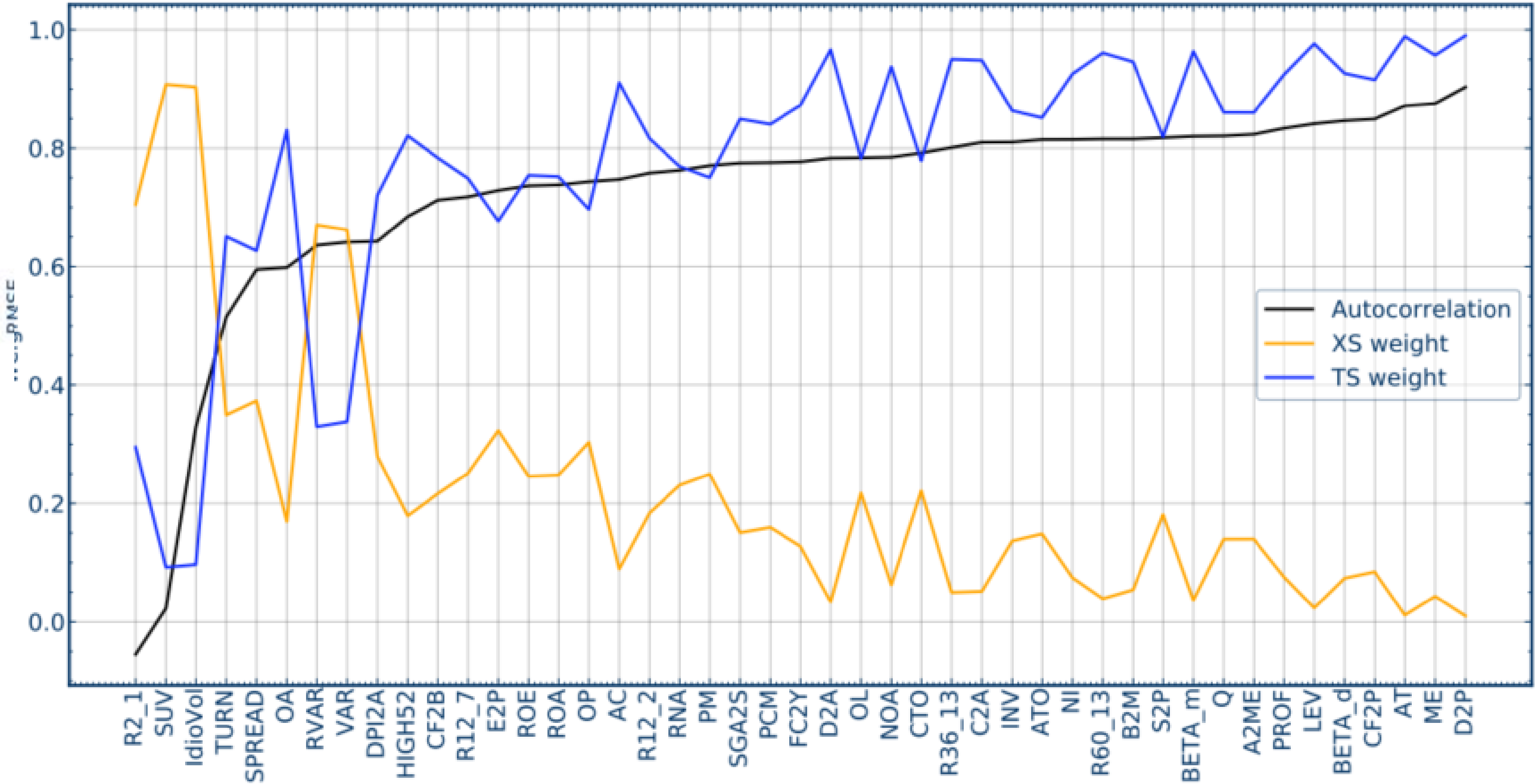
Table 5: Imputation Error for Types of Missingness

	In-Sample			OOS MAR			OOS Block		
Method	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly
				Start of the sample					
				End of the sample					
global BF-XS	-	-	-	-	-	-	-	-	-
global F-XS	-	-	-	-	-	-	-	-	-
global B-XS	0.19	0.21	0.16	0.19	0.20	0.17	0.21	0.21	0.21
global XS	0.24	0.25	0.22	0.27	0.26	0.28	0.25	0.24	0.26
global B	0.21	0.23	0.18	0.20	0.22	0.18	0.23	0.24	0.23
local B-XS	0.20	0.22	0.16	0.19	0.21	0.17	0.22	0.22	0.22
local XS	0.27	0.27	0.26	0.28	0.27	0.30	0.25	0.25	0.26
prev	0.23	0.24	0.21	0.22	0.23	0.21	0.26	0.25	0.26
local B	0.21	0.23	0.18	0.20	0.22	0.18	0.23	0.23	0.23
XS-median	0.35	0.36	0.34	0.33	0.33	0.33	0.32	0.32	0.31
ind-median	0.35	0.36	0.34	0.33	0.33	0.33	0.32	0.32	0.31

5. Imputation

- 5.3. Which information matters?

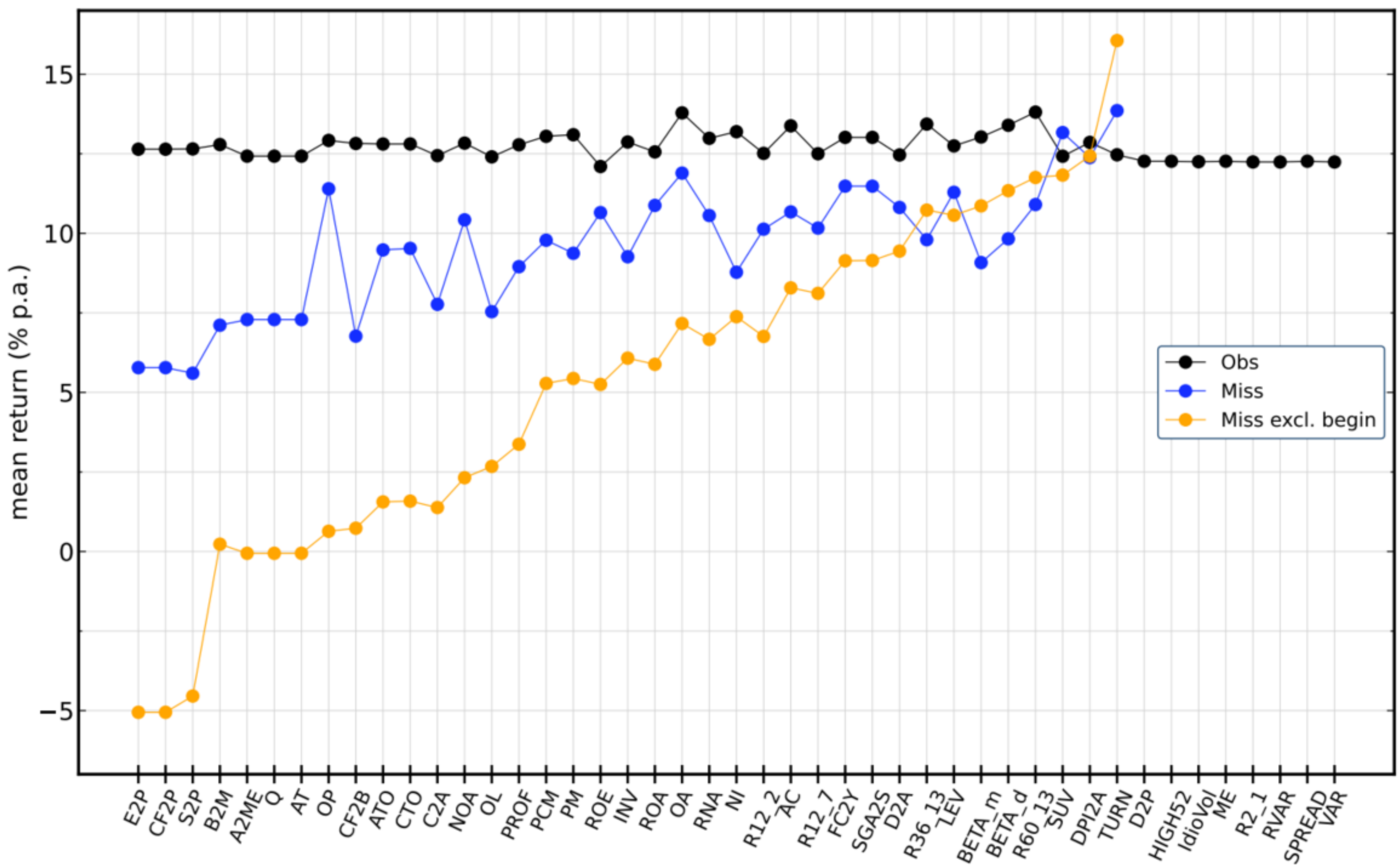
Figure 12: Information used for Imputation



6. Asset Pricing

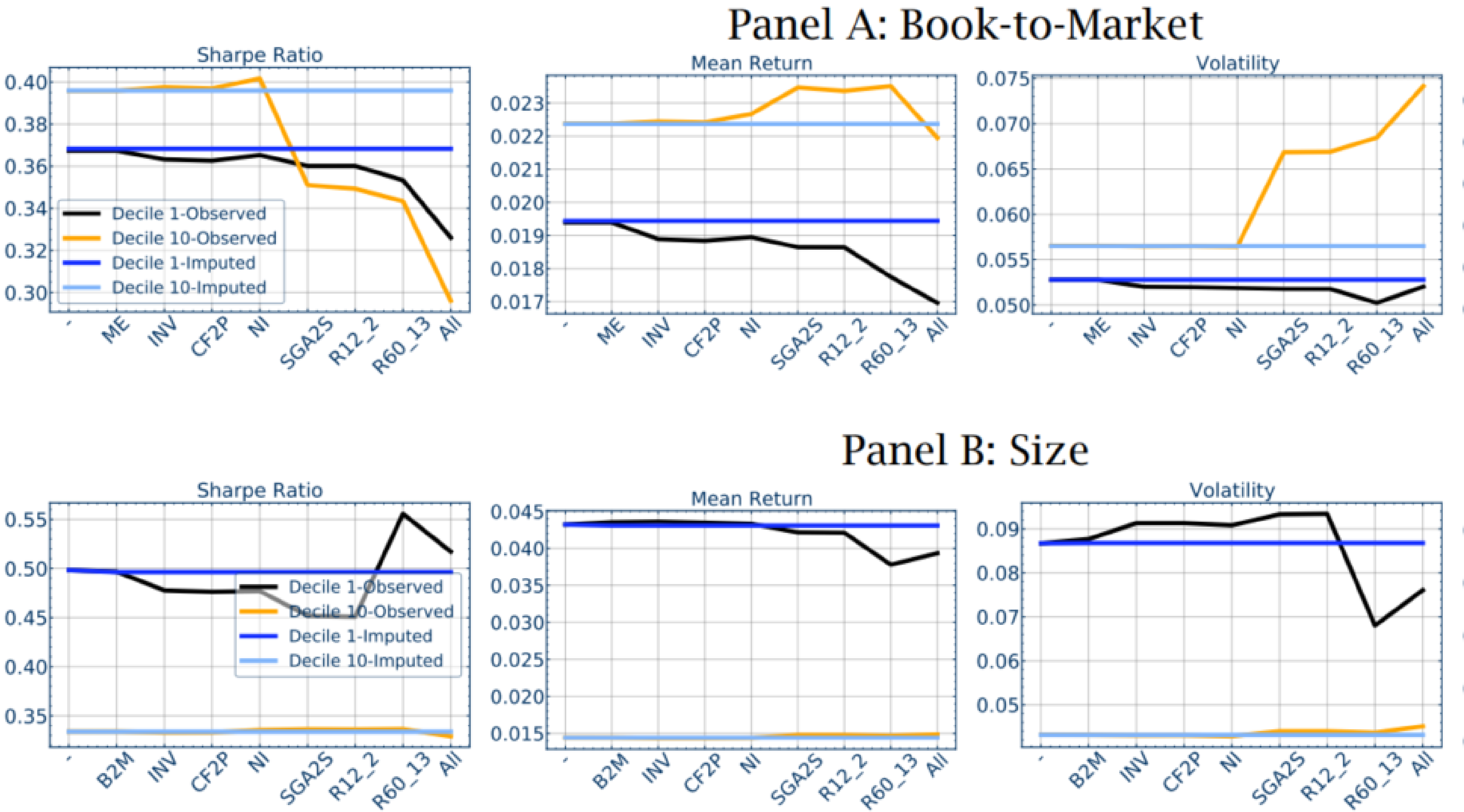
- 6.1. Market strategy with observables

Figure 14: Market-wide investment strtagy



6. Asset Pricing

➢ 6.2. Conditional sorts



7. Conclusion

- This paper focuses on a very widespread yet rarely recognized issue of missing data in firm-specific characteristics.

- We propose a new imputation method, which is easy to use, and substantially outperforms existing alternatives.