Missing Financial Data: Filling the Tensor Blanks

Jiaxi Li

Department of Economics

October 3rd, 2025

THE UNIVERSITY *of* NORTH CAROLINA *at* CHAPEL HILL

1. Introduction

2. Data

3. Method

4. Empirical Results

5. Application to Asset Pricing

Introduction

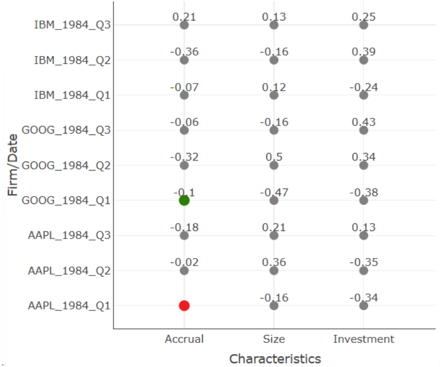
Missing Firm Characteristics

- Firm financial data play a crucial role in various applications, particularly in asset pricing.
- According to Bryzgalova et al. (2024), more than 70% of firms-representing approximately half of the market capitalization-have missing characteristics.
- Relying solely on fully observed firms can lead to biased estimates and efficiency loss.
- Common imputation methods, such as cross-sectional means (medians) or previous values, often result in sizable imputation errors.

Firm Characteristics

Here is a hypothetical **quarterly firm characteristics panel data** represented as a 2D matrix in a visualization:

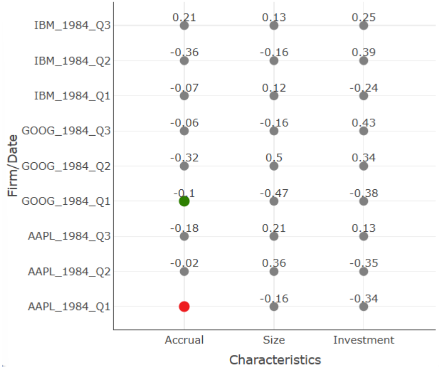
- Two different information dimensions for each row.
- The information can be reorganized into a **3D tensor**.

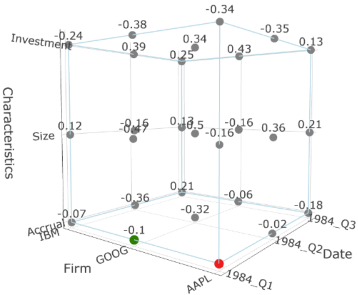


Firm Characteristics

Here is a hypothetical **quarterly firm characteristics panel data** represented as a 2D matrix in a visualization:

- Two different information dimensions for each row.
- The information can be reorganized into a **3D tensor**.





This project

- **New model**: MPTPCA (Mean-projected Tensor PCA) for missing financial data
- **Accuracy**: Adds imputation power beyond time-series & cross-sectional methods
- **Usefulness**: Improves return prediction and portfolio performance

Related Literature

Some missingness firm characteristics imputation literature:

- Freyberger et al. (2024): Proposes a GMM framework for asset pricing that incorporates imputed values and derives inference.
- Chen and McCoy (2024): Uses the Expectation-Maximization (EM) algorithm to impute missing firm characteristics and construct machine-learning-based portfolios for asset pricing.
- Bryzgalova et al. (2024): Combines latent cross-sectional factor models with robust regularized regression to improve characteristic imputation.

Recent return prediction literature:

- Gu et al. (2020): Compared different methods in Machine Learning for Return Prediction. Suggests that Neural Network Model is one of the best in Return Prediction.
- Kelly et al. (2024): Demonstrated that the with a complex model where $p >> n$, the out-of-sample prediction can be improved.
- Shen and Xiu (2025): Showed that machine learning methods (especially with dense regularization) can learn weak signals.

Data

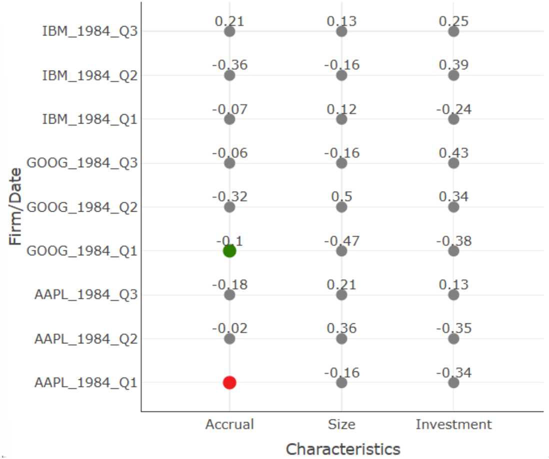
Firm Characteristics Data

- Source: Firm characteristics data from Bryzgalova et al.'s replication dataset (originally from Compustat).
- Dates: Q1 1984 - Q4 2020.
- Characteristics: 45 Characteristics converted into centered rank quantiles and scaled to be in the $[-0.5, 0.5]$ interval
- Frequency: Quarterly
- Time Intervals: Data is grouped into 5-year intervals for tensor estimation.

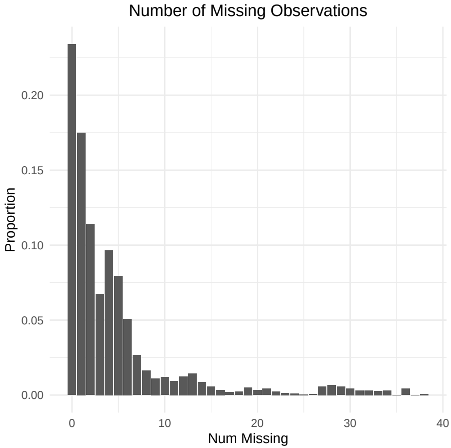
Dataset Dimensions: Date x Firm x Characteristics (148 x 19242 x 45).

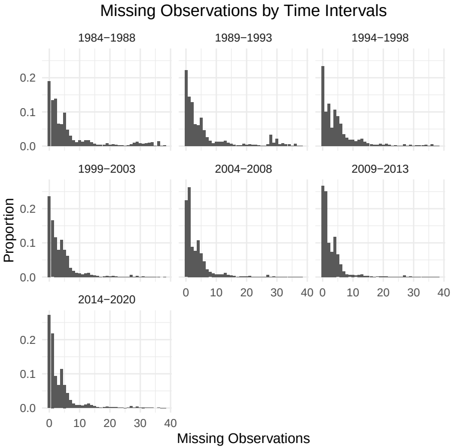
Revisit the Panel Structure

Here is the data structure: each observation is a firm at a specific time. There are 45 different Characteristics while some might be missing.

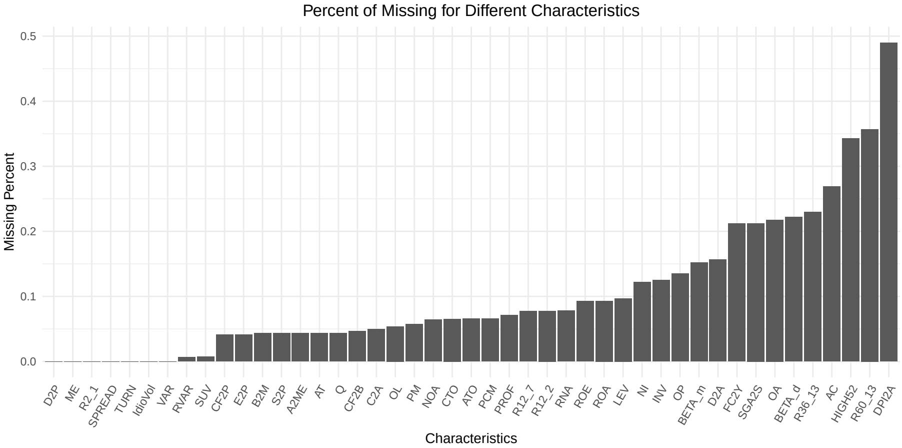


Number of Observations

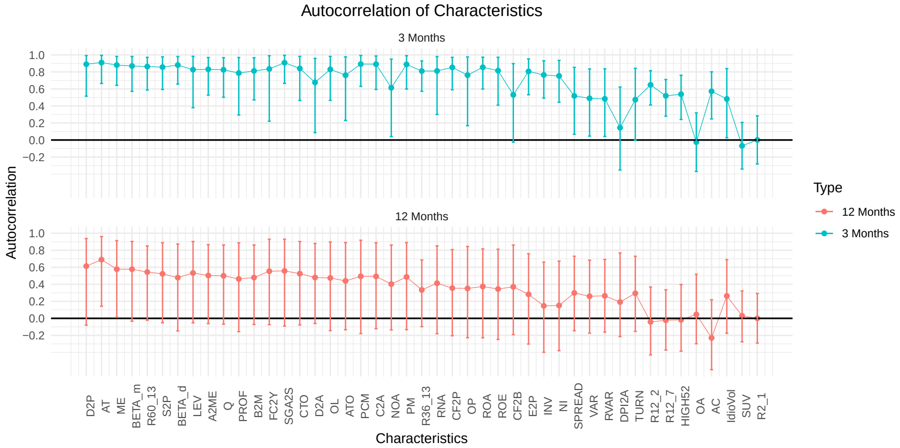




Fraction of Missing Each Characteristic



Characteristics Autocorrelation



Method

PCA with Regularization for Missing Data (XS)

Bryzgalova et al. (2024) introduced a financial data imputation method based on Ridge regression to address missing data issues in a relative small cross section of characteristics. They also

For each time t:

- Estimate sample covariance matrix $\hat{\Sigma}^t$ of C using only observed entries.
- Compute loadings, $\hat{\Lambda}^t$, based on eigenvalue scaled first R principal components of $\hat{\Sigma}^t$: $\hat{\Lambda}^t = \hat{V}^t \hat{D}^t$
- Estimate factors, $\hat{F}^t$, using Ridge regression:

$$\hat{F}_i^t = (\sum_{l=1}^{L} W_{i,l}^t \hat{\Lambda}_l^t (\hat{\Lambda}_l^t)^\top + \gamma I_R)^{-1} (\sum_{l=1}^{L} W_{i,l}^t \hat{\Lambda}_l^t C_{i,l}^t)$$

- Compute the complete panel $\hat{C}_{i,l}^t = \hat{F}_i^t (\hat{\Lambda}_l^t)^\top$.

This is referred to as the Cross-sectional (**XS**) method in Bryzgalova et al. (2024).

Backward XS (BXS)

In Bryzgalova et al. (2024), they also showed including backward information can improve the imputation accuracy (Motivated by a cross-sectional model with factor being an AR(1) process). Here is the Backward Cross-sectional method:

- Obtain the estimation from XS method, $\hat{C}^{XS}_{t,i,l}$.
- For each t, l, regress $C_{t,i,l}$ against $X^{BXS}_{t,i,l} = [\hat{C}^{XS}_{t,i,l} \quad C_{t-1,i,l} \quad \hat{e}^{XS}_{t-1,i,l}]$, where $\hat{e}^{XS}_{t-1,i,l}$ is the residual of the XS method in previous period.

$$\hat{\beta}^{l,t,BXS} = (\sum_{i=1}^{N} W_{t,i,l} X^{XS}_{t,i,l} (X^{XS}_{t,i,l})^\top)^{-1} (\sum_{i=1}^{N} W_{i,l} X^{XS}_{t,i,,l} C_{t,i,l})$$

- Compute the complete tensor $\hat{C}^{BXS}_{t,i,l} = (\hat{\beta}^{l,t,BXS})^\top X^{XS}_{t,i,l}$.

Mean Projected TPCA Model (MPTPCA)

Motivated by Babii et al. (2025), I am writing the characteristics with a tensor-structured model:

$$C_{t,i,l} = \sum_{r=1}^{R} \mu_{r,t} f_{r,i} \lambda_{r,l} + \epsilon_{t,i,l}$$

However, due to strong autocorrelation in characteristics data, time dimension orthogonality may not hold AND we have **missing data issue**. I will model the time dimension with an AR(1) process:

$$\mu_{r,t} = \frac{M_r}{1-\gamma_r} + \gamma_r \mu_{r,t-1} + \xi_{r,t}$$

- $\xi_{r,t}, \epsilon_{t,i,l}$ are white noise; $\xi_{r,t} \perp \epsilon_{t,i,l}$.
- $\gamma_r < 1$ for stationary.

MPTPCA Analysis

Here, I propose a new way to estimate the parameters (MPTPCA):

First, take conditional expectation of the model:

$$E[C_{t,i,l}|i,l] = \sum_{r=1}^{R} M_r f_{r,i} \lambda_{r,l}$$

The averaged matrix of the tensor data actually have a factor structure similar to the tensor data. Aggregating along the time dimension addresses autocorrelation and also the problem of missingness.

Notice that the $f_{r,i}$ and $\lambda_{r,l}$ are the same as the ones in the tensor model.

Potential Problem with Averaging

As I laid out in my field paper, there are two potential problem with aggregating (averaging) a 3D tensor into 2D matrix:

1. **Loss of "Weak" Factors**: $M_r = E[\mu_{r,t}|r] \approx 0$

   This should **NOT** be a big problem for Characteristics data because most of them have a non-zero mean.

2. **3D Factors become not identifiable in 2D**: $M_{r_1} \approx M_{r_2}$

   When the two factors have the same "strength" (singular value) in 2D, they become unidentifiable.

One may additionally fit a tensor model to the residuals of MPTPCA to address both issues.

Rewrite MPTPCA Model

Second, plug in μ_t into the main equation:

$$C_{t,i,l} = \sum_{r=1}^{R} \left(\frac{M_r}{1-\gamma_r} + \gamma_r \mu_{r,t-1} + \xi_{r,t}\right) f_{r,i}\lambda_{r,l} + \epsilon_{t,i,l}$$

Define $\bar{\gamma} \equiv \frac{1}{R}\sum_{r=1}^{R}\gamma_r$

$$C_{t,i,l} = \sum_{r=1}^{R} \left(\frac{M_r}{1-\gamma_r} + (\gamma_r - \bar{\gamma})\mu_{r,t-1} + \xi_{r,t}\right) f_{r,i}\lambda_{r,l} + \epsilon_{t,i,l} + \bar{\gamma}(C_{t-1,i,l} - \epsilon_{t-1,i,l})$$

By redefining $\rho_{r,t} \equiv \frac{M_r}{1-\gamma_r} + (\gamma_r - \bar{\gamma})\mu_{r,t-1} + \xi_{r,t}$ and $\eta_t \equiv \epsilon_{t,i,l} - \bar{\gamma}\epsilon_{t-1,i,l}$, we have

$$C_{t,i,l} = \sum_{r=1}^{R} \rho_{r,t} f_{r,i}\lambda_{r,l} + \bar{\gamma} C_{t-1,i,l} + \eta_{t,i,l}$$

This would be just a regression for estimation since we can get consistent estimates of $f_{r,i}\lambda_{r,l}$ from the averaged 2D model estimation.

MPTPCA Procedure

- Average the tensor along the time dimension to get $\bar{C}_{i,l}$ using all available data.
- Estimate covariance matrix $\hat{\Sigma}$ of $\bar{C}_{i,l}$ using only available entries
- Compute loadings $\hat{\Lambda} = \hat{V}\hat{D}^{1/2}$ from the first R principal components.
- Estimate factors, $\hat{F}$:

$$\hat{F}_i = (\sum_{l=1}^{L} W_{i,l}\hat{\Lambda}_l(\hat{\Lambda}_l)^{\top})^{-1}(\sum_{l=1}^{L} W_{i,l}\hat{\Lambda}_l\bar{C}_{i,l})$$

- Regress $C_{t,i,l}$ on $\{\hat{f}_{r,i}\hat{\lambda}_{r,l}\}_{r=1}^{R}$, $(\mathbf{1}_{\exists t-1,i,l} * C_{t-1,i,l})$ and $(\mathbf{1}_{\nexists t-1,i,l} \times \{\hat{f}_{r,i}\hat{\lambda}_{r,l}\}_{r=1}^{R})$ to obtain $\hat{\rho}_{r,l,t}$, $\hat{\gamma}_{t,l}$ and $\hat{\phi}_{r,t,l}$.
- Compute the final imputed values: $\hat{C}_{t,i,l} = \sum_{r=1}^{R} \hat{\rho}_{r,l,t}\hat{f}_{r,i}\hat{\lambda}_{r,l} + \hat{\gamma}_{t,l}(\mathbf{1}_{\exists,i,l} * C_{t-1,i,l}) + \sum_{r=1}^{R} \hat{\phi}_{r,l,t}(\mathbf{1}_{\nexists,i,l}\hat{f}_{r,i}\hat{\lambda}_{r,l})$

Imputation Methods Summary

Below are the imputation methods to compare in this project:

Method	Estimation
Cross-sectional Median (Median)	$\hat{C}^{median}_{t,i,l} = 0$
Previous Value (PV)	$\hat{C}^{PV}_{t,i,l} = C^{t-1}_{i,l}$
Backward (B)	$\hat{C}^{B}_{t,i,l} = (\hat{\beta}^{l,t,B})^\top C^{t-1}_{i,l}$
Backward-XS (B-XS)	$\hat{C}^{B-XS}_{t,i,l} = (\hat{\beta}^{l,t,B-XS})^\top \begin{bmatrix} \hat{C}^{XS}_{t,i,l} & C_{t-1,i,l} & \hat{e}^{XS}_{t-1,i,l} \end{bmatrix}$
Mean-Projected TPCA (MPTPCA)	$\hat{C}^{MPTPCA}_{t,i,l} = \sum_{r=1}^{R} \hat{\rho}_{r,l,t}\hat{f}_{r,i}\hat{\lambda}_{r,l} + \hat{\gamma}_{t,l}(\mathbf{1}_{\exists,i,l} * C_{t-1,i,l}) + \sum_{r=1}^{R} \hat{\phi}_{r,l,t}(\mathbf{1}_{\nexists,i,l}\hat{f}_{r,i}\hat{\lambda}_{r,l})$
Combined (Combined)	$\hat{C}^{Combined}_{t,i,l} = (\hat{\beta}^{l,t,Combined})^\top \begin{bmatrix} \hat{C}^{XS}_{t,i,l} & \hat{C}^{MPTPCA}_{t,i,l} & C_{t-1,i,l} & \hat{e}^{XS}_{t-1,i,l} & \hat{e}^{MPTPCA}_{t-1,i,l} \end{bmatrix}$

If $t-1$ value does not exist, use the most recent available value.

Method Comparison

We have three different types of imputation methods and there are some trade-offs among them:

1. Backward (B) Model:
   - Take advantage of time-series pattern for each characteristics.
   - It does not utilize cross-sectional information and cannot be applied to new firms.
2. Cross-sectional (XS) Model:
   - Capture the common cross-sectional patterns among all characteristics.
   - It does not utilize the time information. If we have a large amount of missing characteristics, loadings would be poorly estimated, resulting in large out-of-sample error.
3. Tensor Principal Component Analysis (TPCA) Model:
   - Utilize both the common time and cross-sectional pattern.
   - TPCA model captures only the common time and cross-sectional pattern. If factor structure is time-varying, especially with structural break, TPCA may not have a high predictive power.

Empirical Results

Masking

To assess the out-of-sample performance of each method, I introduce additional missing data by selectively masking certain entries.

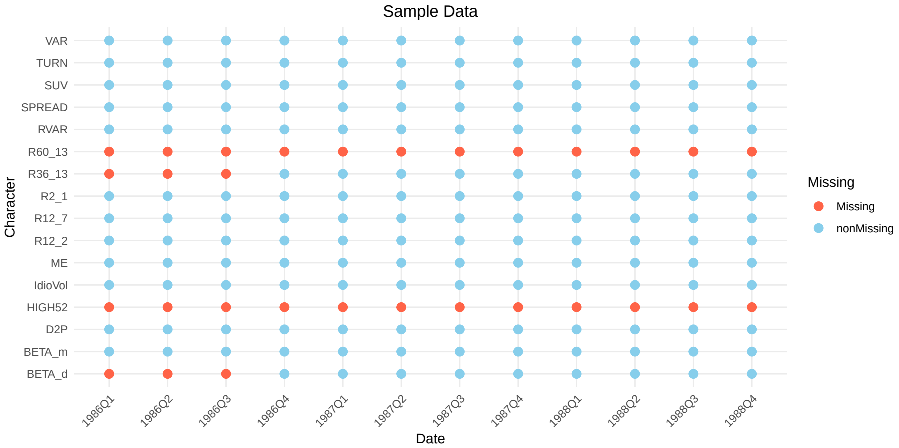
For each Characteristics l:

$$\mathit{Masking\ Amount}_l = 10\% * \mathit{Missing\ Amount}_l$$

- 10% ensures that the overall missing data structure remains largely intact while introducing a sufficient amount of masking for reliable out-of-sample evaluation.
- The formula ensures masking amount is proportional to the actual missing amount for each characteristics.

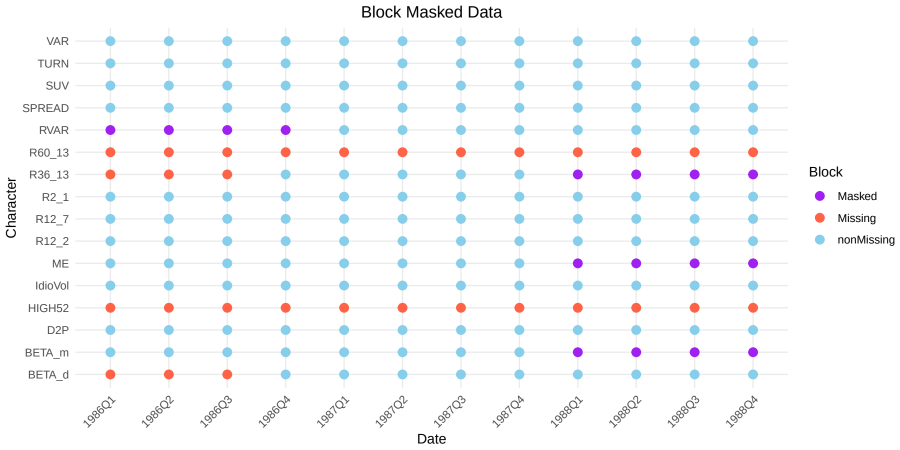
Masking Demonstration

Here is sample data from A.A. Importing Co., Inc. (ANTQ), covering the period from 1984 to 1988. During this time, several instances of block missingness naturally occur.

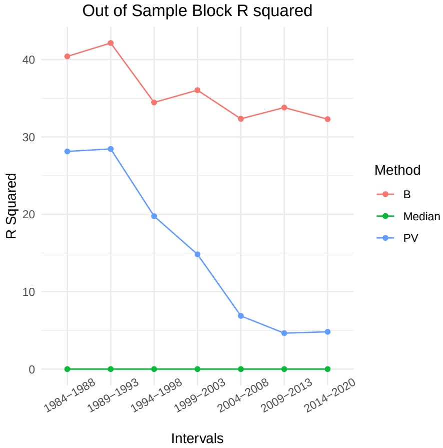


Block Masking Conditioning on Characteristics (Block)

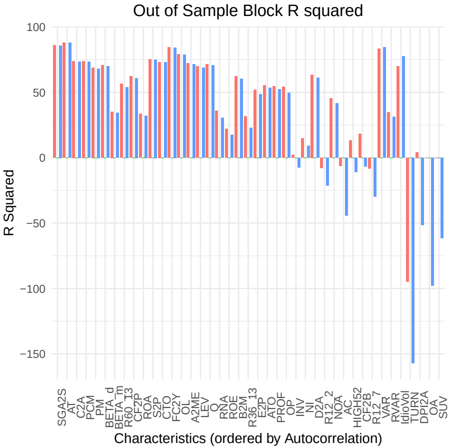
Block Masking Conditioning on Characteristics (Block): randomly masking additional one-year blocks based on the assigned masking probability.



Median, Previous Value and Backward (Block)



Out of Sample Block R squared

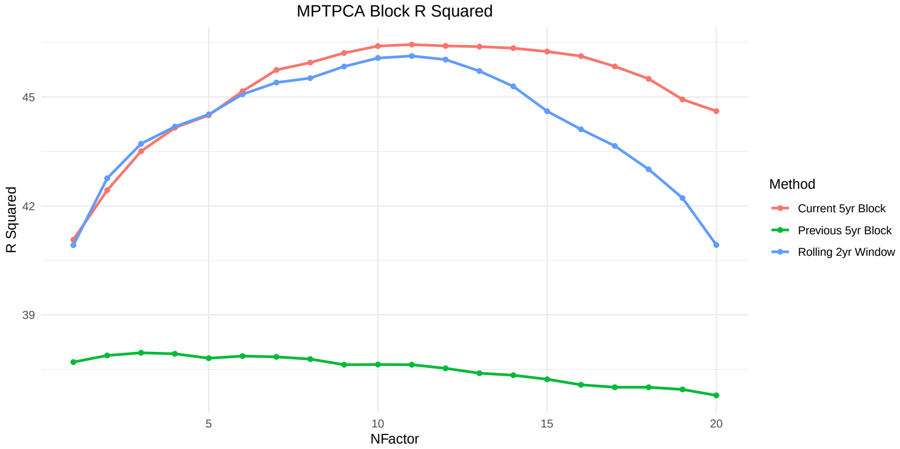


Out of Sample Block R squared

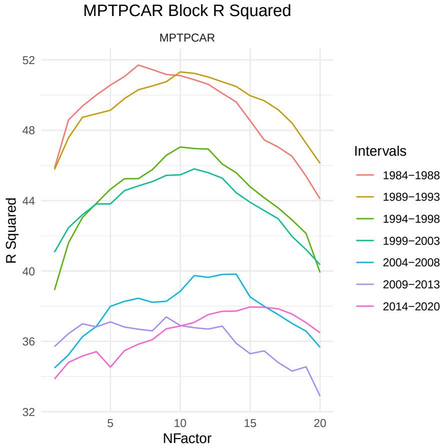
TPCA Models (Block, Overall)

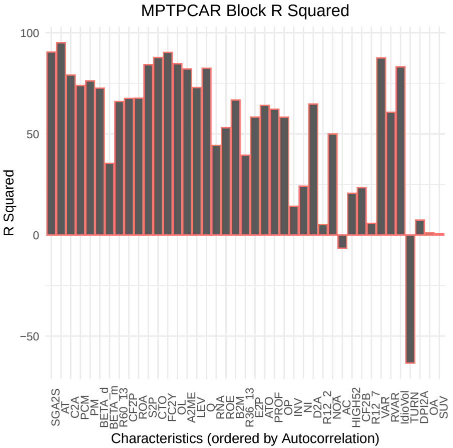
The 2-year-rolling-window-mean MPTPCA is denoted “MPTPCAR” in later section.

The 5-year-previous-mean MPTPCA is denoted “MPTPCA” in later section.

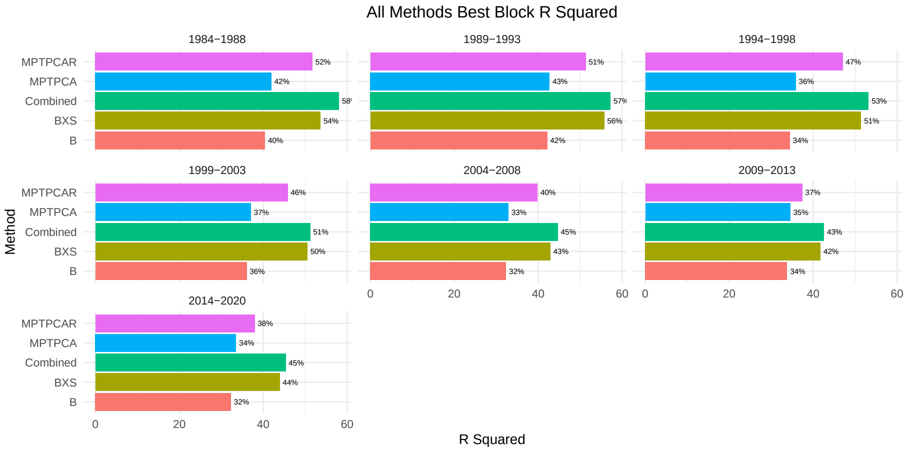


MPTPCAR (Block, Detailed)

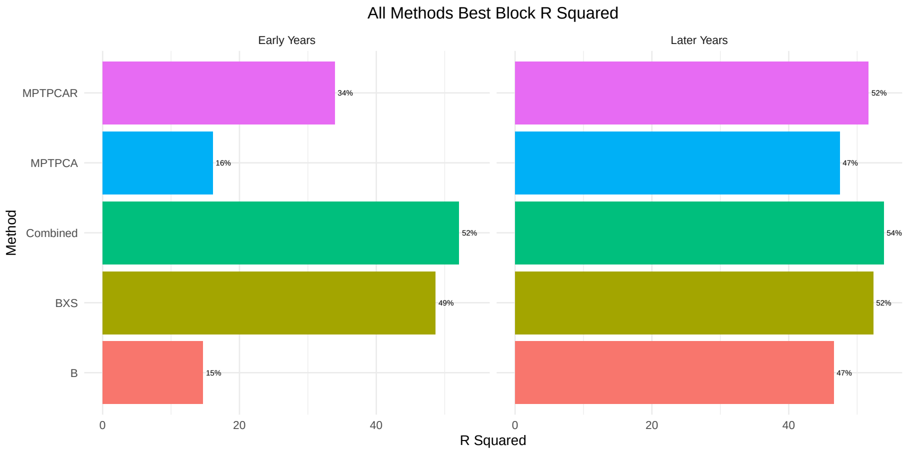




All Method Comparisons (Block, By Time Intervals)



All Method Comparisons (Block, By Time Early_Years)



Application to Asset Pricing

Return Prediction via Neural Network Model

The latent model of return is:

$$r_{i,t+1} = f(C_{i,t}) + \epsilon_{i,t+1},$$

Gu et al. (2020): Neural network (NN) model excels in return prediction but apply 0 to the missing characteristics value.

Our project: see whether imputation would bring value to the return prediction with NN model.

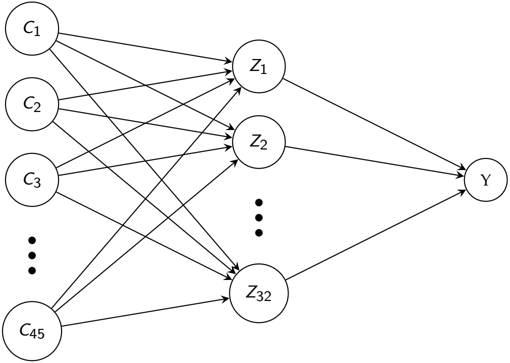
Data (Revisit)

- Source: Firm characteristics data from Bryzgalova et al.'s replication dataset (originally from Compustat), monthly return data from CRSP and monthly risk free rate from Ken French Data Library.
- Dates: Q1 1984 - Q4 2020.
- Characteristics: 45 Characteristics converted into centered rank quantiles and scaled to be in the $[-0.5, 0.5]$ interval
- Frequency: Quarterly

Characteristics Data Dimensions: Date x Firm x Characteristics (148 x 19242 x 45). Return Data Dimensions: Date x Firm (145 x 22263).

Neural Network with 1 Hidden Layer

C's: observed characteristics (with 0 for missing values)



Hidden Layer Hidden Layer Output

Neural Network Model Settings

Here, I would follow similar setting as Gu et al. (2020):

- ReLu as activation function
- Adaptive moment estimation algorithm (Adam) for Stochastic Gradient Descent (SGD)
- L1 regularization (Voided)
- Early Stopping
- Batch Normalization
- Ensemble
- Similar sample spiting: 30% training (1984-1994), 20% validation (1995-2001), 50% testing (2002-2020). Update model once every year for testing.

Special Considerations

- Use 2-year-rolling-window for characteristics imputation via MPTPCA to avoid look-ahead bias.
- Include delisted returns to mitigate survivorship bias.
- Use six-month-lag characteristics to predict returns to avoid look-ahead bias.
- Rely on Early Stopping instead of L1 normalization.
- Use the first training and validation set to determine optimal learning rate and batch size.
- **Do not simply plug in the imputed value.** Rather predict with two steps:
  1. Fit a Neural Network model with 0 as imputation.
  2. Run a Ridge Regression, regressing the residual on the imputation.

Fitting the Residual

In the first step, the neural network may overfit observed characteristics to compensate for missing values.

To address this, I include observed characteristics again in the second-step regression.

Also, the regression function can be seen as:

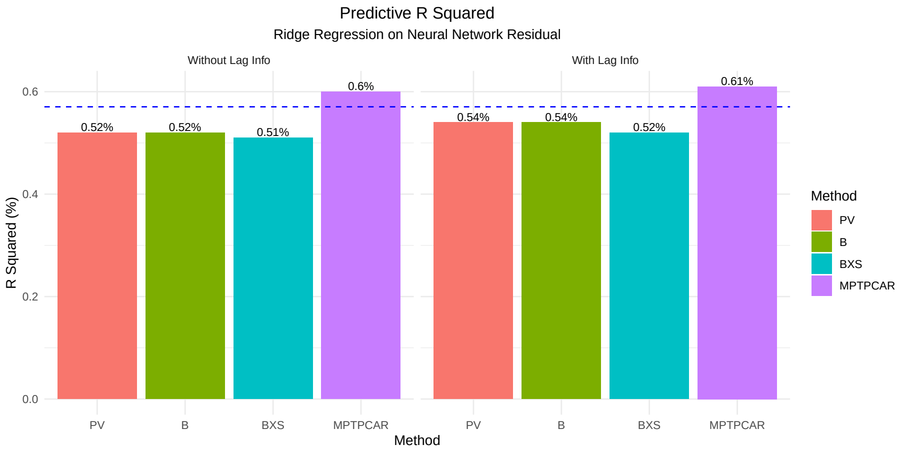
$$\hat{f}(C_{Obs}, C_{Im}) = Pred(IM(C_{Im}))$$

where the return predictor $\hat{f}$ combines:

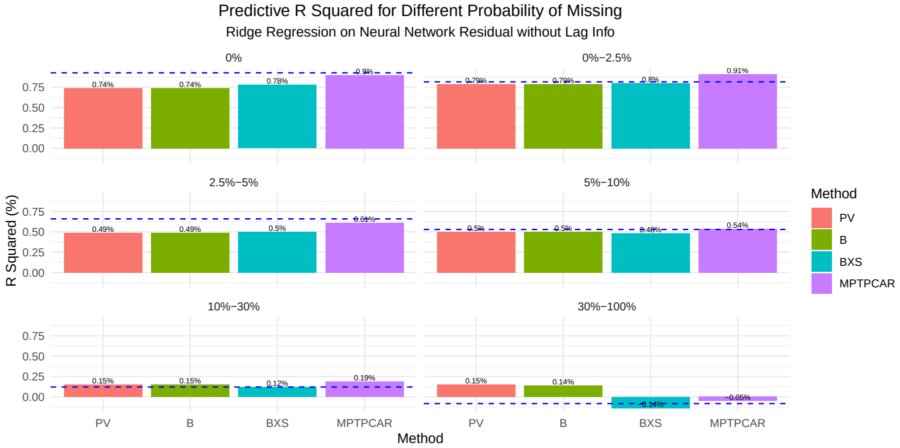
- Imputation function: $IM(X)$
- Prediction function: $Pred(X)$

Thus, the regression performs imputation and prediction simultaneously, using our feed-in imputation as input to its own.

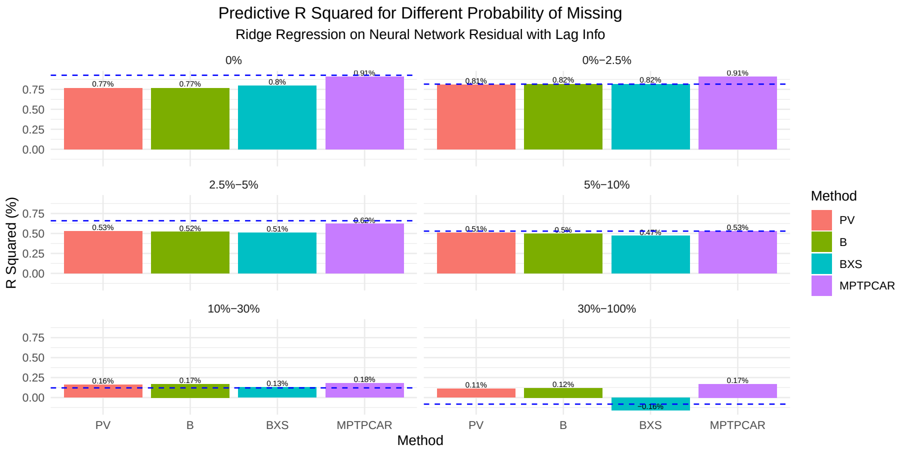
Predicting Results



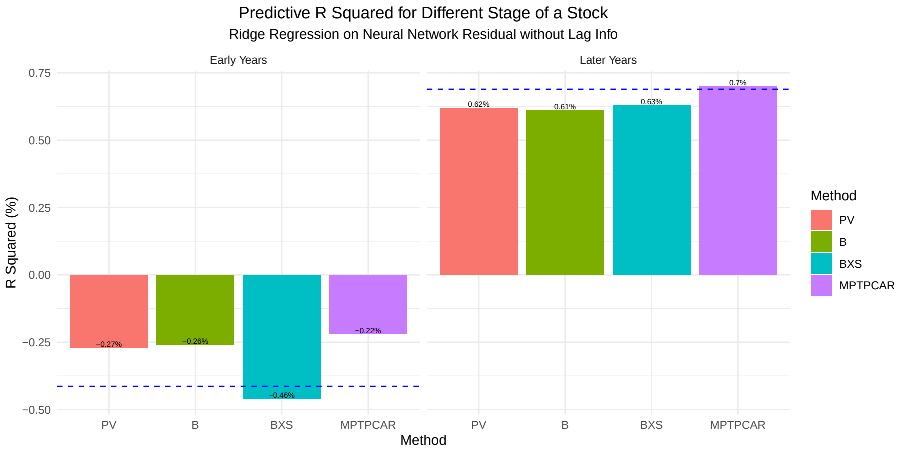
Missing Stratified Predicting Results (Without Lag Info)



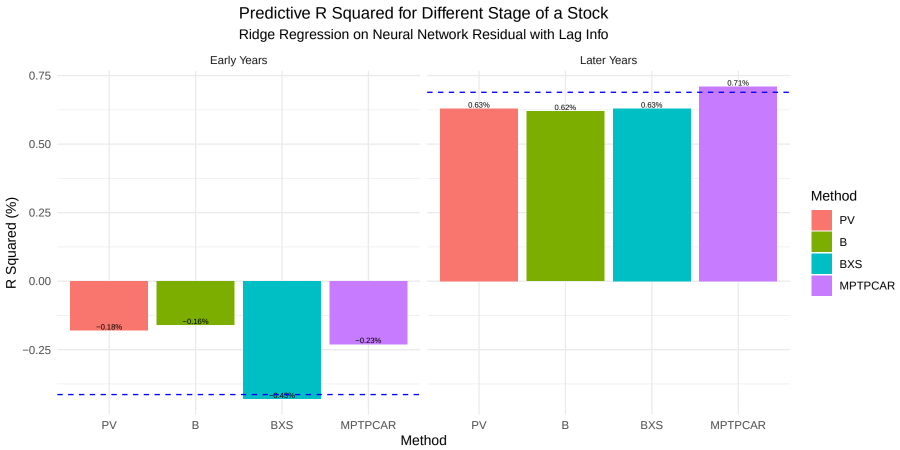
Missing Stratified Predicting Results (with Lag Info)



Stage Stratified Predicting Results (Without Lag Info)

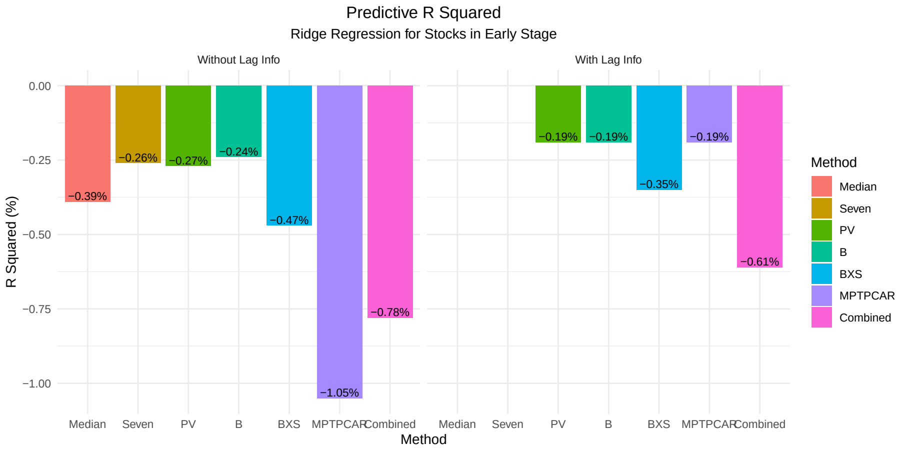


Stage Stratified Predicting Results (with Lag Info)



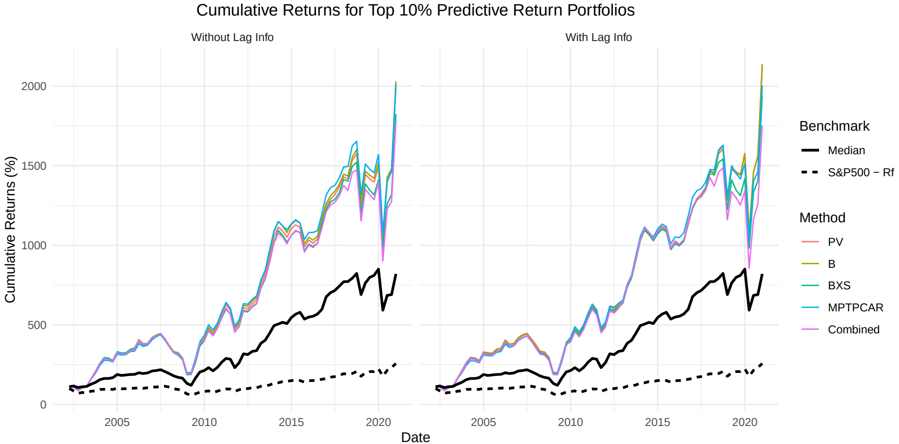
Early Result Only

Here, I only take out the stock returns during early years and try to fit a ridge regression:

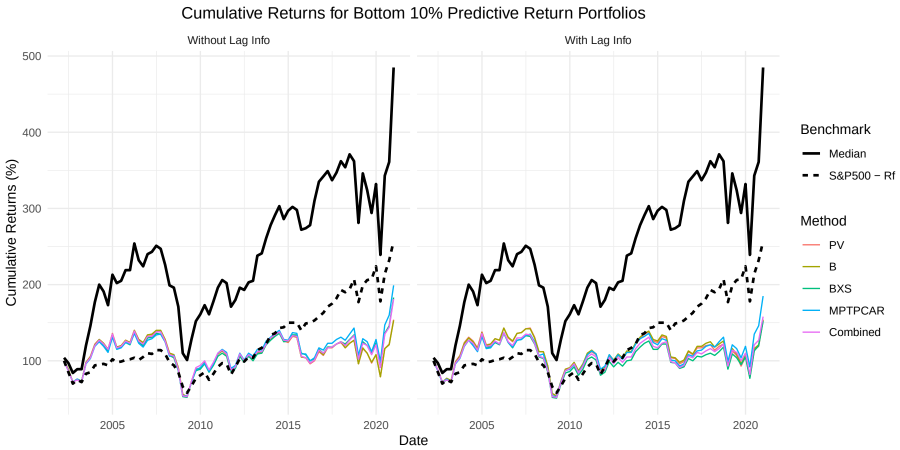


Top 10% Portfolio

Here, I **exclude early-year stocks** and apply different methods to construct **equal-weighted portfolios**. In each quarter, the stocks with predicted returns in the **top 10%** are selected to form a portfolio, which is then held for that quarter.



Bottom 10% Portfolios



Interpretation

There is a bias/variance tradeoff for predictors too. If the true model for return Y based on characteristics X is:

$$r = f(C) + \epsilon$$

where $E[\epsilon] = 0$ and $Var(\epsilon) = \sigma^2$. If we fit a model $\hat{f}(C)$, the expected squared prediction error at a point c_0 is:

$$\mathbb{E}\left[(Y - \hat{f}(c_0))^2\right] = \underbrace{\left(\mathbb{E}[\hat{f}(c_0)] - f(c_0)\right)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[\left(\hat{f}(c_0) - \mathbb{E}[\hat{f}(c_0)]\right)^2\right]}_{\text{Variance}} + \underbrace{\sigma^2}_{\text{Irreducible noi}}$$

The benefit of adding noisy imputation: lower bias (probably worth the extra variance).

Conclusion

MPTPCA method is useful in two senses:

1. The MPTPCA method provides **informative quarterly imputation**: The combination of MPTPCA and cross-sectional method can yield better out-of-sample imputation, especially during the early stage of a stock.
2. The MPTPCA method's imputation is **valuable in quarterly return prediction**: Using the Neural Network model residual, MPTPCA imputation with ridge regression can yield better prediction.
3. Up to this point in the project, none of the methods appear to address the **unpredictability of returns in the early stages**.

Next Steps:

Extend the base model to layer above 3 and work with monthly frequency since MPTPCA should be easily applied to monthly data.

It might be interesting to look into the timing when the firm characteristics starts to become informative for return prediction.

Thank you!

Babii, A., E. Ghysels, and J. Pan (2025). Tensor pca for factor models. Journal of Econometrics, 106077.

Bryzgalova, S., S. Lerner, M. Lettau, and M. Pelger (2024). Missing financial data. Review of Financial Studies.

Chen, A. Y. and J. McCoy (2024, May). Missing values handling for machine learning portfolios. Journal of Financial Economics 155, 103815.

Freyberger, J., B. Hoeppner, A. Neuhierl, and M. Weber (2024, 01). Missing data in asset pricing panels. The Review of Financial Studies.

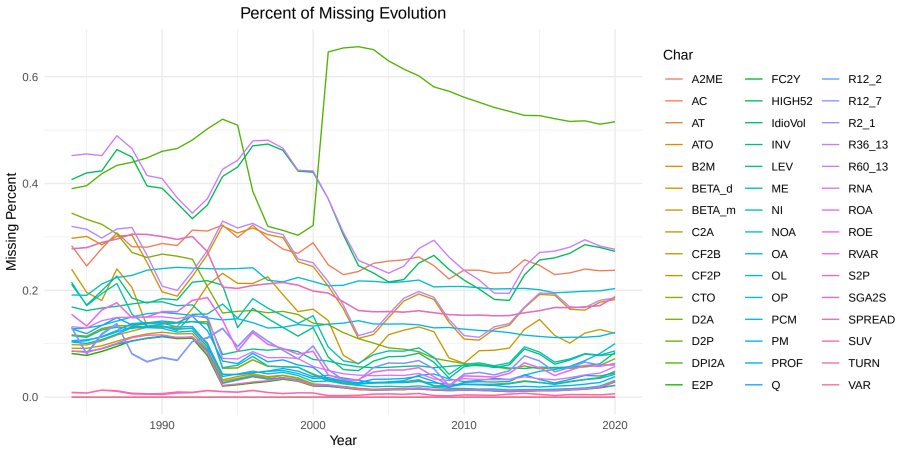
Gu, S., B. Kelly, and D. Xiu (2020, May). Empirical asset pricing via machine learning. The Review of Financial Studies 33(5), 2223–2273. Special Issue: New Methods in the Cross-Section.

Kelly, B., S. Malamud, and K. Zhou (2024). The virtue of complexity in return prediction. The Journal of Finance 79(1), 171–217.

Shen, Z. and D. Xiu (2025). Can machines learn weak signals? Technical report, Chicago Booth Research Paper No. 24-03. Available at SSRN: https://ssrn.com/abstract=4722678 or http://dx.doi.org/10.2139/ssrn.4722678.

Appendix

Missingness of Each Characteristic Across Stocks



Principal Component Analysis

The Principal Component Analysis (PCA) method estimates the two-dimensional factor model as follows:

$$C = FD\Lambda^\top + E$$

where:

- F is a $N \times R$ matrix of factors,
- Λ is an $L \times R$ matrix of loadings,
- D is a diagonal $R \times R$ matrix of eigenvalues,
- E is an $N \times L$ residual matrix.

Estimation steps:

- The columns of F are the principal components obtained from PCA on $CC^\top$ (unit vectors).
- The columns of Λ are the principal components from PCA on $C^\top C$ (unit vectors).
- The diagonal values of D are the square root of eigenvalues.

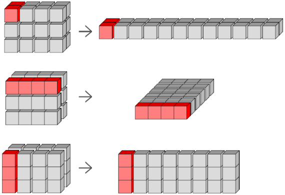
Tensor PCA

Following Babii et al. (2025)'s special case of CP decomposition, the data tensor is modeled as:

$$C = \sum_{r=1}^{R} \sigma_r M_r \otimes F_r \otimes \Lambda_r + U$$

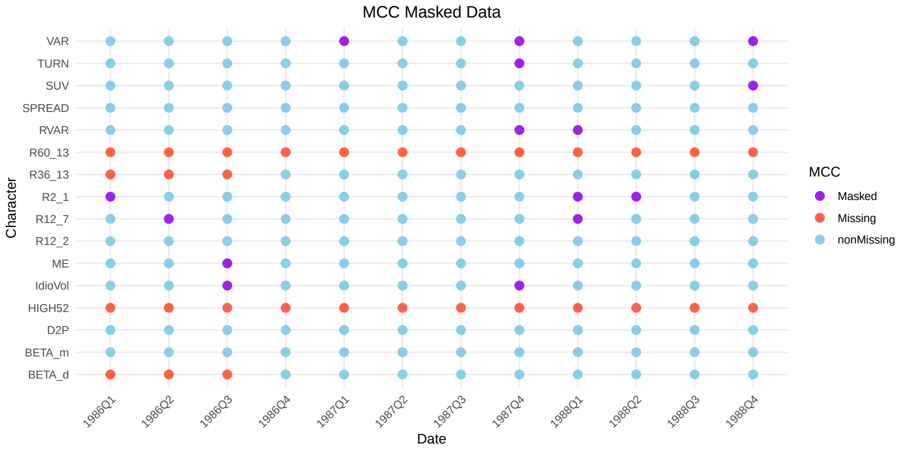
where C is a $T \times N \times L$ tensor, M, F, Λ contain time, firm and characteristics dimension information.

By unfolding the tensor along each dimension and performing PCA on the unfolded matrix, we can obtain the patterns along each dimension.

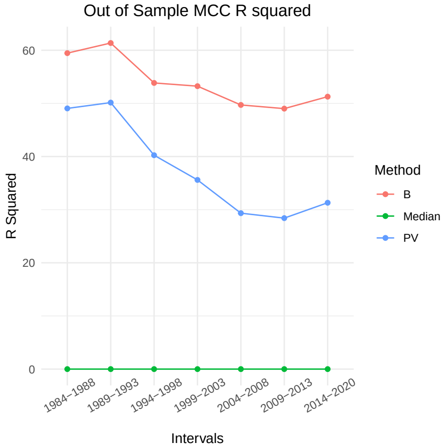


Masking Conditioning on Characteristics (MCC)

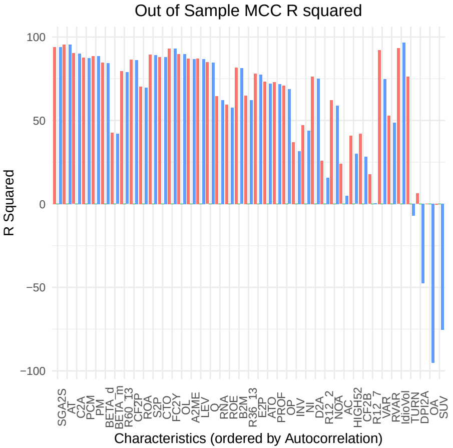
Masking Conditioning on Characteristics (MCC): randomly masking additional individual entries based on the assigned masking probability.



Median, Previous Value and Backward (MCC)

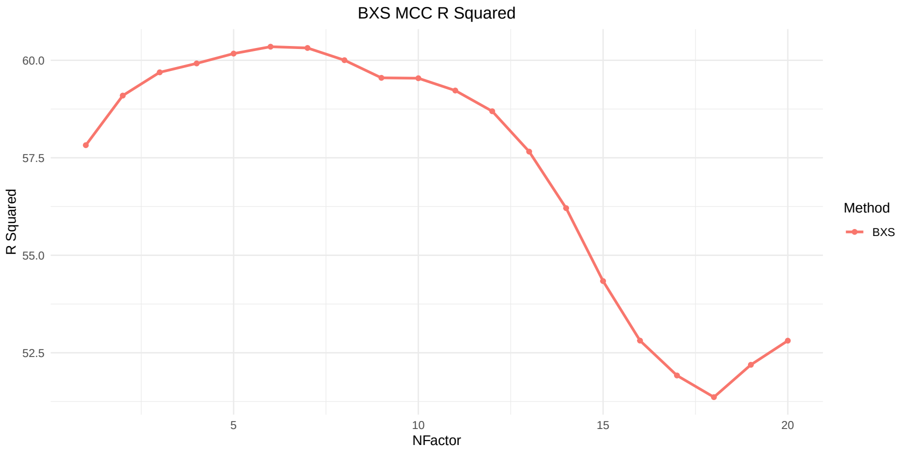


Out of Sample MCC R squared

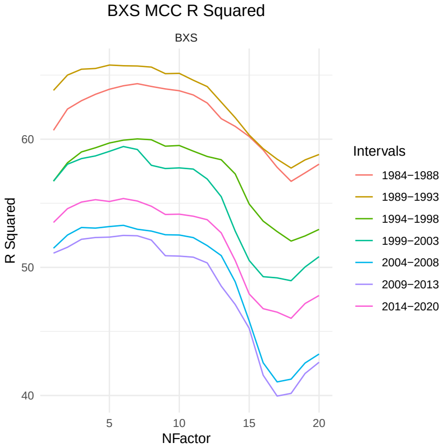


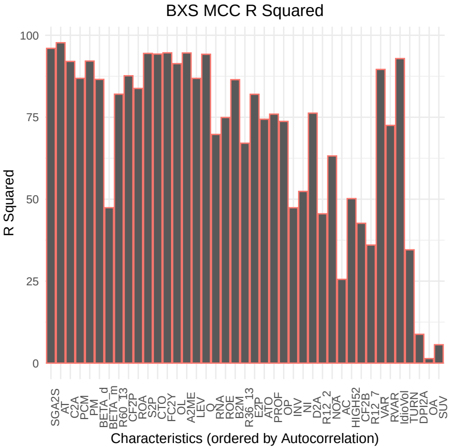
Out of Sample MCC R squared

Backward Cross-Sectional (MCC, Overall)

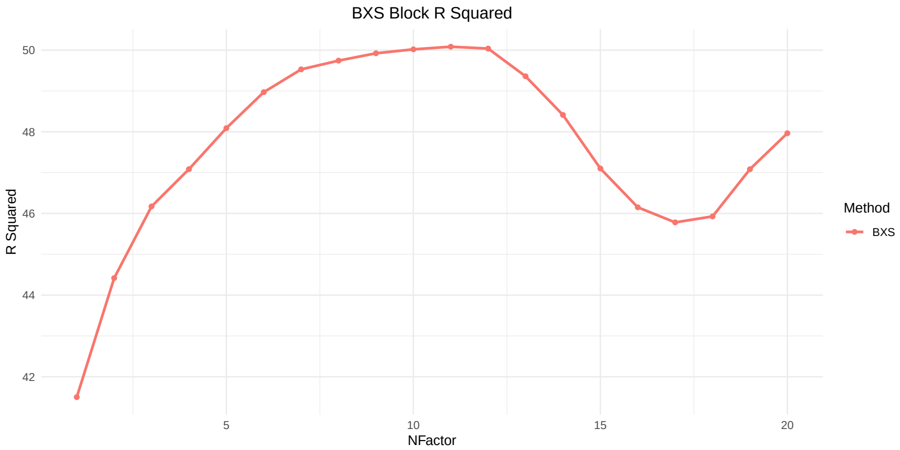


Backward Cross-Sectional (MCC, By Detailed)

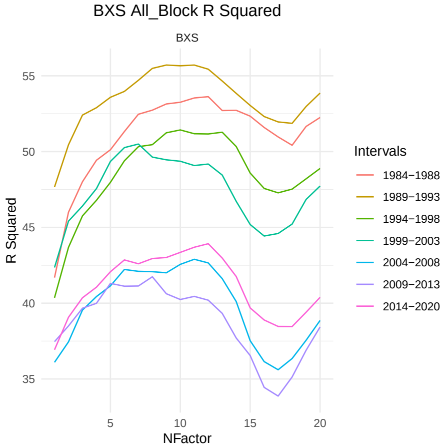


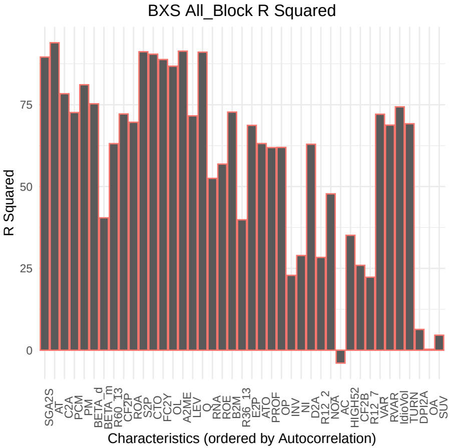


Backward Cross-Sectional (Block, Overall)

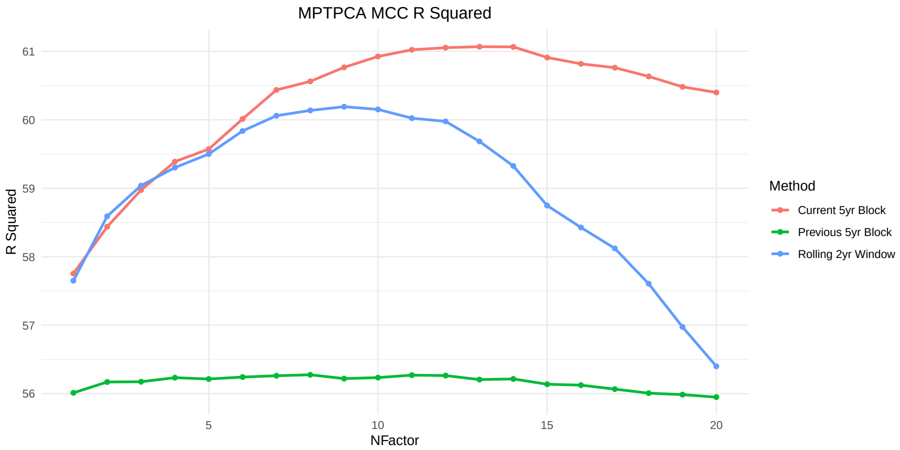


Backward Cross-Sectional (Block, Detailed)

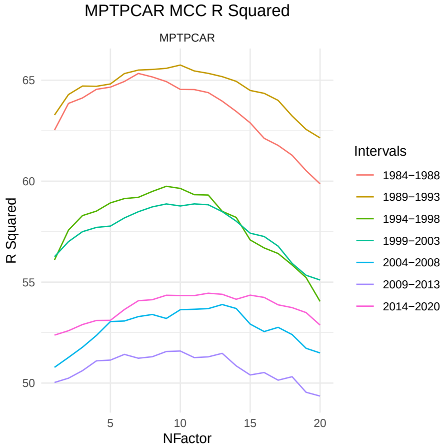


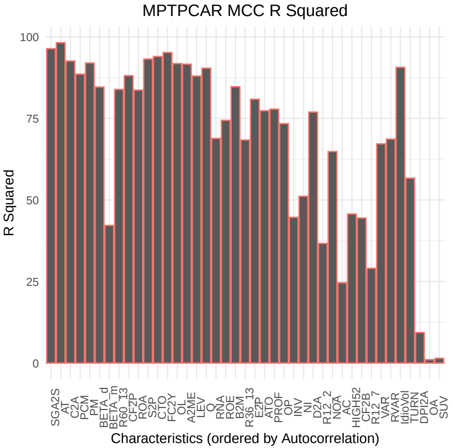


TPCA Models (MCC, Overall)

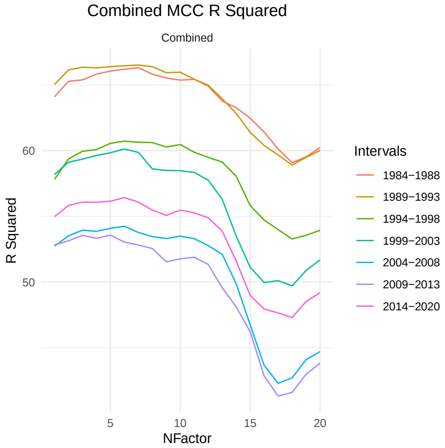


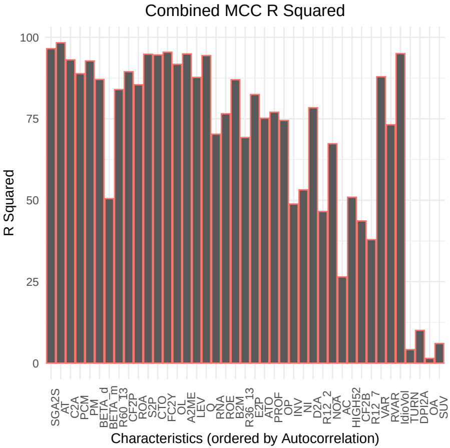
MPTPCAR (MCC, Detailed)



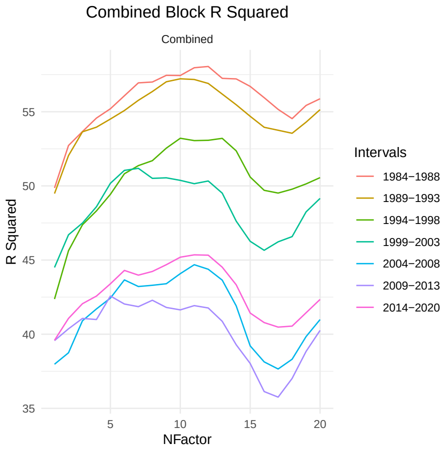


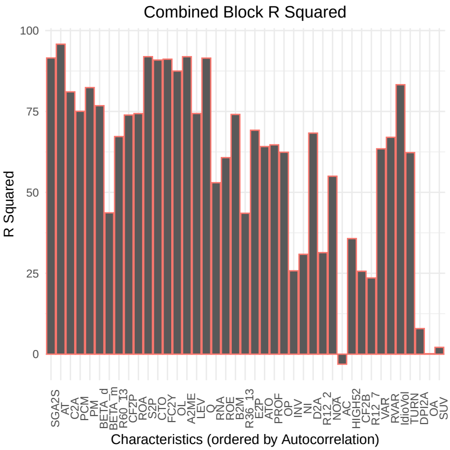
Combined (MCC, Detailed)



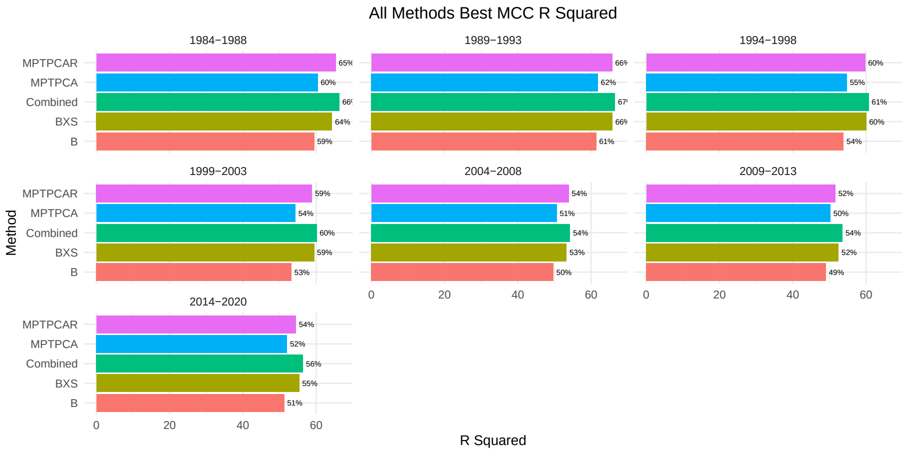


Combined (Block, Detailed)

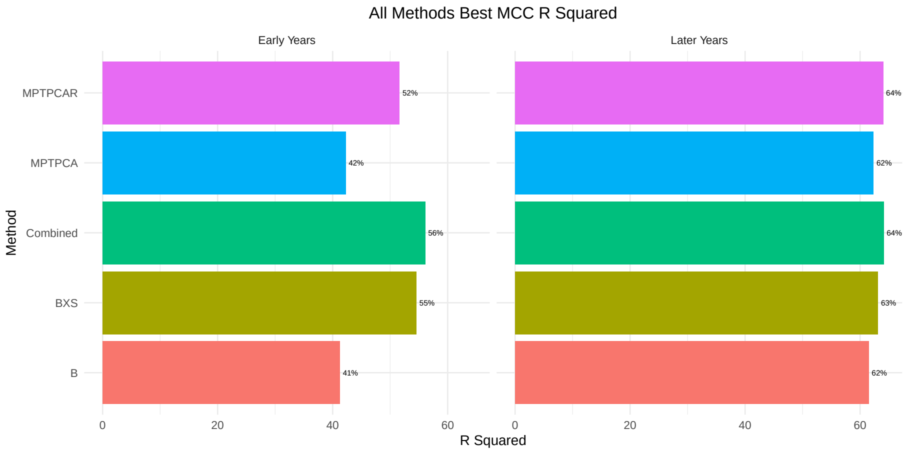

Combined Block R Squared


Combined Block R Squared

All Method Comparisons (MCC, By Time Intervals)



All Method Comparisons (MCC, By Time Early_Years)


All Methods Best MCC R Squared
Early Years
Later Years
Method
MPTPCAR
MPTPCA
Combined
BXS
B
R Squared