A Tensor Principal Component Analysis on Intraday Stock Returns

Jiaxi Li

Department of Economics

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

Format

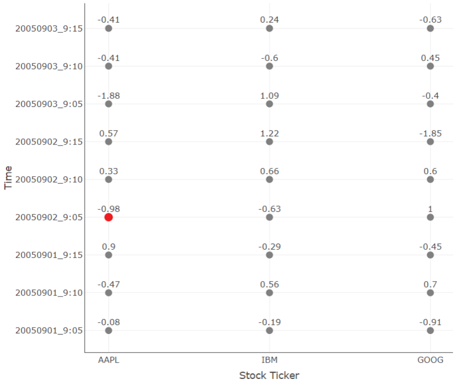
1. Introduction
2. Data
3. Model
4. Simulation
5. Empirical Results
6. Conclusion

Introduction

What is Tensor?

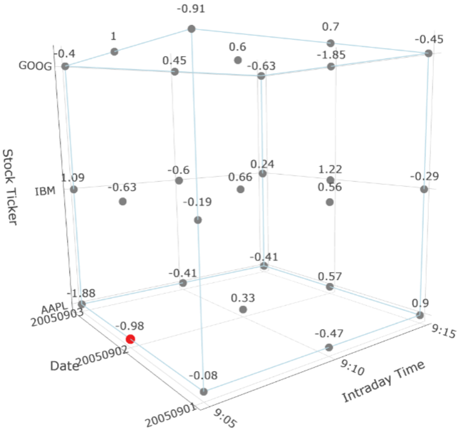
- A tensor is a multidimensional array.

A two-dimensional tensor is a panel dataset. One example is the following intraday stock return panel represented in a 2D plot:



3D Tensor

For the above panel data, we can regroup them into a 3D Tensor:



Factor Model for Panel Data

For a 2-Dimensional Factor Model,

$$y_{i,t} = \sum_{r=1}^{R} \beta_{i,r} f_{t,r} + u_{i,t}$$

where $f_{t,r}$ is a factor driving co-movement, $\beta_{i,r}$ is the factor exposure. $i = 1, \ldots, N$ is cross-section and $t = 1, \ldots, T$ is time.

In a matrix representation, $\mathcal{B}$ is a NxR matrix and F is a TxR matrix

$$Y = \sum_{r=1}^{R} \mathcal{B}_r \otimes F_r + U$$

where $\otimes$ is the tensor outer product.

2 Dimensional Factor Model Estimation

Some literature on 2D Factor Model estimation in Asset Pricing:

- Low Frequency Return:
  - Observable Factor: Ross (1976), Fama and French (1993), and later evolved into Factor Zoo (Chen and Zimmermann, 2022).
  - Latent Factor: Connor and Korajczyk (1986), Kelly et al. (2019).
  - Both Factors: Andreou et al. (2023).
- High Frequency Return:
  - Observable Factor: Aït-Sahalia, Jacod, and Xiu (2021).
  - Latent Factor: Aït-Sahalia and Xiu (2019).

Factor Model for 3D Tensor Data

For a 3 Dimensional Factor Model (with intraday as the extra dimension),

$$y_{i,j,t} = \sum_{r=1}^{R} \beta_{i,r}\gamma_{j,r}f_{t,r} + u_{i,j,t}$$

where $f_{t,r}$ is a factor driving co-movement, $\gamma_{j,r}$ is a common intraday pattern, $\beta_{i,r}$ is the factor exposure. $i = 1, \ldots, N$ is cross-section, $j = 1, \ldots, P$ is the intraday period and $t = 1, \ldots, T$ is date.

In a matrix representation, $\mathcal{B}$ is a NxR matrix, Γ is a PxR matrix and F is a TxR matrix

$$Y = \sum_{r=1}^{R} \mathcal{B}_r \otimes \Gamma_r \otimes F_r + U$$

where $\otimes$ is the tensor outer product. Referred to as Canonical Polyadic (CP) Decomposition, see Lettau (2023) for a factor model based on Tucker Decomposition.

3 Dimensional Factor Model

Why we can expect an intraday 3D factor model?

- Intraday Market Beta Variation: Andersen et al. (2021)
- Systematic Intraday Factor Beta Variation: Andersen et al. (2023)

Estimation of 3D Factor Model in Asset Pricing:

- CP Decomposition: Babii, Ghysels, and Pan (2022)
  - Tensor Principal Component Analysis (TPCA)
  - Alternating Least Square (ALS)[1]
- Tucker Decomposition: Lettau (2023)

[1] The result can be unstable.

Potential Problem with Dimension Aggregation

Dimension Aggregation: Aggregate one (or more) dimension of the tensor to make the tensor one (or more) dimension less.

Example: Aggregate intraday returns into daily returns.

$$y_{i,t} = \sum_{j=1}^{d_\gamma} y_{i,j,t}$$

Suppose d_γ is the length of dimension γ (Intraday Dimension).

Potential Problem: a loss of factors in the analysis.

Potential Problem with Dimension Aggregation

If the true model is 3D with

$$y_{i,j,t} = \sum_{r=1}^{R} \sigma_r \beta_{i,r} \gamma_{j,r} f_{t,r} + u_{i,j,t}$$

Suppose d_γ is the length of dimension γ. In an extreme case, for $\forall r$, $\sum_{j=1}^{d_\gamma} \gamma_{jr} = 0$, aggregating along dimension j would make the model:

$$\sum_{j=1}^{d_\gamma} y_{i,j,t} = \sum_{r=1}^{R} \sigma_r \beta_{i,r} (\sum_{j=1}^{d_\gamma} \gamma_{j,r}) f_{t,r} + \sum_{j=1}^{d_\gamma} u_{i,j,t}$$

$$\sum_{j=1}^{d_\gamma} y_{i,j,t} = \sum_{j=1}^{d_\gamma} u_{i,j,t}$$

We lose all factors in this dimension aggregation even if the signal strength σ_r is very large.

Issues with Intraday Tensors

Issues along the Intraday Dimension:

- Relatively small sample size
- Relatively large standard deviation of noise
- Intraday Seasonality (Heteroskedasticity)

The combination of these issues might cause an estimation bias in the TPCA model.

Potential Solution: Weighted TPCA.

Note: Heteroskedasticity is only an issue for small samples and a relatively low signal-noise ratio.

This Project

- Simulation:
  - TPCA and Alternating Least Square (ALS) on data under strong Heteroscedasticity with a small sample
  - Weighted TPCA to fix the issue of Heteroscedasticity
- Estimation:
  - Apply the TPCA method to estimate a Tensor factor model for Intraday Stock Returns
  - Using the weighted TPCA method to estimate the Tensor factor model for Intraday Stock Returns

Data

Intraday Stock Data

- Source: TAQ - Millisecond Consolidated Trades dataset at Wharton Research Data Service (WRDS).
- Dates: From 1/2/2009 to 7/17/2023.[2]
- Stocks: 78 permanent stocks of the frequently traded S&P 100 stocks.[3]
- Intraday Subsampling: 5-minute interval during the trading hours.[4]

The resulting data has dimension Date x Intraday x Stock (3315 x 78 x 78).

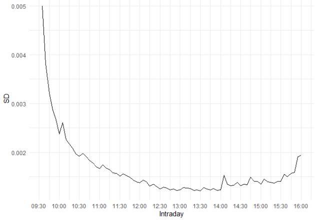
[2] 4/5/2012 data is also removed due to a dataset issue.

[3] Remaining in the list of S&P 500 during the whole sample period. The list was obtained on 7/9/2023 from Wikipedia. There are 82 stocks in the permanent list but 4 have some data issues with mergers.

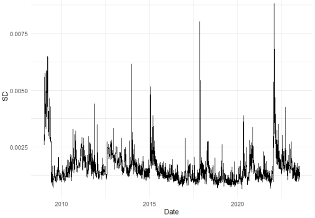
[4] The data cleaning procedure follows Barndorff-Nielsen et al. (2009).

Volatility and Intraday Seasonality

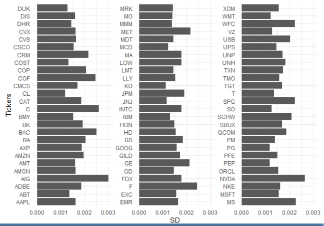
Each Intraday Standard Deviation



Each Date Standard Deviation



Each Stock Standard Deviation



Model

Principal Component Analysis (PCA)

With the Principal Component Analysis method (PCA),[5] we can estimate the 2 Dimensional factor model as:

$$Y = \mathcal{B}DF^T + U$$

where $\mathcal{B}$ is a NxR matrix and F is a TxR matrix and D is a diagnal RxR matrix.

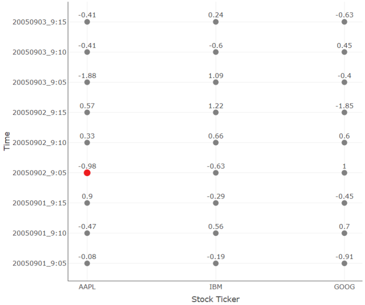
- The columns of $\mathcal{B}$ are the principal components from PCA on YY^T (unit vectors).
- The columns of F are the principal components from PCA on Y^TY (unit vectors).
- The diagonal values of D are the square root of eigenvalues (in descending order).

[5]One can also estimate the PCA with Singular Value Decomposition.

2D Model

$$y_{i,t} = \sum_{r=1}^{R} \sigma_r \beta_{i,r} f_{t,r} + u_{i,t}$$

where $f_{t,r}$ is a factor driving co-movement, $\beta_{i,r}$ is the factor exposure and σ_r represents the signal strength.



Tensor Principal Component Analysis (TPCA)

Based on Babii, Ghysels, and Pan (2022), if we generalize the above 2D model into a 3D tensor model:

$$Y = \sum_{r=1}^{R} \sigma_r \mathcal{B}_r \otimes \Gamma_r \otimes F_r + U$$

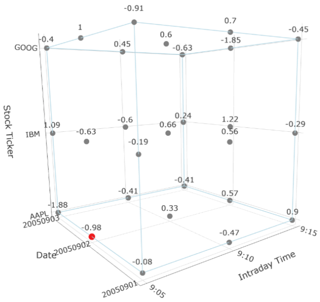
Under Orthogonal Assumption that $\mathcal{B}^T \mathcal{B} = \Gamma^T \Gamma = F^T F = I$, we can estimate the above model with the following algorithm for each dimension (I will explain with the Intraday dimension):

1. Unfold the Tensor into a Matrix along the Intraday Dimension
2. Estimate the factor loadings via PCA (eigenvectors form Γ)
3. Estimate the factor signal strength via PCA (eigenvalues are σ_r^2)

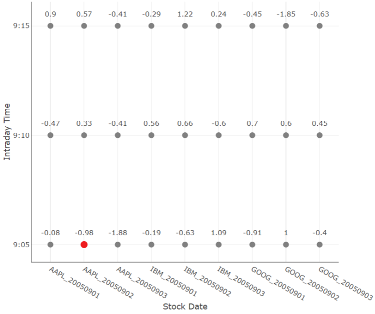
TPCA Tensor Matricization (Unfolding)

This would be the unfolding of Intraday Stock Return tensor along the Intraday Dimension (and only extract the pattern across this dimension):

(a) Tensor Data



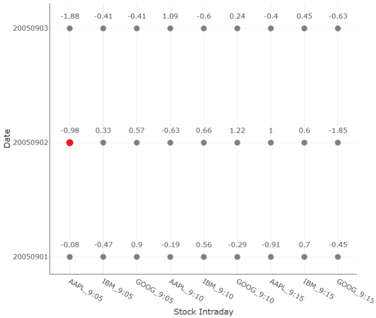
(b) Unfolding along Intraday Dimension for Γ estimation



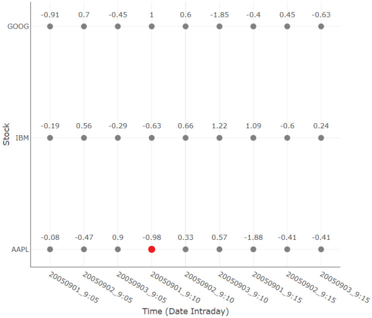
TPCA Tensor Matricization (Unfolding)

Here are the other two dimensions unfolding:

(a) Unfolding along Date Dimension for F estimation



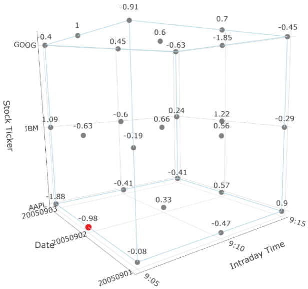
(b) Unfolding along Stock Dimension for $\mathcal{B}$ estimation



3D Model

$$y_{i,j,t} = \sum_{r=1}^{R} \sigma_r \beta_{i,r} \gamma_{j,r} f_{t,r} + u_{i,j,t}$$

where $f_{t,r}$ is a factor driving co-movement, $\gamma_{j,r}$ is the common intraday pattern, $\beta_{i,r}$ is the factor exposure and σ_r represents the signal strength.



Heteroskedasticity along Intraday Dimension

Babii, Ghysels, and Pan (2022) shows that under i.i.d. noise and orthoganol factors and loadings, TPCA can **consistently** identify the factors and loadings.

However, in **small sample**, when the heteroskedasticity is severe and noise is relatively large, we may not estimate the factors/loadings correctly in that direction.

For a 3-dimensional Factor Model:

$$y_{i,j,t} = \sum_{r=1}^{R} \sigma_r \beta_{i,r} \gamma_{j,r} f_{t,r} + w_j * u_{i,j,t} \tag{1}$$

where $f_{t,r}$ is a factor driving co-movement, $\gamma_{j,r}$ is a common intraday pattern, $\beta_{i,r}$ is the factor exposure. $i = 1, \ldots, N$ is cross-section, $j = 1, \ldots, P$ is the intraday period and $t = 1, \ldots, T$ is date. w_j is the intraday seasonal weights, where $\frac{1}{d_\gamma} \sum_{j=1}^{d_\gamma} w_j^2 = 1$.[6]

[6] This is to make sure $Var(u_{i,j,t}) = Var(w_j * u_{i,j,t})$.

Weighted TPCA

A weighted version of TPCA seems to be a natural candidate. Divide everything in Eq 1 by w_j:

$$(y_{i,j,t}/w_j) = \sum_{r=1}^{R} \sigma_r \beta_{i,r} (\gamma_{j,r}/w_j) f_{t,r} + u_{i,j,t}$$

Let $\tilde{y}_{i,j,t} = y_{i,j,t}/w_j$, $\gamma_{j,r} = \frac{w_j \tilde{\gamma}_{j,r}}{\sum_{j=1}^{d_\gamma} w_j^2 \tilde{\gamma}_{j,r}^2}$, $\sigma_r = (\sum_{j=1}^{d_\gamma} w_j^2 \tilde{\gamma}_{j,r}^2) * \tilde{\sigma}_r$:

$$\tilde{y}_{i,j,t} = \sum_{r=1}^{R} \tilde{\sigma}_r \beta_{i,r} \tilde{\gamma}_{j,r} f_{t,r} + u_{i,j,t}$$

$$\sum_{j=1}^{d_\gamma} w_j^2 \tilde{\gamma}_{j,r}^2 = \sum_{j=1}^{d_\gamma} (\sum_{j=1}^{d_\gamma} w_j^2 \tilde{\gamma}_{j,r}^2)^2 \gamma_{j,r}^2 = (\sum_{j=1}^{d_\gamma} w_j^2 \tilde{\gamma}_{j,r}^2)^2 \sum_{j=1}^{d_\gamma} \gamma_{j,r}^2 = (\sum_{j=1}^{d_\gamma} w_j^2 \tilde{\gamma}_{j,r}^2)^2$$

$$\Rightarrow \sum_{j=1}^{d_\gamma} w_j^2 \tilde{\gamma}_{j,r}^2 = 1, \ \gamma_{j,r} = w_j \tilde{\gamma}_{j,r}, \ \sigma_r = \tilde{\sigma}_r$$

Weighted TPCA

Weighted data $\tilde{y}_{i,j,t}$ has i.i.d. noise and the TPCA result on $\tilde{y}_{i,j,t}$ preserves signal strength σ_r of the original model:

$$\tilde{y}_{i,j,t} = \sum_{r=1}^{R} \sigma_r \beta_{i,r} \tilde{\gamma}_{j,r} f_{t,r} + u_{i,j,t}$$

where $\gamma_{j,r} = w_j \tilde{\gamma}_{j,r}$

All the results from Babii, Ghysels, and Pan (2022) can apply to the weighted version of TPCA.
TPCA result on a heteroskedastic y would still be consistent but the weighted TPCA can achieve better efficiency.

Simulation

Heteroskedastic Model

The Data Generating Process (DGP) for the Heteroskedastic Model is as follows:

$$y_{i,j,t} = \sum_{r=1}^{3} \sigma_r \beta_{i,r} \gamma_{j,r} f_{t,r} + w_j * u_{i,j,t}$$

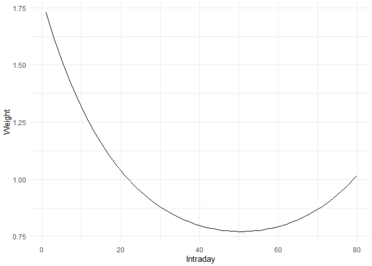
Where the dimension of y is 2000 x 80 x 80. Each $\beta_{i,r}$, $\gamma_{j,r}$, $f_{t,r}$ is simulated using a standard normal distribution and then normalized to 1. $\sigma_1 = 3, \sigma_2 = 2, \sigma_3 = 1$. $u \sim N(0, 0.04)$. $w_j = \frac{e^{101-j}+e^j+0.5}{\sqrt{\frac{1}{80}*\sum_{j=1}^{80}(e^{101-j}+e^j+0.5)^2}}$.

The focus would be γ_j since all the other dimensions can be identified well.

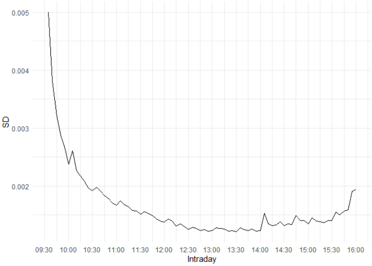
Weights

The weighting function is set to mimic the standard deviation of each intraday period:

(a) Weights in simulation



(b) Intraday Seasonality



TPCA Result

The result of the TPCA has an Intraday pattern that usually focuses on certain intraday period:

TPCA Result in Eigenvectors

Eigenvector representations of the TPCA result of the intraday dimension would be:

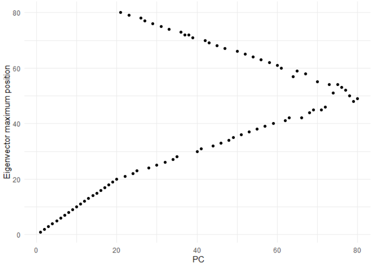
$$
\begin{array}{ccc}
\gamma_1 & \gamma_2 & \gamma_3 \\
PC1 & PC2 & PC3 \\
\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} &
\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} &
\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}
\end{array}
$$

The maximum value positions in each vector: 1, 2, 3 ...
The maximum value in each vector: 1, 1, 1 ...

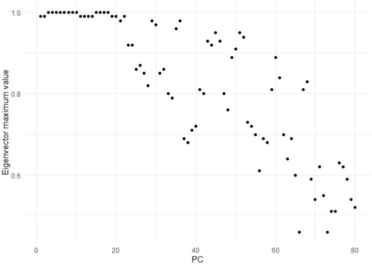
TPCA Intraday

Here are the plots for the position where the maximum in eigenvector occurs and the value of the maximum eigenvector:

(a) Position of eigenvector maximum

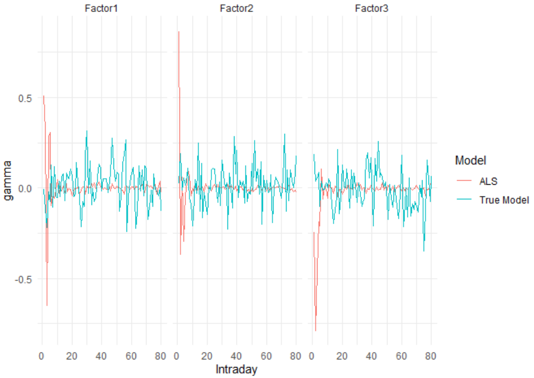


(b) Value of eigenvector maximum



ALS Result

If we set the number of factor to be three (true number), the result of ALS also has an Intraday pattern that usually focuses on a certain intraday period:



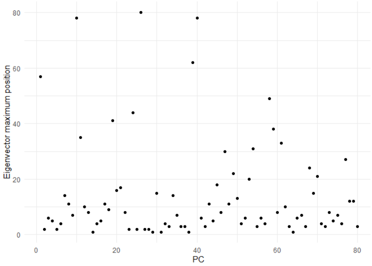
Weighted TPCA Result

The result of weighted TPCA aligns well with the true model:

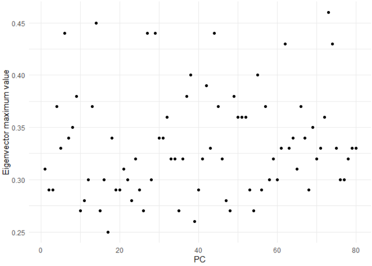
Weighted TPCA Intraday

Here are the plots for the position where the maximum in eigenvector occurs and the value of the maximum eigenvector:

(a) Position of eigenvector maximum



(b) Value of eigenvector maximum



Estimation Error Comparison

Here is a table of the Mean Squared Error (MSE) of the Intraday Pattern Estimates from the three models:

	MSE for Γ				
	Homoskedastic		**Heteroskedastic**		
Factors	**TPCA**	**ALS**	**TPCA**	**ALS**	**WTPCA**
Γ_1	0.0087	0.0247	0.0244	0.0236	0.0071
Γ_2	0.0180	0.0226	0.0204	0.0234	0.0179
Γ_3	0.0246	0.0247	0.0236	0.0224	0.0235

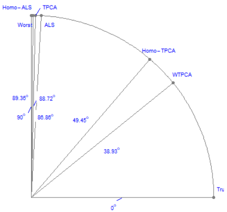
Note: Since both the estimates and the true values are normal vectors, the maximum MSE would be $\frac{2}{N} = 2/80 = 0.025$. I would show a plot of angles between the estimated and true normal vectors to illustrate the goodness of fit.

Quarter Circle Plots

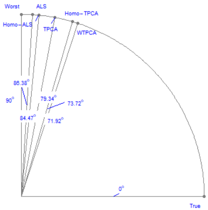
Below are the quarter-circle plots indicating the goodness of fit of normal vectors:

- 0^o indicates a perfect fit
- 90^o indicates a worst fit (orthogonal)

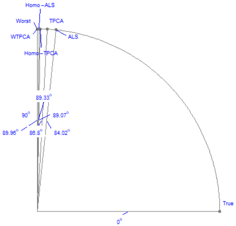
(a) First Factor



(b) Second Factor

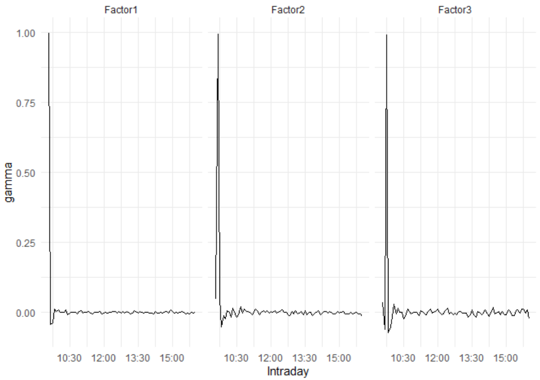


(c) Third Factor



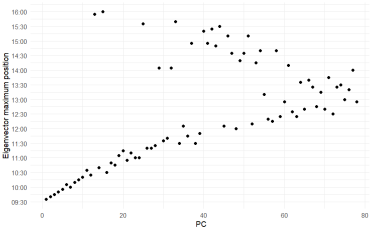
Empirical Results

The result of the unconditional Factor Model has an Intraday pattern that usually focuses on certain intraday period returns, especially at the beginning of trading hours.
Here is a sample for the first factor's intraday pattern:

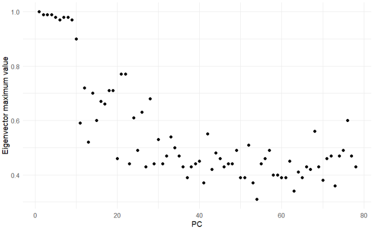


Here are the plots for the position where the maximum in eigenvector occurs and the value of the maximum eigenvector:

(a) Position of eigenvector maximum



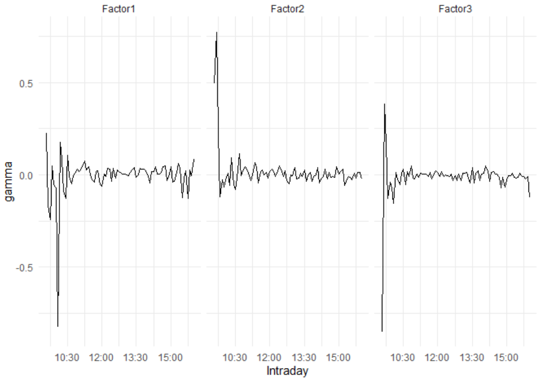
(b) Value of eigenvector maximum



The first few eigenvectors (factors' intraday patterns) almost solely focus on one of the intraday periods. This seems to align with the simulation result with TPCA on Heteroskedastic data.

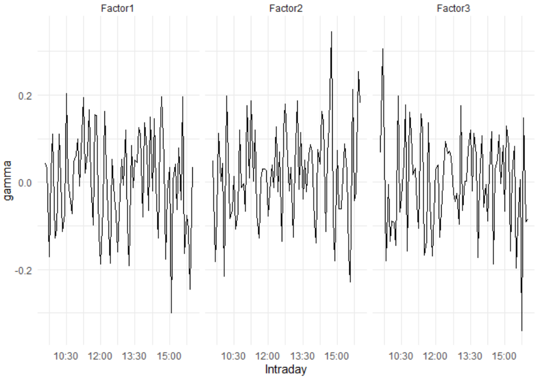
ALS Intraday

Here are the plots for the first two factor's intraday component from ALS Estimation while setting the number of factor as Three:

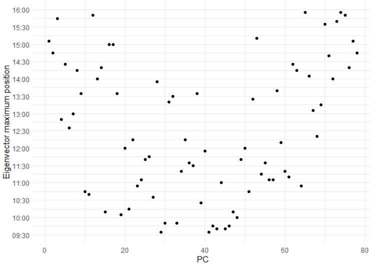


Weighted TPCA Intraday

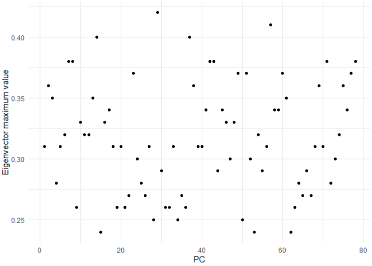
I used the scaled unconditional standard deviation of each intraday period as the weights for the Weighted TPCA. The result no longer focuses on single dimensions.



Here are the plots for the position where the maximum in eigenvector occurs and the value of the maximum eigenvector:



(a) Position of eigenvector maximum



(b) Value of eigenvector maximum

Conclusion

Conclusion and extension

Three main Conclusions:

1. Small sample with large heteroskedasticity might result in estimation bias in TPCA.
2. Intraday Stock return Tensor might suffer this issue.
3. Weighted version of TPCA can resolve this issue. It can preserve the signal strength and therefore preserve all the result from Babii, Ghysels, and Pan (2022).

Potential next steps:

1. Check the number of factors in the resulting factor model.
2. Check whether the model is overfitting.
3. Compare the factors generated with the existing factors in the factor zoo.

Thank you!

Citation I

Andersen, T. G., R. Riva, M. Thyrsgaard, and V. Todorov (2023). Intraday cross-sectional distributions of systematic risk. Journal of Econometrics 235(2), 1394–1418.

Andersen, T. G., M. Thyrsgaard, and V. Todorov (2021). Recalcitrant betas: Intraday variation in the cross-sectional dispersion of systematic risk. Quantitative Economics 12(2), 647–682.

Andreou, E., P. Gagliardini, E. Ghysels, and M. Rubin (2023, February). Spanning latent and observable factors. Available at SSRN: https://ssrn.com/abstract=4349003 or http://dx.doi.org/10.2139/ssrn.4349003.

Aït-Sahalia, Y., J. Jacod, and D. Xiu (2021). Inference on risk premia in continuous-time asset pricing models. Working paper.

Aït-Sahalia, Y. and D. Xiu (2019). Principal component analysis of high-frequency data. Journal of the American Statistical Association 114(525), 287–303. Theory and Methods.

Citation II

Babii, A., E. Ghysels, and J. Pan (2022). Tensor principal component analysis. arXiv preprint arXiv:2212.12981.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2009). Realized kernels in practice: Trades and quotes. Econometrics Journal 12, C1–C32.

Chen, A. Y. and T. Zimmermann (2022). Open source cross-sectional asset pricing. Critical Finance Review 11(2), 207–264.

Connor, G. and R. A. Korajczyk (1986, March). Performance measurement with the arbitrage pricing theory: A new framework for analysis. Journal of Financial Economics 15(3), 373–394.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3–56.

Kelly, B. T., S. Pruitt, and Y. Su (2019, December). Characteristics are covariances: A unified model of risk and return. Journal of Financial Economics 134(3), 501–524.

Citation III

Lettau, M. (2023, September). High-dimensional factor models and the factor zoo. Working Paper 31719, National Bureau of Economic Research.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. Journal of Economic Theory 13(3), 341–360.