THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

A Weighted Tensor Principal Component Analysis on Intraday Stock Returns

By

Jiaxi Li

December 15th, 2023

A paper submitted in fulfillment of the Field Paper requirements
for the Economics Ph.D.

Faculty Advisor: Eric Ghysels
Co-advisor: Andrii Babii
Field Paper Committee: Eric Ghysels, Andrii Babii, Peter Hansen

Abstract

Substantial research has delved into extracting the factor structure of the U.S. stock market, and many papers utilize the intraday returns since their availability. Employing the recently developed Tensor Principal Component Analysis method (TPCA introduced by Babii et al. in 2022), this paper tries to derive the underlying factors of intraday five-minute returns from the permanent constituents of the S&P 100. However, the data seem to suffer a small sample, strong intraday heteroskedasticity, and large error. In simulation, such problems can lead to insensible TPCA results. I propose a weighted version of TPCA that can potentially mitigate the issue and preserve the asymptotic properties and test implications of the TPCA.

1 Introduction

After Ross (1976) laying the foundation of the multifactor model in Asset Pricing, significant endeavors have been directed toward uncovering the underlying factor structure of asset returns. Since then, researchers have used monthly and daily stock or portfolio returns in their analysis. There are generally two types of factors: Observable Factors and Latent Factors. Due to the data availability, early work has been focusing on the monthly and daily stock or portfolio returns.

To produce the observable factors, researchers use finance theory or stylized facts to make an argument that a particular firm characteristic would help explain or predict asset returns. They then use these characteristics to form the observable factors and test their validity. Fama and French (1993) observe that small-cap and high book-to-market-ratio stocks tend to outperform the market and construct Small-Minus-Big (SMB) and High-Minus-Low (HML) factors. Along with the market factor, the so-called Fama French three factors can better explain the cross-sectional returns. This has led to the development of a complex array of observable factors, often referred to as the "factor zoo". In Chen and

Zimmermann (2022), they document 319 observable factors and verify their validity. This vast body of factors prompts discussions about the need to trim this zoo. Since the observable factors are built upon observable characteristics, they have clear economic interpretations.

Generating the so-called latent factors often involves some versions of the statistical method called Principal Component Analysis (PCA). PCA would use the covariance matrix to extract the principal components that explain the most variations in the data. In the factor literature, these principal components from different dimensions will inform us of the factors and corresponding loadings. Connor and Korajczyk (1986) first propose using the PCA to generate the latent factors and derive the asymptotic properties of the estimates. More recently, Kelly et al. (2019) use an instrumented PCA to uncover the time-varying factor dynamics. Andreou et al. (2022) use both individual stock and portfolio monthly returns to generate three factors that have a better out-of-sample performance. In General, the latent factors usually perform better statistically than the observable factors, but their economic meanings can be hard to interpret.

There are also recent literature working on testing the observable and the latent factor concurrently. Andreou et al. (2023) test whether the latent factors and the observable factors generate the same space. If they generate the same space, we can have factors with both good statistical properties and better economic explanations.

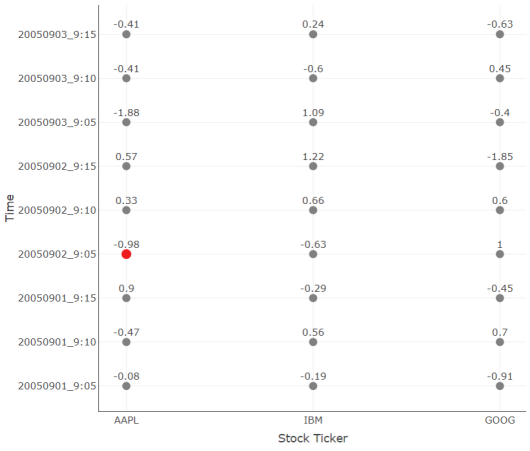
At the same time, technological advancements have enabled day trading and the availability of intraday data, potentially offering deeper insights into factor exploration. Intraday data can also help to generate the observable and latent factors. As an obvious extension of the previous literature, some researchers use the panel data of the stock intraday returns where we have time and stock as the two dimensions (as presented in Figure 1 Panel (a)). For example, Aït-Sahalia et al. (2021) use intraday returns to generate both continuous and

the jump factors of the Fama French Five factors, and test their validity in a continuous-time asset pricing model. Aït-Sahalia and Xiu (2019) use PCA to generate a time-varying (latent) factor model using one-week-rolling-window intraday S&P 100 stock returns.

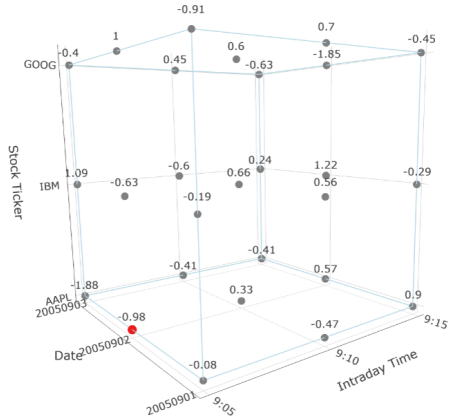
Figure 1: Two-dimensional Panel and Three-dimensional Tensor

Panel (a) presents a made-up sample of stock intraday return data in a panel, where the time dimension includes both date and intraday period. Panel (b) presents the same stock intraday return data in a three-dimensional tensor. As an example, the red point in both data represents that on September 2nd, 2005, AAPL had a -0.58 five-minute return at 9:05.

(a) Panel Data



(b) Tensor Data



However, we can also transform the stock intraday returns into a three-dimensional tensor (as Figure 1 Panel (b)) by introducing an intraday time dimension, besides traditional dimensions of date and stock. Essentially, the tensor form would contain the same information as the panel form. However, it can help us extract the intraday variation of stock returns in addition to the cross-sectional variation and time-series variation.

One may wonder why we can expect an intraday variation in returns. Andersen et al. (2021) show that there is a systematic intraday market beta variation. Andersen et al. (2023) provide further evidence that size and value factors in the Fama French Five factors

also exhibit significant systematic variation. If such intraday variations of the factor loading occur, we can have a model with consistent intraday variation.

Regarding the decomposition of tensor data, one way is called Canonical Polyadic (CP) Decomposition. The workhorse of its estimation is the Alternating Least Square (ALS), as noted in Kolda and Bader (2009). However, it relies on the numerical method to minimize squared errors which can result in numerical instability.[1] Another promising avenue for CP decomposition is the Tensor Principal Component Analysis (TPCA), as proposed by Babii et al. (2022). This method involves the unfolding of tensor data followed by PCA on the unfolded structure, enabling the extraction of factors from intraday data. Their paper also provides an example of TPCA on the daily sorted portfolio return tensor.

The other tensor decomposition method is called Tucker Decomposition. Lettau (2023) proposes a method to estimate such decomposition and applies the method to intraday return tensor as an example. However, this method also involves numerical approximation in the estimation. Tucker decomposition result is also harder to interpret than the CP decomposition.

In this study, I will apply the TPCA method to analyze the intraday returns of the permanent members of the S&P 100 index at a five-minute interval, aiming to reveal and comprehend their underlying factor structure, including time-series variation, cross-sectional variation, and intraday variation.

Unfortunately, the obtained intraday stock return tensor suffers from a combination of a small sample, large errors, and severe heteroskedasticity along the intraday dimension. I will further show in a simulation that this combination can cause a severe estimation

[1]The fit can be different if we start from different initial values, trapping us in a local minimum instead of the global minimum.

problem.[2] I then propose a weighted version of TPCA to mitigate estimation problems in theory and simulation. This weighted TPCA (WTPCA) can nicely reduce the problem in the small sample, produce consistent results asymptotically, and preserve all the testability of the TPCA. Furthermore, I will analyze the actual intraday returns and provide some evidence that the TPCA estimation seems to suffer the same problem and that WTPCA works better. This suggests that when conducting TPCA with tensor data, weighting by estimated unconditional (or conditional) volatility is desired.

The remainder of this paper is structured as follows: Section 2 (Data) presents the dataset and outlines its key characteristics. Section 3 (Models) compares the two-dimensional and three-dimensional factor models for intraday stock data and delves into different methods to reveal the latent factor structure. Section 4 (Simulation) applies and compares different estimation methods on simulated data with a combination of a small sample, large errors, and severe heteroskedasticity along the intraday dimension. The outcomes of the empirical analysis are depicted in Section 5 (Empirical Results). Finally, Section 6 (Conclusion) concludes the paper and suggests some future direction for the research with TPCA or WTPCA.

2 Data

This section outlines the process of data cleaning and the transformation of the data into a tensor form. I will present the summary statistics of five-minute intraday returns and exhibit evidence hinting at the presence of small sample sizes, significant errors, and notable heteroskedasticity issues within the intraday dimension.

[2]Note: Heteroskedasticity is only an issue with small samples and low signal-to-noise ratio. Asymptotically, we would have a consistent estimation of the model.

2.1 Intraday Stock Returns

The high-frequency stock price data is from the TAQ - Millisecond Consolidated Trades dataset at Wharton Research Data Service (WRDS). I extract the price of permanent members of the frequently traded S&P 100 stocks spanning from 1/2/2009 to 7/17/2023.[3] Permanent members refer to stocks consistently listed in the S&P 500 throughout the entire sample period, comprising large-cap stocks exempt from liquidity issues. This selection avoids companies that were delisted or went bankrupt, introducing unavoidable survivorship bias. However, with 82 stocks persisting in the S&P 500 over the entire period, the bias appears manageable.[4][5] Due to mergers and acquisitions, accurate prices are available for only 78 out of the 82 permanent S&P 100 stocks, as detailed in Appendix I.

The data cleaning process adheres to the procedure outlined by Barndorff-Nielsen et al. (2009) and utilizes the High-frequency package from Boudt et al. (2022). Consistent with the approach in Aït-Sahalia and Xiu (2019), only the price information from the exchange with the highest transaction records is employed. Additionally, stock information with additional suffixes is eliminated to prevent the inclusion of preferred stock prices in the sample. To address balanced panel concerns and mitigate issues related to stock splits and dividend issuance, the data is exclusively selected from trading hours (9:30 to 16:00 EST), and prices are sampled at five-minute intervals.

After extracting prices from TAQ, I compute five-minute log returns to mitigate microstructure noise while ensuring a relatively large sample of intraday returns. Each full

[3] The list is obtained on 7/9/2023 from Wikipedia. 4/5/2012 data is also removed due to a dataset issue.

[4] An alternative sampling approach could involve obtaining intraday returns for all S&P 100 stocks each quarter and analyzing with TPCA (similar to Aït-Sahalia and Xiu (2019), though it is beyond the paper's scope.

[5] Another consideration is the permanent members of S&P 500 stocks, offering a larger cross-section but with only 286 permanent ones, implying a more pronounced survivorship bias.

trading day encompasses 78 five-minute intervals. Considering the large-cap nature of these stocks, a high number of missing returns may indicate unusual market conditions (such as half-day or circuit breakers). Days with more than three missing returns for all stocks are therefore excluded and empty entries are assigned a value of zero. The resulting dataset consists of 3,315 days, each comprising 78 intraday periods and 78 stocks. However, the original data is structured in panel form, yielding a total of 258,570 periods (3315×78) with 78 stock returns for each period. An illustrative example is provided in Figure 1 Panel (a). Many studies, including Aït-Sahalia and Xiu (2019), analyze intraday stock returns in this manner.

2.1.1 Summary Statistics of Intraday Panel

Upon obtaining the returns, I compute the unconditional moments of each stock's five-minute returns.[6] The means and medians of the five-minute returns are consistently zero. However, the relatively large standard deviations suggest a notable degree of variability or error. The significant kurtosis could stem from the dynamic and time-varying nature of volatility. Examining the lag-one autocorrelations (five minutes) reveals the mean-reverting behavior in intraday stock returns. Figure 20 (available in Appendix II) illustrates positive correlations among the selected stocks, indicating a tendency for all stocks to have a positive loading on the major systematic risk, typically associated with market risk.

2.1.2 Tensor Data formation

To construct a Tensor, the approach involves configuring the dimensions: Date and Intraday (intraday five-minute intervals) constitute the first two dimensions, while the third dimen-

[6]The tables in Appendix II present the unconditional moments for each stock.

sion, Stock, corresponds to individual stock tickers. This results in a Tensor with dimensions of Date $\times$ Intraday $\times$ Stock ($3315 \times 78 \times 78$). Both the stock and intraday dimensions have relatively small sample sizes.

An example of a three-dimensional tensor would be panel (b) of Figure 1. This three-dimensional tensor contains the same information as the two-dimensional panel data in panel (a). However, the two-dimensional representation would combine the date and intraday dimensions into one single time dimension. An analysis of this Panel data would overlook the intraday variation of stock returns.

Figure 2: Volatility Conditioning on Intraday Periods

The standard deviation of five-minute returns for each intraday period (not condition on stock or date). There is a substantial difference in the volatility levels. High volatility occurs at the beginning (and end) of trading days.

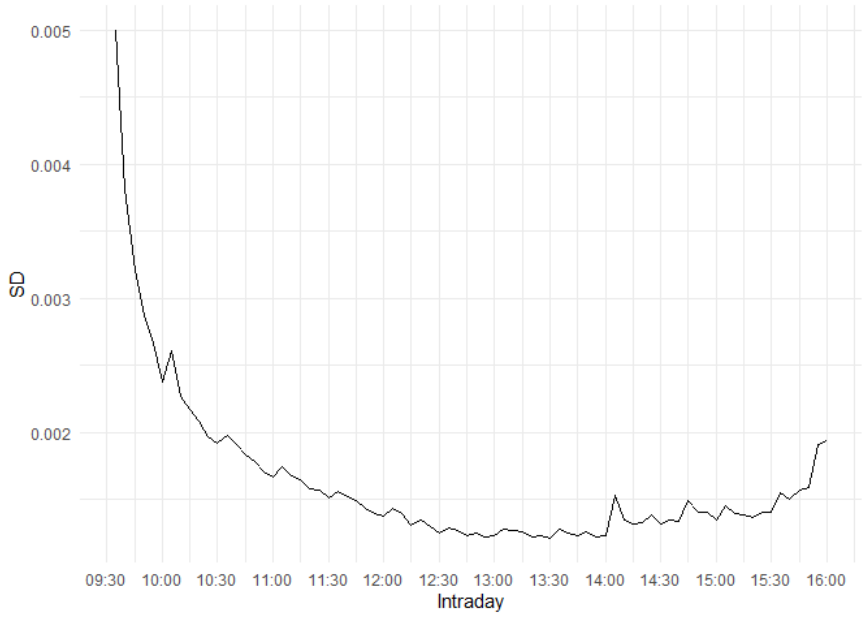
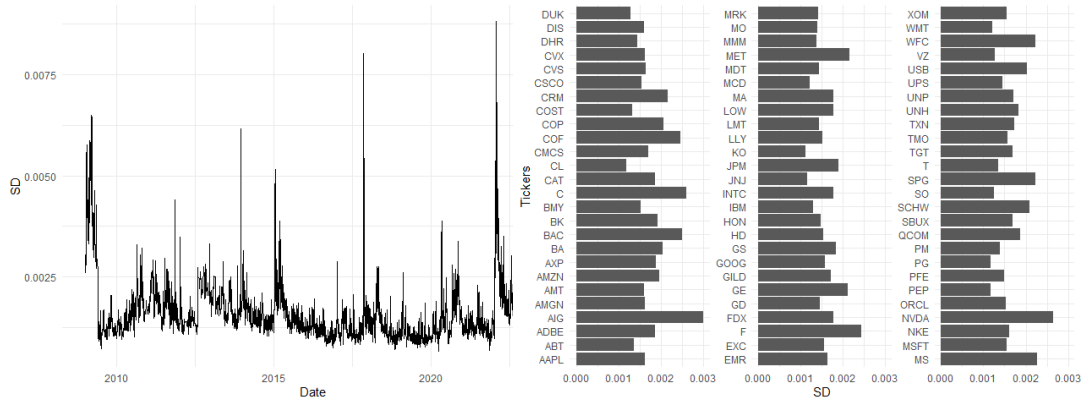


Figure 3: Volatility Conditioning on Date/Stock Dimension

Panel (a) presents the volatility conditioning on dates. Panel (b) presents the volatility conditioning on stocks. We do not have a large dispersion for stocks but for dates. However, we have many dates so this dispersion would not cause an estimation error.

(a) Volatility Conditioning on Dates (b) Volatility Conditioning on Stocks



2.1.3 Volatility of Returns Conditioning on Each dimension

I further calculate the volatility of five-minute returns for each stock, on each date, and during each intraday period. This involves estimating the volatility of returns conditioned on each dimension of the three-dimensional tensor. Figure 2 illustrates return volatility conditioned on each intraday period. Notably, there is a significant dispersion of volatility across different intraday periods, with elevated volatility observed at the beginning and end of the trading day. This could be attributed to the heightened trading volume during these periods, leading to increased volatility in returns. Conversely, Figure 3 shows substantial dispersion in volatility conditioned on dates. However, given the ample number of dates, heteroskedasticity is unlikely to pose an estimation problem. The volatility conditioned on each stock does not exhibit significant sensitivity to heteroskedasticity.

In summary, concerning the intraday dimension, there appears to be a confluence of small sample sizes, substantial errors, and pronounced heteroskedasticity issues (referred to as "the

combined issues" in subsequent sections).

3 Models

In this section, I introduce three distinct models for intraday returns: the two-dimensional factor model, the three-dimensional factor model, and the three-dimensional factor model with heteroskedasticity, along with their respective underlying assumptions. I also outline the estimation techniques for each model: PCA for the two-dimensional factor model, TPCA for the three-dimensional factor model, and WTPCA for the three-dimensional factor model with heteroskedasticity. Additionally, I discuss the risks associated with dimension aggregation within a three-factor model, underscoring the importance of incorporating higher-frequency data in factor analysis. It is worth noting that WTPCA effectively preserves the inherent asymptotic properties and testability present in TPCA.

3.1 Two-dimensional Factor Model

When analyzing a two-dimensional factor model on intraday return data, we do not assume any common intraday variation.[7] Combining the date and intraday dimensions into a unified time dimension results in a structure with time and stock dimensions, akin to the low-frequency two-dimensional factor model.

[7]It is important to recognize that such variation might exist in either factors or loadings. Another approach is to extract principal components along the time dimension and attempt to recover intraday variation. However, this method could introduce error propagation and may not yield consistent results.

3.1.1 Model

For a two-dimensional factor model on intraday returns,

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

where $f_{t,j,r}$ is a factor driving co-movement, $\beta_{i,r}$ is the factor exposure. $i = 1, \ldots, N$ is stock, and $j, t = 1, \ldots, (P \times T)$ is time, containing T date and P intraday period information. r represents the number of factors and we have a total number of R factors. σ_r is the signal strength of the r-th factor. We would assume that β_{i,r_1} and β_{i,r_2} are independent while f_{t,r_1} and f_{t,r_2} are independent, $E[u_{i,j,t}] = 0$. A stronger assumption of the $u_{i,t}$ would be that $u_{i,j,t}$ is i.i.d. normal with mean 0.

In a matrix representation, $\mathcal{B}$ is a $N \times R$ matrix and F is a $T \times R$ matrix

$$Y = \sum_{r=1}^{R} \sigma_r \mathcal{B}_r \otimes F_r + U \tag{2}$$

where $\otimes$ is the tensor outer product. $\mathcal{B}_r$ is the r-th column vector of the factor matrix $\mathcal{B}$; F_r is the r-th column vector of the factor matrix F. All column vectors are unit vectors.

3.1.2 Estimation Method

In the observable factor literature, researchers typically construct $f_{t,r}$ based on theories or empirical evidence and employ regression or other statistical methods to estimate the coefficients (β's). Subsequently, they conduct tests to assess whether the factor effectively explains the cross-section of returns.

To extract the latent factors in Eq. 2, researchers can utilize the PCA method:

- The columns of $\mathcal{B}$ are the unit eigenvectors on YY^T.
- The columns of F are the unit eigenvectors on Y^TY.
- The signal strength σ_r is the square root of eigenvalues (in descending order).

The unit eigenvectors are called principal components (PCs).[8] The PCs along the time dimension are the factors and the ones along the stock dimension are the corresponding cross-sectional factor loadings.

3.2 Three-dimensional Factor Model

In the analysis of a three-dimensional model, we implicitly assume the presence of a common intraday variation. The three dimensions in this context are stock, intraday time, and date. It is important to note that the intraday variation can be attributed to either the factor, cross-sectional loading, or their product. The literature on systematic intraday beta variation suggests that we can have a three-dimensional model.

3.2.1 Model

For a three-dimensional factor model on intraday returns, if we assume a CP decomposition:

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

where $y_{i,j,t}$ is the stock i's five-minute return at date t intraday period j and $u_{i,f,t}$ is the noise for the corresponding return. $f_{t,r}$ 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.

[8] One can also generate $\mathcal{B}$, F, and σ_r with Singular Value Decomposition (SVD). SVD is the estimation method for the PCs in this paper.

$i = 1, \ldots, N$ is stock, $j = 1, \ldots, P$ is intraday period and $t = 1, \ldots, T$ is date. We would assume that β_{i,r_1} and β_{i,r_2} are independent, γ_{j,r_1} and γ_{j,r_2} are independent, f_{t,r_1} and f_{t,r_2} are independent, $E[u_{i,t}] = 0$.

In a matrix representation, $\mathcal{B}$ is a $N \times R$ matrix, Γ is a $P \times R$ matrix and F is a $T \times R$ matrix

$$Y = \sum_{r=1}^{R} \mathcal{B}_r \otimes \Gamma_r \otimes F_r + U \tag{4}$$

All notations are similar to the two-dimensional factor model. All column vectors are unit vectors.

3.2.2 Estimation Method

The ALS method can be employed for the above three-dimensional model estimation. In ALS, values are assigned to assumed dimensions, and a least squares problem is fitted for the remaining single estimation dimension. This process is iteratively repeated, alternating the estimation dimension and utilizing the estimated values for the assumed dimension. The procedure continues until a specified criterion, such as a small reduction in mean squared error, is met. It's essential to note that this method involves numerical minimization and may lead to a local minimum rather than a global minimum.

Another valid method would be TPCA developed by Babii et al. (2022). According to their paper, the TPCA algorithm for this data is as follows:

- Unfold the tensor Y along the 3 dimensions into $Y_{(1)}$, $Y_{(2)}$, $Y_{(3)}$. We can see the illustration of unfolding along the intraday dimension in Figure 4.[9]

[9]The unfolding's along date and stock dimension are presented in Appendix III Figure 21.

- Estimate M_j as $\hat{M}_j = (\hat{m}_{j,1}, \ldots, \hat{m}_{j,R})$ via PCA, i.e. take $\hat{m}_{j,r}$ is the unit norm eigenvector of $Y_{(j)}Y_{(j)}^T$ corresponding to the r-th largest eigenvalue.

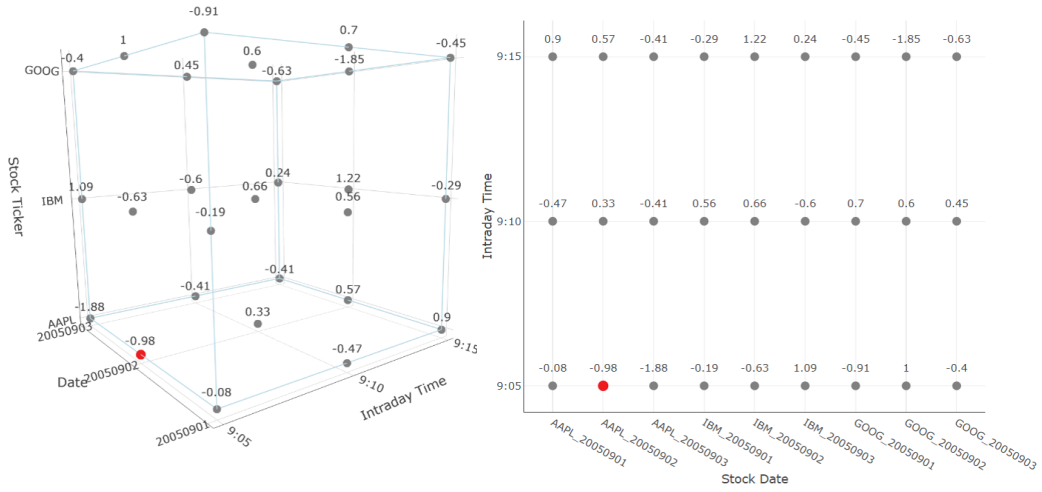
- Recover the scale components, $\hat{\sigma}_r$, as the square root of the r-th largest eigenvalue for the unfolded matrix along the largest dimension (date dimension in my sample).

Figure 4: Unfolding along Intraday Dimension

Panel (a) presents the previous example tensor data. Panel (b) presents the matricization of the tensor along the intraday dimension. The two data contain the same information, but PCA on the unfolded matrix will provide us with the intraday variation.

(a) Tensor Data

(b) Unfolding along Intraday Dimension for γ estimation



To analyze patterns in a target dimension, TPCA's unfolding consolidates all other dimensions into a single dimension and applies PCA to extract the variation in the target dimension. Subsequently, TPCA repeats this unfolding and PCA process for each dimension, extracting variations along each dimension independently. As each dimension estimation is an independent step, this method avoids error propagation.

3.3 Two vs. Three Dimension

If the true data-generating process (DGP) adheres to a three-dimensional factor model, what are the implications of estimating a two-dimensional factor model by amalgamating the date and intraday period into a single time dimension? Comparing the number of parameters, the two-dimensional model involves $(N + J \times T) \times R$ parameters, whereas the three-dimensional model has $(N + J + T) \times R$ parameters. The two-dimensional model requires the estimation of a significantly larger number of parameters. In the context of a squared error minimization problem, introducing an excessive number of variables in a two-dimensional model may lead to overfitting.

If the true DGP is a two-dimensional model, attempting to fit a three-dimensional model would lead to model misspecification. However, if there genuinely is no intraday pattern, we can anticipate the intraday dimension to be a constant $\frac{1}{\sqrt{P}}$.[10] Therefore, fitting a three-dimensional model should still provide us a sensible result.

3.4 Dimension Aggregation

I define dimension aggregation as the process of combining one (or more) dimension of a tensor to reduce its overall dimensionality.

An illustrative example of dimension aggregation involves consolidating individual stocks into portfolios. According to Andreou et al. (2022), this practice may lead to the diversification of risks and potentially obscure pertinent risk- or return-related features associated with individual assets. This highlights one potential drawback of dimension aggregation in data analysis.

In the context of intraday returns, another instance of dimension aggregation is combining

[10] This is because each Γ_r is a unit vector with dimension P.

along the intraday dimension to transform a three-dimensional tensor of intraday returns into a panel of daily returns. The resulting data structure is given by:

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

where P represents the length of the intraday dimension.

There are two potential pitfalls associated with conducting PCA on aggregated lower-frequency daily returns:

1. Loss of weak factors in daily data that are strong factors in intraday data.

2. Mixing factors with similar signal strength in daily data, which may have disparate signal strengths in intraday data.

3.4.1 Loss of Weak Factors

A weak factor is characterized by a small signal strength, denoted as σ_r, whereas a strong factor exhibits a large signal strength. Let's consider the DGP as given by Eq. 3, assuming simplicity with only one factor ($R = 1$). Suppose $\sigma = 10^4$, indicating a strong factor driving the variation of returns. However, for the sake of illustration, let $\sum_{j=1}^{P} \gamma_j = 10^{-7}$, signifying a small cumulative strength of intraday factors. If we aggregate the data along the intraday dimension:

$$
\begin{aligned}
y_{i,t} &= \sum_{j=1}^{P} y_{i,j,t} = \sum_{j=1}^{P} (\sigma \beta_i \lambda_j f_t + U_{i,j,t}) \\
&= 10^4 \beta_i \sum_{j=1}^{P} \lambda_j f_t + \sum_{j=1}^{P} U_{i,j,t} \\
&= 10^{-3} \beta_i f_t + \sum_{j=1}^{P} U_{i,j,t}
\end{aligned}
$$

It's evident that dimension aggregation, as demonstrated in this case along the intraday dimension, has the potential to diminish the signal strength in the daily factor model. The weak factor present in the daily data may then become overshadowed by noise, leading to a potential loss of factors. In essence, analyzing data at a higher frequency might aid in recovering some of the strong factors that might be obscured in the lower-dimension data.

When assuming i.i.d. noise, $\sum_{j=1}^{P} U_{i,j,t}$ scales the noise standard deviation up by $\sqrt{P}$. Additionally, $|\sum_{j=1}^{P} \lambda_j| \leq \sqrt{P}$. This implies that while the noise is scaled up by $\sqrt{P}$, the signal is always scaled up by less than $\sqrt{P}$. Consequently, dimension aggregation consistently results in a decrease in the signal-to-noise ratio.

It's important to note that the signal-to-noise ratio remains constant only when $|\sum_{j=1}^{P} \lambda_j| = \sqrt{P}$, which corresponds to $|\lambda_j| = \frac{1}{\sqrt{P}}$. As discussed earlier, this situation occurs when there are no intraday patterns, essentially constituting a two-dimensional factor model. Even in cases where the two-dimensional model is misspecified, using higher-frequency data in a higher-dimensional model does not lead to a decrease in the signal-to-noise ratio. In other words, in many cases, analyzing higher frequency returns is consistently preferable in terms of signal-to-noise ratio.

3.4.2 Mixing Factors with Similar Strength

Suppose we have the DGP as given by Eq. 3 with two factors. Let's assume that $\sigma_1 = 10\sigma_2$, enabling each estimation method to independently identify the two factors. Additionally, $\sum_{j=1}^{P} \gamma_{j,1} = 0.1 \sum_{j=1}^{P} \gamma_{j,1}$. Then, $\sigma_1 \sum_{j=1}^{P} \gamma_{j,1} = \sigma_2 \sum_{j=1}^{P} \gamma_{j,2} \equiv \hat{\sigma}$. If we aggregate the data along the intraday dimension:

$$\begin{aligned} y_{i,t} &= \sum_{j=1}^{P} y_{i,j,t} = \sum_{j=1}^{P} (\sigma_1 \beta_{i,1} \lambda_{j,1} f_{t,1} + \sigma_2 \beta_{i,2} \lambda_{j,2} f_{t,2} + U_{i,j,t}) \\ &= \hat{\sigma} \beta_{i,1} f_{t,1} + \hat{\sigma} \beta_{i,2} f_{t,2} + \sum_{j=1}^{P} U_{i,j,t} \end{aligned}$$

PCA on the aggregated daily data cannot independently identify the two factors with the same signal strength. Any basis of the space spanned by β_1, β_2 can be a valid result in the daily PCA.[11] In other words, higher-dimensional data can potentially help us separate signals that are mixed in lower frequency data.

3.5 Three-dimensional Factor Model with Heteroskedasticity

In real intraday return data, the DGP might follow a three-dimensional factor model with heteroskedasticity. In this section, I will define such a model and present a weighted version of TPCA capable of estimating all parameters while maintaining consistency and testability.

[11] Note that the converse can also be true. Higher dimension/frequency data PCs with the same signal strength might have different signal strengths in aggregated data.

3.5.1 Model

For a general three-dimensional factor model with heteroskedasticity on intraday returns:

$$y_{i,j,t} = \sum_{r=1}^{R} \sigma_r \beta_{i,r} \gamma_{j,r} f_{t,r} + w_{i,j,t} u_{i,j,t} \tag{5}$$

where $y_{i,j,t}$ is the stock i's five-minute return at date t intraday period j, $u_{i,f,t}$ is the noise for the corresponding return, and $w_{i,f,t}$ is the weighting factor which is the source of heteroskedasticity. $f_{t,r}$ 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. $i = 1, \ldots, N$ is stock, $j = 1, \ldots, P$ is intraday period and $t = 1, \ldots, T$ is date. We would assume that β_{i,r_1} and β_{i,r_2} are independent, γ_{j,r_1} and γ_{j,r_2} are independent, f_{t,r_1} and f_{t,r_2} are independent, $E[u_{i,j,t}] = 0$. $E[w_{i,j,t}^2] = 1$.[12]

Let's consider a special case where heteroskedasticity is dimension-separable and independent,

$$w_{i,j,t} = w_{1,i} w_{2,j} w_{3,t} \tag{6}$$

where $w_{1,i}, w_{2,j}, w_{3,t}$ are the heteroskedasticity along each of the stock, intraday, and date dimensions respectively. $E[w_{1,i}^2] = E[w_{2,j}^2] = E[w_{3,t}^2] = 1$.

For a more specific case, where we only have intraday dimension heteroskedasticity:

$$w_{i,j,t} = w_j \tag{7}$$

[12] When $E[w_{i,j,t}^2] = 1$, $Var[w_{i,j,t} u_{i,j,t}] = Var[u_{i,j,t}]$. The model has the same signal-to-noise ratio as the three-dimensional model with homoskedastic error.

where $E[w_j^2] = 1$.

3.5.2 Weighted TPCA

The weighted TPCA (WTPCA) can solve a model with heteroskedasticity along a single dimension or where heteroskedasticity is dimension-separable and independent. I will demonstrate the procedure and proof in the single-dimension case, where dimension-separable and independent case is almost identical.

Suppose we have a DGP as Eq. 5 and 7,

$$y_{i,j,t} = \sum_{r=1}^{R} \sigma_r \beta_{i,r} \gamma_{j,r} f_{t,r} + w_j u_{i,j,t} \tag{8}$$

We can utilize the following weighted procedure to perform the TPCA (WTPCA):

- Estimate weighted data: $\tilde{y}_{i,j,t} = \frac{y_{i,j,t}}{w_j}$.
- Perform the TPCA method on the weighted data and extract the signal strength and pattern in each dimension: $\tilde{\sigma}, \beta, \tilde{\gamma}, f$.
- Using the weighting vector to transform the result of weighted data back to parameters in the unweighted model. $\sigma = \tilde{\sigma}$, $\gamma_{j,r} = w_j \tilde{\gamma}_{j,r}$.

Proof that such a procedure can yield correct and consistent results can be found in Appendix IV. It's noteworthy that $\sigma = \tilde{\sigma}$, a crucial outcome as it ensures the estimate of σ remains unchanged, preserving the order of factors. This guarantees that all asymptotic properties and testable implications of TPCA apply to WTPCA.

This result can also be easily extended to the case where heteroskedasticity is dimension-separable and independent.

In Appendix IV, I also demonstrate that even with an incorrectly chosen weighting vector k_j, consistent estimation is still achievable. This implies that adjusting for heteroskedasticity is always preferable when applying TPCA.

4 Simulation

In this section, I initially showcase the TPCA and ALS estimates on data featuring homoskedastic large errors within small samples. Subsequently, I apply both methods to similar data but with intraday-dimension heteroskedasticity. It becomes apparent that neither TPCA nor ALS analysis on data with heteroskedasticity produces sensible results. However, WTPCA applied to the heteroskedastic data yields results akin to the TPCA result obtained from homoskedastic data. This suggests that WTPCA has the capability to alleviate issues related to severe heteroskedasticity, large noise, and small sample sizes.

4.1 Homoskedastic Data

4.1.1 DGP

The DGP for the homoskedastic data is as follows:

$$y_{i,j,t} = \sum_{r=1}^{3} \sigma_r \beta_{i,r} \gamma_{j,r} f_{t,r} + u_{i,j,t} \tag{9}$$

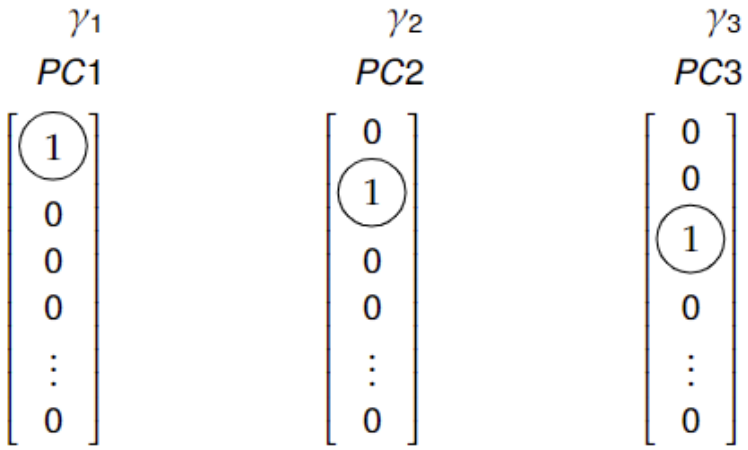
Where the dimension of y is $80 \times 80 \times 2000$. Each $\beta_{i,r}$, $\gamma_{j,r}$, $f_{t,r}$ is simulated using a standard normal distribution and then normalized to 1. $\sigma_1 = 3, \sigma_2 = 2, \sigma_3 = 1$. $u \sim N(0, 0.04)$. The sample size mimics the actual intraday data I have. In this DGP, we have a relatively large error and a small sample.

4.1.2 Focus Plots

To elucidate the emphasis of each Principal Component (PC) in the PCA analysis and its degree of focus, I will introduce two "Focus Plots": "Position of Focused Dimension" and "Value of Focused Dimension". Let's consider the estimated PCs along the intraday dimension to be the Identity Matrix Columns (as depicted in Figure 5). In this case, the j-th PC would have the largest value 1 at the j-th position. The resulting "Position of Focused Dimension" and "Value of Focused Dimension" plots are showcased in Figure 6.

Figure 5: Example of PCs as the Identity Matrix Columns

Suppose that we have estimated PCs along the intraday dimension forming the basis of the identity matrix. The first three PCs would look like the following plot. The circled number is where the largest (absolute) value in the eigenvector occurs.

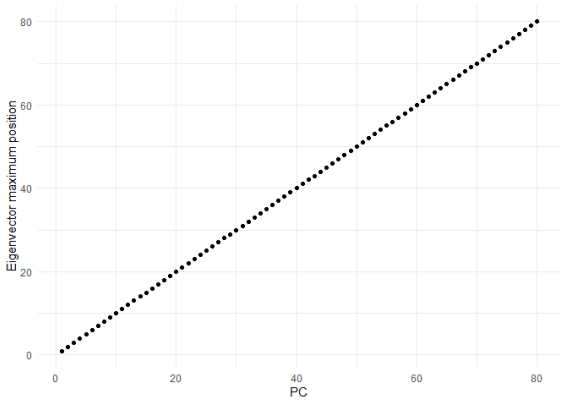


The "Position of Focused Dimension" for the r-th PC signifies which intraday period contributes the most to it. Meanwhile, the "Value of Focused Dimension" for the r-th PC illustrates the maximum contribution from a single intraday period. The value of the focused dimension ranges from $\frac{1}{\sqrt{P}}$ to 1, where $\frac{1}{\sqrt{P}}$ indicates an equal contribution from all intraday

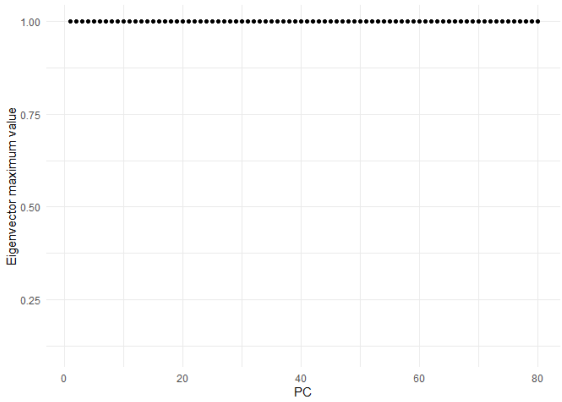
Figure 6: Example of Focus Plots

Panel (a) presents the "Position of Focused Dimension". Panel (b) presents the "Value of Focused Dimension. The value of focused dimension ranges from $\frac{1}{\sqrt{P}}$ to 1, where $\frac{1}{\sqrt{P}}$ indicates an equal contribution from all intraday periods and 1 indicates a single focus on one dimension. In this particular example, we have r-th PC solely focusing on the r-th intraday period.

(a) Position of Focused Dimension



(b) Value of Focused Dimension



periods, and 1 indicates a single focus on one dimension.[13] For instance, in the PCs depicted in Figure 5, the r-th PC (the r-th factor) would concentrate solely on the r-th intraday period.

This example represents a special case where each intraday period has a distinct factor structure, meaning the first factor can only explain the first five-minute returns, and so on. However, this might not be a sensible model for actual return data, as we anticipate factors that can explain the common variation in returns.

4.1.3 Estimations With TPCA and ALS

Both TPCA and ALS methods are applied to the simulated data. Only the results of the γ's are reported, as this is the dimension where heteroskedasticity will be introduced. I only

[13] P is the size of the intraday dimension.

display the estimates for the first three factors, considering that the true DGP involves three factors.

Figure 7: TPCA and ALS on Simulated Homoskedastic Data

Panel (a) presents the intraday dimension estimation from TPCA on homoskedastic data and the true γ. Panel (b) presents the intraday dimension estimation from ALS on the same homoskedastic data and the true γ. Both methods yield sensible results for the three factors.

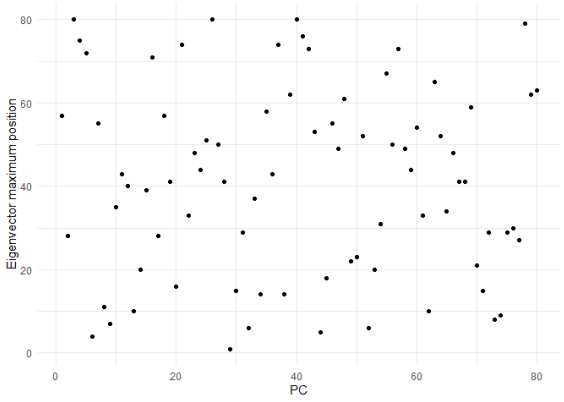
(a) TPCA Estimates

(b) ALS Estimates

Figure 8: TPCA Focus Plots for Simulated Homoskedastic Data

Panel (a) presents the "Position of Focused Dimension" for the TPCA γ estimation of the simulated homoskedastic data. Panel (b) presents the "Value of Focused Dimension" for such an estimation.

(a) Position of Focused Dimension



(b) Value of Focused Dimension

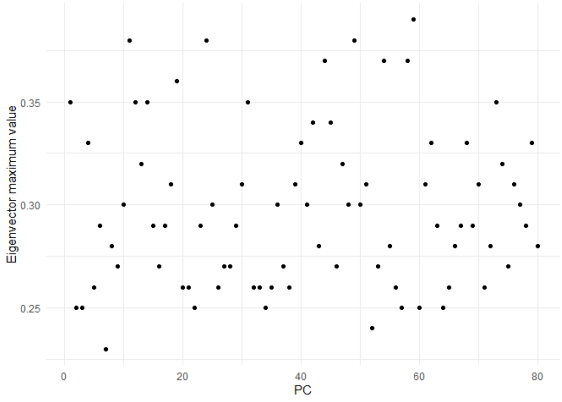


Figure 7 demonstrates that both TPCA and ALS produce sensible results for the intraday dimension, up to three factors. The Focus Plots for the TPCA estimates (Figure 8) reveal that TPCA does not yield a result focusing on a single intraday period. This aligns with expectations, considering that the DGP does not emphasize a single intraday period.

4.1.4 Goodness of Fit Comparison

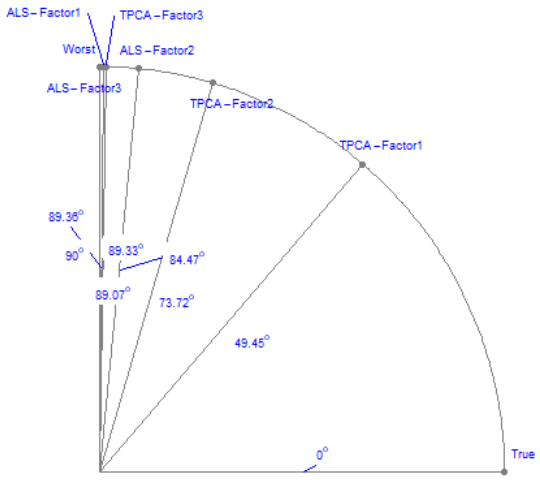
Subsequently, I proceed to quantitatively assess the goodness of fit of the γ's. Given that the γ's are all normalized to 1, a conventional Mean Squared Error (MSE) report might not offer a comprehensive understanding of the absolute goodness of fit. Instead, I calculate the angle between the fitted and actual γ's and illustrate these angles in the "Quarter Circle Plot." As both the fitted and actual γ vectors are normalized, the MSE serves as a direct mapping to the angle between them. The optimal estimate would exhibit a 0^o angle to the true γ vector (perfect alignment), while the least favorable estimate would present a 90^o angle (orthogonal) to the true γ vector.

As illustrated in Figure 9, TPCA provides γ estimates with less error compared to ALS.[14] The ALS-estimated γ_1 is nearly orthogonal to the true γ_1. In the TPCA case, the fit for subsequent factors is compromised owing to a lower signal-to-noise ratio. The ensuing analysis entails applying the same evaluations to the heteroskedastic data and comparing the results with the homoskedastic case.

[14] This discrepancy might arise from ALS not leveraging the orthogonality of the factors in each dimension.

Figure 9: Quarter Circle Plot for TPCA and ALS on Simulated Homoskedastic Data

The quarter circle plot presents the angle between the estimated and true γ. Since both are normal vectors, the estimation error would be a one-to-one mapping to the angle between them. The best estimate would be 0^o (perfectly aligned) to the true γ vector while the worst estimate would be 90^o (orthogonal) to the true γ vector.



4.2 Heteroskedastic Data

4.2.1 DGP

The DGP for the heteroskedastic data is almost identical to the homoskedastic case in Eq. 9. The only new component is the heteroskedasticity weights, w_j:

$$y_{i,j,t} = \sum_{r=1}^{3} \sigma_r \beta_{i,r} \gamma_{j,r} f_{t,r} + w_j u_{i,j,t} \tag{10}$$

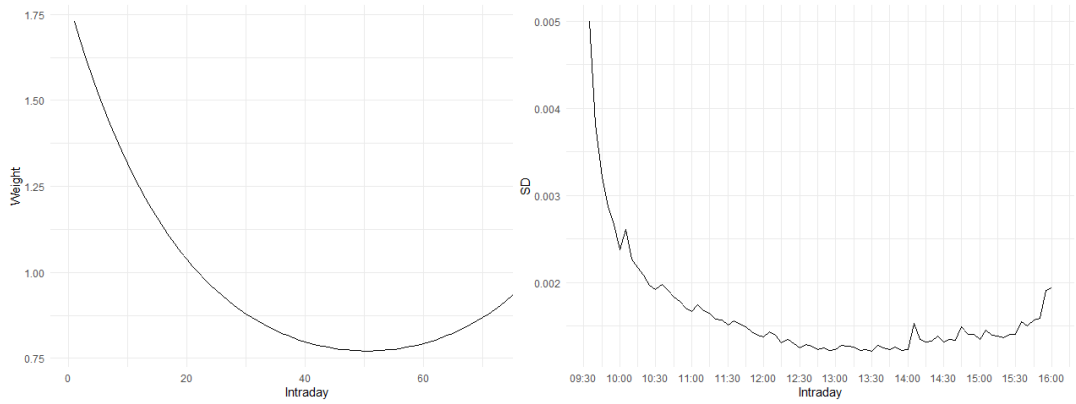
To ensure comparability, I will use the same values of σ_r, $\beta_{i,r}$, $\gamma_{i,r}$, $f_{t,r}$ and $u_{i,j,t}$ as the simulated homoskedastic data. $w_j = \frac{e^{101-j}+e^j+0.5}{\sqrt{\frac{1}{80}*\sum_{j=1}^{80}(e^{101-j}+e^j+0.5)^2}}$. As depicted in Figure 10, w_j is

expressed in this form to mirror the standard deviation of actual returns along the intraday dimension.

Figure 10: Wights and Intraday Heteroskedasticity

Panel (a) plots the w_j for the heteroskedasticity data simulation. Panel (b) plots the standard deviation of actual returns along the intraday dimension. They are very similar.

(a) Weights in Simulation (b) Intraday Heteroskedasticity



4.2.2 Estimations With TPCA and ALS

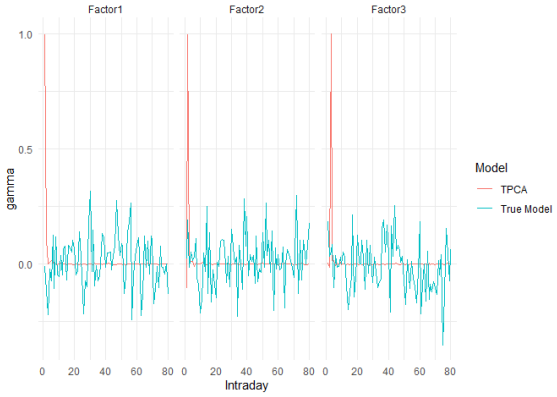
Following the approach in the previous section, I will present the first three factors' γ from both TPCA and ALS. Figure 11 reveals that neither TPCA nor ALS produces sensible results; they all excessively focus on one of the dimensions with more volatile error. This data issue can lead to estimation problems.

The Focus Plots for the TPCA estimates (Figure 12) further highlight that TPCA excessively focuses on single dimensions. The "Position of Focused Dimension" plot closely resembles a rotation of the heteroskedasticity weight plot in Figure 10 Panel (a). It's worth noting that in Figure 10 Panel (a), the last period has roughly the 21st highest volatility, and in Figure 12, the 21st PC concentrates on the last intraday period. This alignment appears

Figure 11: TPCA and ALS on Simulated Heteroskedastic Data

Panel (a) presents the intraday dimension estimation from TPCA on heteroskedastic data and the true γ. Panel (b) presents the intraday dimension estimation from ALS on the same homoskedastic data and the true γ. Neither method yields sensible results for the three factors. They all focus too much on single dimensions.

(a) TPCA Estimates

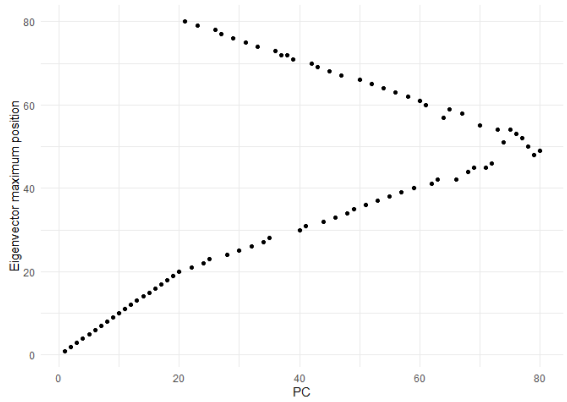


(b) ALS Estimates

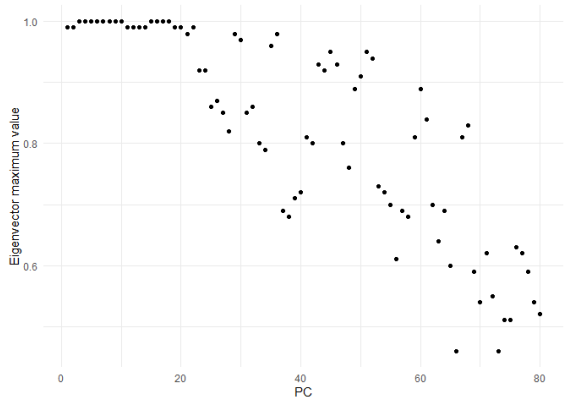
Figure 12: TPCA Focus Plots for Simulated Heteroskedastic Data

Panel (a) presents the "Position of Focused Dimension" for the TPCA γ estimation of the simulated heteroskedastic data. Panel (b) presents the "Value of Focused Dimension" for such an estimation. This further presents that the TPCA estimates falsely focus too much on single dimensions.

(a) Position of Focused Dimension



(b) Value of Focused Dimension



to be more than a coincidence. The "Value of Focused Dimension" plot exhibits values close to 1 for the first 20 PCs. The combined issues can contribute to a false focus on the noisy periods.

4.2.3 WTPCA Estimation

In contrast to the outcomes of TPCA and ALS, WTPCA would yield more accurate estimates, as depicted in Figure 13. The fits for the last two factors are inferior to the first one, but this is inevitable due to larger errors. This result mirrors the TPCA result on comparable homoskedastic data in Figure 7 Panel (a), confirming the efficacy of WTPCA on heteroskedastic data with combined issues. The Focus Plots for the WTPCA estimates (Figure 14) further suggest that WTPCA no longer excessively focuses on single dimensions.

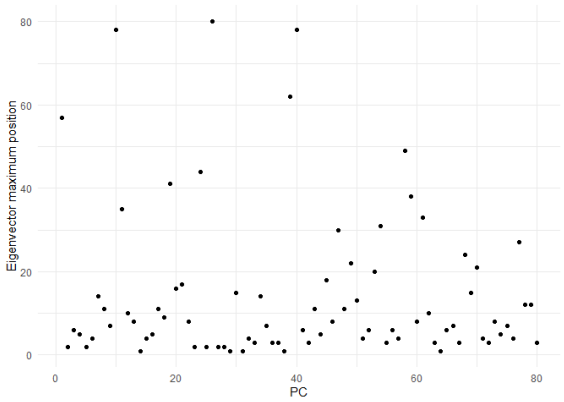
Figure 13: WTPCA on Simulated Heteroskedastic Data

This plot presents the first three factor's intraday dimension estimates with the WTPCA method. It seems that the result is more aligned with the true process compared to the TPCA and ALS estimates.

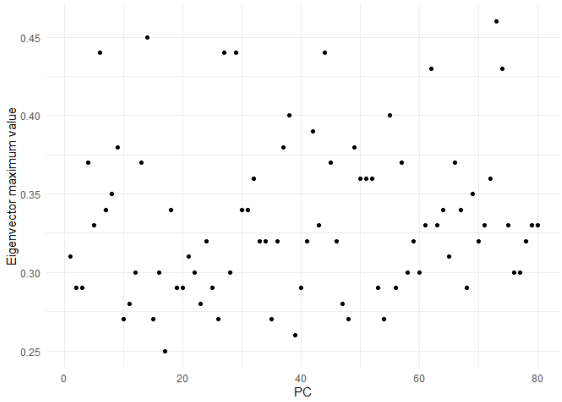
Figure 14: WTPCA Focus Plots for Simulated Heteroskedastic Data

Panel (a) presents the "Position of Focused Dimension" for the WTPCA γ estimation of the simulated heteroskedastic data. Panel (b) presents the "Value of Focused Dimension" for such an estimation. This further indicates that the WTPCA can mitigate the data issue.

(a) Position of Focused Dimension



(b) Value of Focused Dimension

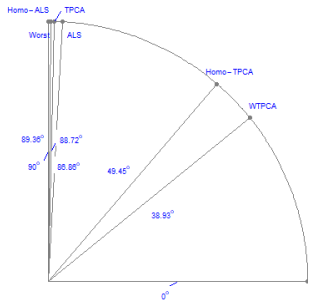


4.2.4 Goodness of Fit Comparison

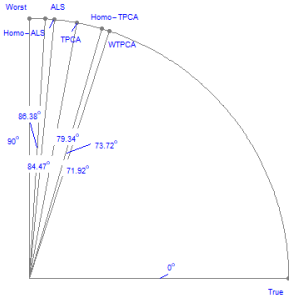
Figure 15: Quarter Circle Plot for All Simulated Data Estimates

These quarter circle plots present the estimation error of each method on different data.

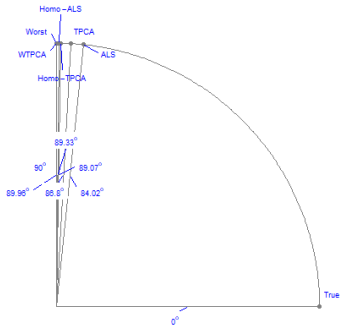
(a) First Factor



(b) Second Factor



(c) Third Factor



As depicted in Figure 15, in the heteroskedastic case, WTPCA provides the most accurate γ estimates, while TPCA and ALS yield considerably poorer results. WTPCA appears to

effectively alleviate the combined issues.[15]

5 Empirical Results

In this section, I present the results of TPCA on true intraday returns, focusing on the intraday dimension estimates for up to three factors. It appears that the results are similar to the simulation outcome when faced with the combined issues. The empirical estimates with WTPCA seem to provide more sensible results.

5.1 TPCA Intraday Results

The TPCA on high-frequency returns produces similar results in the Date and Stock dimensions as the PCA on daily data.[16] In previous sections, I argued that the combined issues might arise in the intraday dimension, potentially leading to estimation problems. Therefore, in this section, I will specifically examine the intraday dimension estimates.

The TPCA intraday dimension estimates for the first three factors appear to be nonsensical. They excessively focus on single dimensions, similar to the TPCA result on simulated heteroskedastic data.[17]

The Focus Plots for the TPCA estimates (Figure 8) exhibit similarities to those for simulated heteroskedastic data TPCA (Figure 12). The "Position of Focused Dimension" plot closely resembles a rotation of the heteroskedasticity weight plot in Figure 10 Panel (b). Notably, in Figure 10 Panel (b), the last period has approximately the 12th highest

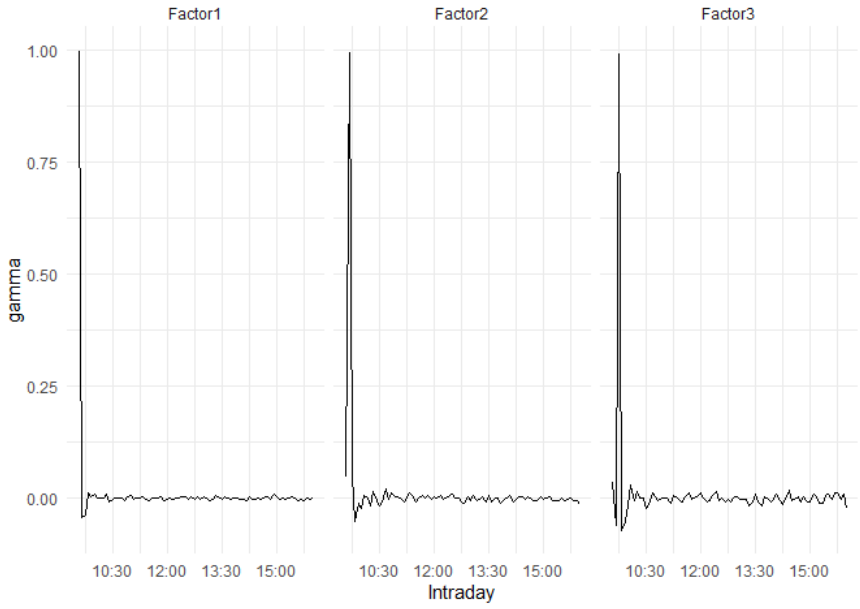
[15] In general, for all these methods, the later factor estimates have a larger angle with the true value. This indicates that subsequent factors exhibit a less favorable fit due to a lower signal-to-noise ratio. This pattern is also observed in the fits for homoskedastic data.

[16] This result is not presented in this paper.

[17] In Appendix V Figure 22, we can observe that ALS also focuses on single dimensions. This further supports the notion that the intraday return sample might suffer from the combined issues.

Figure 16: TPCA Intraday Results on Actual High-frequency Returns

This plot presents the first three factor's intraday dimension estimates with the TPCA method on actual high-frequency returns. The result looks similar to the simulation result where we have the combined issues.



volatility, and in Figure 8, the 12th PC focuses on the last intraday period. Such alignment was also observed for simulated heteroskedastic data. The "Value of Focused Dimension" plot indicates values close to 1 for the first few PCs, further suggesting that TPCA excessively focuses on single dimensions, and the intraday return sample might suffer from the combined issues.

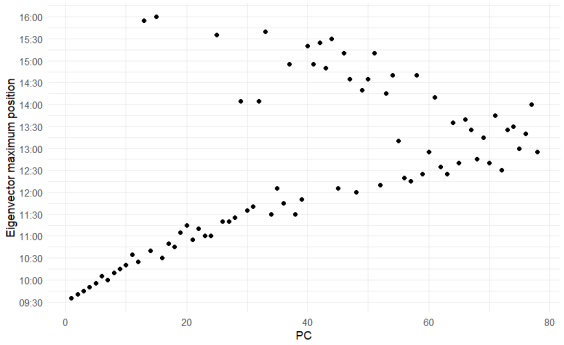
5.2 WTPCA Intraday Results

Figure 18 reveals that the WTPCA intraday dimension estimates for the first three factors are more sensible than TPCA (and ALS). While we cannot quantify the goodness of fit, the WTPCA estimates no longer excessively focus on single dimensions.

Figure 17: TPCA Focus Plots for Actual High-frequency Returns

Panel (a) presents the "Position of Focused Dimension" for the TPCA γ estimation of the actual data. Panel (b) presents the "Value of Focused Dimension" for such an estimation. This further indicates that the TPCA estimates focus too much on single dimensions. These plots look similar to the Focus Plots of TPCA on simulated heteroskedastic data (in Figure 12)

(a) Position of Focused Dimension



(b) Value of Focused Dimension

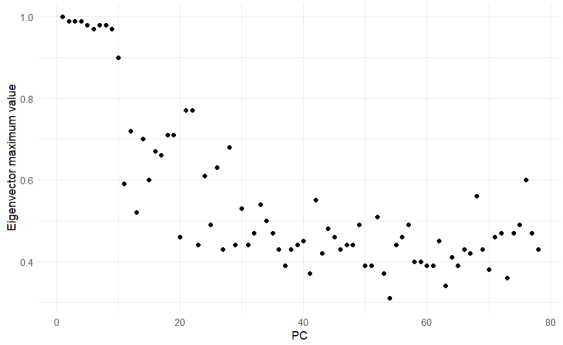


Figure 18: WTPCA Intraday Results on Actual High-frequency Returns

This plot presents the first three factor's intraday dimension estimates with the WTPCA method on actual high-frequency returns. The result looks similar to the simulation result where we have the combined issues but apply the WTPCA method.

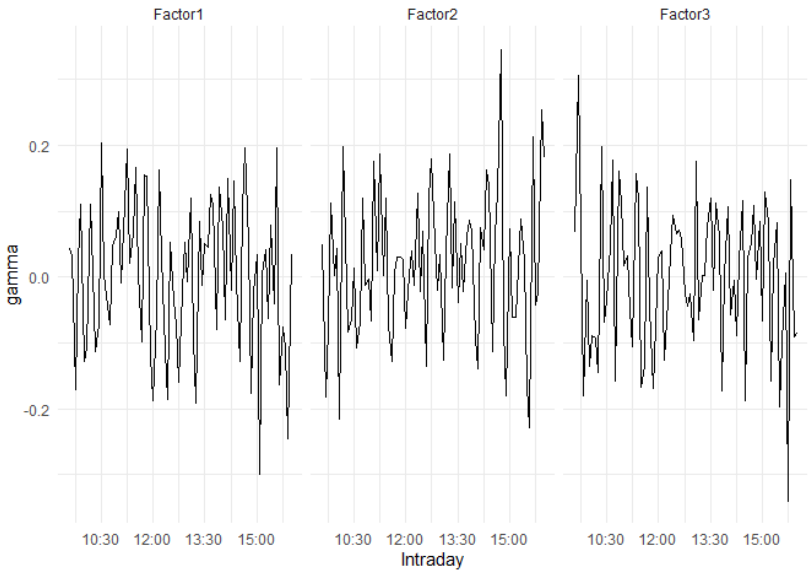
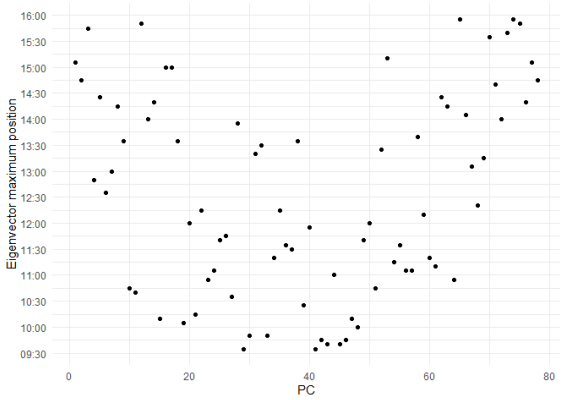


Figure 19: WTPCA Focus Plots for Actual High-frequency Returns

Panel (a) presents the "Position of Focused Dimension" for the WTPCA γ estimation of the actual data. Panel (b) presents the "Value of Focused Dimension" for such an estimation. This further indicates that the WTPCA estimates no longer focus on single dimensions.

(a) Position of Focused Dimension



(b) Value of Focused Dimension

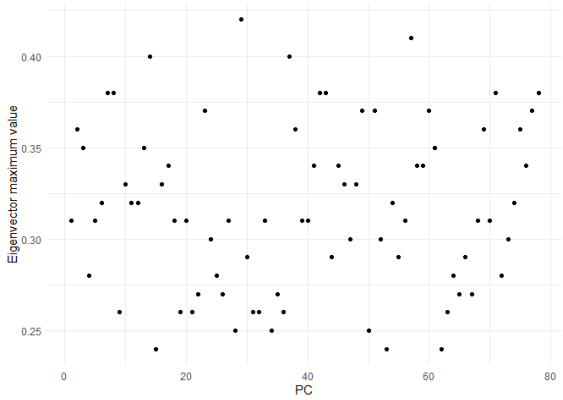


Figure 19 looks similar to the Focus Plots of WTPCA on simulated heteroskedastic data (in Figure 14). This further indicates that WTPCA might help to mitigate the combined issues in the actual data.

6 Conclusion

Arguing against dimension aggregation, this paper establishes that, under i.i.d. noise, utilizing higher frequency is always preferable in terms of signal-to-noise ratio. Given evidence of intraday beta variation, a three-dimensional model may be suitable for intraday returns, and Tensor Principal Component Analysis (TPCA), developed by Babii et al. (2022), can be employed for estimation.

However, there is evidence suggesting that intraday return data might face challenges such as small sample sizes, significant noise, and severe heteroskedasticity. This evidence is

twofold: the return standard deviation for each intraday period exhibits significant variation, and the TPCA fit of actual data resembles the TPCA fit on simulated data with combined issues. Such data issues can lead to substantial estimation errors with TPCA, primarily due to a false focus on noisy periods. To address this, a weighted version of TPCA is proposed, which shows potential in mitigating the combined issues in both simulated and actual data.

Under dimension-separable and independent heteroskedasticity, Weighted Tensor Principal Component Analysis (WTPCA) is favored in small samples and can deliver consistent results in large samples. Even when using the wrong weighting vector or when heteroskedasticity is not dimensionally separable or independent, WTPCA can still yield consistent factor estimates.

Further exploration could involve testing the number of factors with WTPCA on intraday returns and assessing whether employing a time-varying weighting vector (such as the previous month's return standard deviation for each intraday period) improves estimates. Additionally, WTPCA can be applied to subsamples of intraday return tensors, such as samples on dates with macroeconomic announcements, to uncover the relationship between these events and asset returns.

References

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

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

Andreou, E., Gagliardini, P., Ghysels, E., and Rubin, M. (2022). Three common factors.

Andreou, E., Gagliardini, P., Ghysels, E., and Rubin, M. (2023). 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., Jacod, J., and Xiu, D. (2021). Inference on risk premia in continuous-time asset pricing models. Working paper.

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

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

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

Boudt, K., Kleen, O., and Sjørup, E. (2022). Analyzing intraday financial data in r: The highfrequency package. *Journal of Statistical Software*, 104(8).

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

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

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

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

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3):455–500.

Lettau, M. (2023). 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.

Appendix

Appendix I: List of S&P 100 Permanent Members

This list of S&P 100 stocks was obtained on 7/9/2023. Permanent members are the ones remaining on the S&P 500 list for the whole duration of the sample period (1/2/2009 to 7/17/2023).

The list of 78 permanent stocks with correct prices:

AAPL	ABT	ADBE	AIG	AMGN	AMT	AMZN	AXP	BA
BAC	BK	BMY	C	CAT	CL	CMCS	COF	COP
COST	CRM	CSCO	CVS	CVX	DHR	DIS	DUK	EMR
EXC	F	FDX	GD	GE	GILD	GOOG	GS	HD
HON	IBM	INTC	JNJ	JPM	KO	LLY	LMT	LOW
MA	MCD	MDT	MET	MMM	MO	MRK	MS	MSFT
NKE	NVDA	ORCL	PEP	PFE	PG	PM	QCOM	SBUX
SCHW	SO	SPG	T	TGT	TMO	TXN	UNH	UNP
UPS	USB	VZ	WFC	WMT	XOM			

The list of four permanent stocks with issues:

- "LIN": Linde PLC. It was Praxair (PX) before the merger on 10/30/2018. However, the "PX" prices after the merger (10/30/2018) are incorrect, and "LIN" is not included in TAQ.
- "NEE": NextEra Energy. There was a name change from FPL Group (FPL) in March 2010. However, "NEE" data started on 1/2/2015 in TAQ, and "FPL" is not included

in TAQ.

- "RTX": Raytheon Technologies Corporation. It was Raytheon Company (RTN) until 4/3/2020 when it merged with UTC. "RTN" was discontinued after the merger and "RTX" is not included in TAQ.

- "WBA": Walgreens Boots Alliance. It was Walgreens (WAG) before December 31, 2014, but "WAG" disappeared from TAQ at the end of 2012. "WBA" is not included in TAQ.

Appendix II: Preliminary Analysis of Intraday Returns

Table 1: Summary Statistics for the five-minute returns (1)

This table describes the moments of the intraday five-minute returns for all permanent S&P 100 stocks. SD is the standard deviation and AutoCorr is the lag-one autocorrelation (five-minute lag).

Ticker	Mean	Median	SD	Skewness	Kurtosis	AutoCorr
AAPL	0	0	0.0016	-0.0076	20.5899	-0.0152
ABT	0	0	0.0014	-0.3351	30.0177	-0.0193
ADBE	0	0	0.0019	0.0254	20.1755	-0.0083
AIG	0	0	0.0030	-1.0262	287.2658	-0.0130
AMGN	0	0	0.0016	0.1625	56.1765	-0.0316
AMT	0	0	0.0016	-0.0692	34.7056	-0.0106
AMZN	0	0	0.0020	-0.0512	13.6472	-0.0212
AXP	0	0	0.0019	0.1273	32.1240	-0.0053
BA	0	0	0.0020	-0.2619	49.1425	-0.0097
BAC	0	0	0.0025	-1.1688	80.6296	-0.0163
BK	0	0	0.0019	0.6211	90.7197	-0.0218
BMY	0	0	0.0015	-0.4177	192.2175	-0.0310
C	0	0	0.0026	-0.6577	169.7712	-0.0101
CAT	0	0	0.0019	-0.0470	15.9951	-0.0028
CL	0	0	0.0012	0.3101	44.3317	-0.0236
CMCS	0	0	0.0017	-0.1024	30.8251	-0.0135
COF	0	0	0.0025	0.4366	57.6793	0.0046
COP	0	0	0.0021	-0.0798	23.9140	-0.0034
COST	0	0	0.0013	0.0308	26.5822	-0.0293

Table 2: Summary Statistics for the five-minute returns (2)

This table describes the moments of the intraday five-minute returns for all permanent S&P 100 stocks. SD is the standard deviation and AutoCorr is the lag-one autocorrelation (five-minute lag).

Ticker	Mean	Median	SD	Skewness	Kurtosis	AutoCorr
CRM	0	0	0.0022	0.4845	36.3043	-0.0030
CSCO	0	0	0.0015	0.1466	25.1647	-0.0272
CVS	0	0	0.0016	-0.6763	736.0063	-0.0324
CVX	0	0	0.0016	0.2990	30.8511	-0.0142
DHR	0	0	0.0015	-0.0320	25.2582	-0.0256
DIS	0	0	0.0016	-0.2658	27.8542	-0.0090
DUK	0	0	0.0013	-0.1589	20.4663	-0.0170
EMR	0	0	0.0016	0.5845	31.8259	-0.0232
EXC	0	0	0.0015	-0.2454	54.5299	-0.0123
F	0	0	0.0024	-0.4543	73.3754	-0.0376
FDX	0	0	0.0018	0.0554	22.4163	-0.0151
GD	0	0	0.0015	-0.0176	25.3157	-0.0173
GE	0	0	0.0021	-0.2798	32.8719	-0.0141
GILD	0	0	0.0017	-0.1122	27.3831	-0.0168
GOOG	0	0	0.0016	0.0846	15.6812	-0.0222
GS	0	0	0.0018	-0.1508	30.8631	-0.0155
HD	0	0	0.0015	0.0733	34.7498	-0.0177
HON	0	0	0.0015	-0.2726	28.3429	-0.0166
IBM	0	0	0.0013	0.0869	22.6435	-0.0121
INTC	0	0	0.0018	0.4066	24.6842	-0.0219

Table 3: Summary Statistics for the five-minute returns (3)

This table describes the moments of the intraday five-minute returns for all permanent S&P 100 stocks. SD is the standard deviation and AutoCorr is the lag-one autocorrelation (five-minute lag).

Ticker	Mean	Median	SD	Skewness	Kurtosis	AutoCorr
JNJ	0	0	0.0012	-0.3990	312.2674	-0.0471
JPM	0	0	0.0019	0.0416	29.0428	-0.0163
KO	0	0	0.0011	-0.1706	27.0291	-0.0242
LLY	0	0	0.0015	-0.3532	38.2672	-0.0127
LMT	0	0	0.0014	-0.0948	25.9915	-0.0202
LOW	0	0	0.0018	-0.2608	40.5455	-0.0162
MA	0	0	0.0018	1.2807	162.2622	-0.0216
MCD	0	0	0.0012	1.4011	111.2488	-0.0176
MDT	0	0	0.0014	-0.2315	45.4742	-0.0173
MET	0	0	0.0022	0.0846	38.4602	-0.0071
MMM	0	0	0.0014	0.2226	74.1085	-0.0110
MO	0	0	0.0014	-1.2886	109.6418	-0.0104
MRK	0	0	0.0014	-0.2274	40.6357	-0.0156
MS	0	0	0.0023	0.2422	26.8784	-0.0083
MSFT	0	0	0.0016	0.0901	15.2209	-0.0185
NKE	0	0	0.0016	0.5382	47.3168	-0.0240
NVDA	0	0	0.0026	0.1313	14.5301	-0.0043
ORCL	0	0	0.0015	-0.1081	28.2406	-0.0187
PEP	0	0	0.0012	-0.1794	48.7742	-0.0312
PFE	0	0	0.0015	-0.2670	27.4956	-0.0234

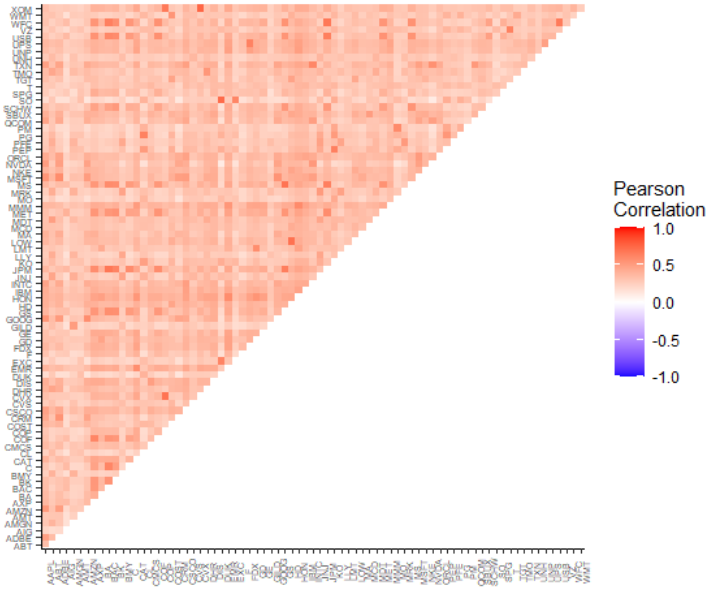
Table 4: Summary Statistics for the five-minute returns (4)

This table describes the moments of the intraday five-minute returns for all permanent S&P 100 stocks. SD is the standard deviation and AutoCorr is the lag-one autocorrelation (five-minute lag).

Ticker	Mean	Median	SD	Skewness	Kurtosis	AutoCorr
PG	0	0	0.0012	0.3343	50.5874	-0.0340
PM	0	0	0.0014	-0.2718	66.8321	-0.0229
QCOM	0	0	0.0019	0.7415	41.6805	-0.0124
SBUX	0	0	0.0017	0.2260	25.8865	-0.0102
SCHW	0	0	0.0021	-0.5707	39.3666	-0.0103
SO	0	0	0.0013	-0.2648	31.8505	-0.0121
SPG	0	0	0.0022	-0.0055	64.8528	-0.0079
T	0	0	0.0013	-3.5889	348.1169	-0.0175
TGT	0	0	0.0017	-0.0622	24.1193	-0.0193
TMO	0	0	0.0016	-0.0809	27.5885	-0.0135
TXN	0	0	0.0017	0.1614	18.6079	-0.0304
UNH	0	0	0.0018	-0.3069	90.2466	-0.0071
UNP	0	0	0.0017	-0.1464	49.6376	-0.0235
UPS	0	0	0.0014	0.4949	50.3701	-0.0192
USB	0	0	0.0020	-0.1336	85.4469	0.0005
VZ	0	0	0.0013	-1.2708	94.3267	-0.0228
WFC	0	0	0.0022	0.5448	72.7320	-0.0077
WMT	0	0	0.0012	-0.6047	55.1182	-0.0201
XOM	0	0	0.0016	0.4804	43.2163	-0.0091

Figure 20: Correlation of the S&P 100 stock returns

The 78 stock five-minute return correlations are presented. There is not a single pair of stocks whose correlation is negative. This suggests that all stocks tend to have a positive loading on the major systematic risk (usually the market risk).



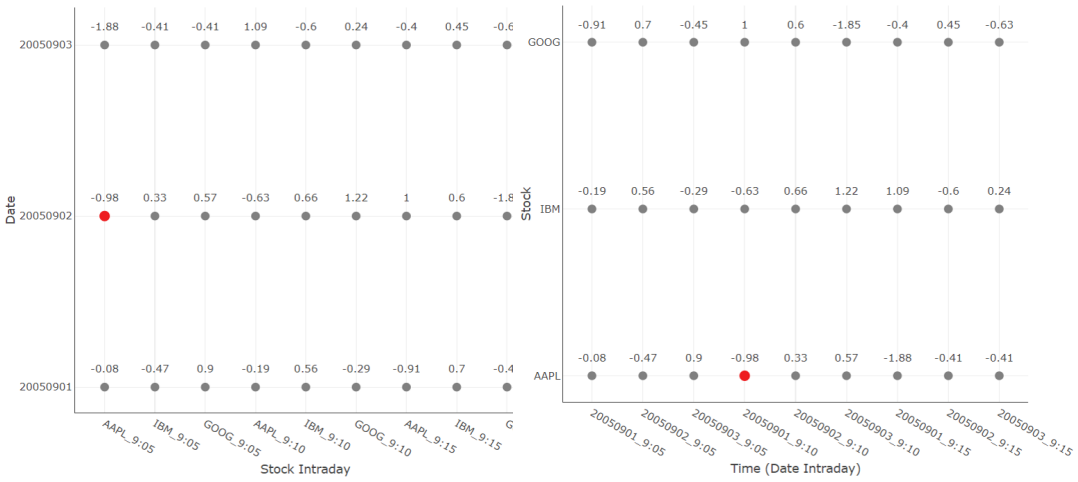
Appendix III: Tensor Unfolding along Date and Stock Dimensions

Figure 21: Tensor Unfolding along Date and Stock Dimensions

Panel (a) presents the unfolding of the tensor (Figure 4 Panel (a)) along the Date dimension. Panel (b) presents the unfolding of the same tensor along the Date dimension.

(a) Unfolding along Date Dimension for F estimation

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



Appendix IV: Proof of WTPCA Procedure

The original model is as Eq. 8. The model for the weighted data:

$$\tilde{y}_{i,j,t} = \sum_{r=1}^{R} \tilde{\sigma}_r \beta_{i,r} \tilde{\gamma}_{j,r} f_{t,r} + u_{i,j,t} \tag{11}$$

We have the same noise structure as in the homoskedastic case. Mathematically, only the σ and γ are affected in the model after the adjustment.

Note that we can first divide the weighting vector from Eq. 8 and then normalize the γ:

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

Since both vector of γ and vector of $\tilde{\gamma}$ are normal, we would have:

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

The summed square of $w_j \tilde{\gamma}$ would yield:

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

This result is powerful since it indicates that both the signal strength estimates and order are preserved in WTPCA. TPCA asymptotic properties are, therefore, also preserved. Tests

based on the asymptotic properties of TPCA would also work for WTPCA.

If we use a wrong weighting vector k_j instead of w_j. Since we need to have the weighting vector $E[k_j] = 1$, similar to the previous proof:

$$\bar{y}_{i,j,t} \equiv \frac{y_{i,j,r}}{k_j} = \sum_{r=1}^{R} \bar{\sigma}_r \beta_{i,r} \bar{\gamma}_{j,r} f_{t,r} + \frac{w_j}{k_j} u_{i,j,t} \tag{12}$$

We would have $\sigma_r = \bar{\sigma}_r$ and $\gamma_{j,r} = k_j \bar{\gamma}_{j,r}$. Since TPCA can consistently estimate the $\bar{\gamma}_{j,r}$, even with the presence of $\frac{w_j}{k_j}$ heteroskedasticity, the WTPCA procedure can consistently estimate the $\gamma_{j,r}$ even we use the wrong weighing vector.

Appendix V: ALS Intraday Dimension Estimates

Figure 22: ALS Intraday Results on Actual High-frequency Returns

This plot presents the first three factors' intraday dimension estimates with the ALS method on actual high-frequency returns. Again, the result looks similar to the simulation result where we have the combined issues.