Stony Brook University

OFFICIAL COPY

The official electronic file of this thesis or dissertation is maintained by the University Libraries on behalf of The Graduate School at Stony Brook University.

© All Rights Reserved by Author.

NEW DEVELOPMENT ON MARKET MICROSTRUCTURE AND MACROSTRUCTURE: PATTERNS OF U.S. HIGH FREQUENCY DATA AND A UNIFIED FACTOR MODEL FRAMEWORK

A Dissertation Presented

by

Xu Dong

to

The Graduate School

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Quantitative Finance

Stony Brook University

Dec 2013

Stony Brook University

The Graduate School

Xu Dong

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Andrew Mullhaupt – Dissertation Advisor
Research Professor
Department of Applied Mathematics and Statistics, Stony Brook University

Svetlozar Rachev – Chairperson of Defense
Director of Quantitative Finance Program
Department of Applied Mathematics and Statistics, Stony Brook University

Haipeng Xing – Committee Member
Associate Professor
Department of Applied Mathematics and Statistics, Stony Brook University

Noah Smith – External Committee Member
Assistant Professor
College of Business, Stony Brook University

This dissertation is accepted by the Graduate School

Charles Taber
Interim Dean of the Graduate School

Abstract of the Dissertation

NEW DEVELOPMENT ON MARKET MICROSTRUCTURE AND MACROSTRUCTURE: PATTERNS OF U.S. HIGH FREQUENCY DATA AND A UNIFIED FACTOR MODEL FRAMEWORK

by

Xu Dong

Doctor of Philosophy

in

Applied Mathematics and Statistics

Quantitative Finance

Stony Brook University

2013

In this thesis we study the problem of modeling cross-sectional volatility structure of U.S. stock market. The thesis contains two parts. In the first part we identify a particular volatility spike phenomenon, that the cross-sectional volatility is significantly stronger at every 5 minute. An empirical spike study is conducted on individual stock level by applying Lee-Mykland jump detector. By constructing spike ratio statistics, the evolving paths of this spike phenomenon is studied during the 1993 to 2009 time period. In the second part, we model the volatility structure using factor structures. Particular, we propose two approaches to build a better model for small sample size problem, which is a common issue for high dimension models. We first study the fisher information matrix and derive Jeffrey's prior for the factor model. We extend the EM algorithm method for factor parameter estimation by applying the Jeffrey's prior, which turns to be more robust in the sense that risk would not be underestimated under small sample size. Next we extend the statistical factor model to be able to process empirical information.The fundamental(empirical) factor model and the statistical(hidden) factor model are widely used in factor analysis. The fundamental factor model is empirical, biased, and has small estimation variance. On the other hand, the statistical factor model is subjective, unbiased, but has a bigger estimation variance. In this

paper, a new factor model called EH Factor Model (Empirical-Factor–and–Hidden-Factor Factor Model) is introduced to combine the two models under a unified framework. The EH model allows to include exogenous information to help reducing the number of parameters, and meanwhile it maintains a statistical factor structure so that the idiosyncratic variance is kept. In this way, it reduces the estimation variance and the bias. We also compare the EH model with a widely used hybrid approach, which firstly apply the fundamental factor model and then treat the residuals with the statistical factor model. Compare with this approach, The EH model does not require to assume that the column space of statistical factor exposures is orthogonal to the fundamental factor exposures, and is also shown to be more robust in a sense of less estimation variance and bias.

Keywords: factor model; EH factor model; fundamental factors; statistical factors; EM algorithm

Contents

Acknowledgements

I need to first thank Prof. Andrew Mullhaupt, my academic advisor. He was always available and patient for any discussion not only limited to academic research and he was a belovedprofessor by all students at Stony Brook. His knowledge is not only deep but also broad, that his lectures covers from matrix analysis, signal processing and high dimensional statistics. I started to build my knowledge in maths from literally zero ground, and after five years, I am very skillful in several fields, such as matrix analysis and portfolio selection. These knowledge is comprehensive and powerful, but the most unique thing he taught is the perspective. With his over twenty years' experience as both portfolio manager and mathematician, he brought us a high level of how to apply structures in mathematics in understanding finance market. There is both structures for real world finance and structures for mathematics. Only by a applying a good view that one can use mathematics to kill finance problems. Under his guidance, students can build their view of the market as a high dimensional structure where good tools can be applied. Students can also learn things like how to use signal processing to enlarge the signal from a noisy background and how to use matrix analysis and optimization to build high speed portfolio selector. He was also generous and kind in opening doors for the students to the real Wall Street work and research. He is the absolutely the most cool professor in my life experience. I am thankful to all his help in answering my questions and feel fortunate that I had the opportunity to learn from him.

I want to thank Prof. Ann Tucker who gives me the chance to study in this program. Without her support I would never have the chance to become the first student of quantitative finance track with full scholarship. I remember clearly the first year studying with her that I learned very much about how to understand how macro-events impact on the market. Her view of how single event interact with market in short term and long term builds me an early insight which later grows mature under Prof Mullhaupt's signal processing view.

I want to thank Prof. Rachev for his enormous effort in directing such a wonderful program to attract so many prominent professors, and his friendliness and patience in communicating with us. I remember that he continuously support my research and give me huge encouragement when my mathematic maturity was at beginning level. I also learned a lot from him how to cooperate with other students and how to work as a group to overcome challenges using everyone's wisdom.

Other professors like Prof. Haipeng Xing, Prof. Alan Tucker, Prof Jiaqiao and Prof.

Jiao also gives me great help in the way of informative discussions. I also want to express my appreciation to many colleagues including (but not limited to) Xiaoping Zhou, Tengjie Jia, Rong Lin, Riyu Yu, Xiao Yu, Barret Shao, Yikang Chai, Tianyu Lu, Xiang Shi, Ning Ma and Ke Zhang, with whom I had chance to work together or discuss about problems and ideas during my research. Also, the working stuff of AMS department is of huge importance and helpful to me. I want to thank Christine, Laurie and Janice, they and the professors are equally important fundamentals for our department.

I would like to thank the support from my family members, Zhaojun Dong, Changqing Chen, Renxin Chen, Huilan Liu, Xiaobo Chen, Mingyang Chen, Xuejiao Wang, Qiuping Zhen, Kang Li, Changyin Chen, and Luyi Dong, without whom I would not succeed in accomplishments.

Last but the most important, I want to thank Cecilia Du, thank you so much for your accompany all the way.

Vita, Publications and/or Fields of Study

I was born in Chongqing, P.R. China in 1984. I received my B.S in Biological Science from Xiamen University in 2007.

I was admitted to the PhD program in Applied Mathematics and Statistics at Stony Brook University with full tuition scholarship in 2009. My concentration is quantitative finance and my academic advisor is Professor Andrew Mullhaupt. My research topics in PhD include factor model, portfolio optimization and covariance estimation. The related research work include three working papers:

1. How to use empirical information in hidden factor model? A factor model that combines fundamental factors and statistical factors.

2. Cross-sectional spike volatility phenomenon in U.S. stock market.

3. Small eigenvalue problem: the Jeffrey's prior for EM algorithm on statistical factor model.

1 Introduction

Dimensionality is an important feature of financial data. Under observation of single time series with dimension $1 \times n$, much attention is paid on modeling and prediction through a broad class of models such as ARMA-GARCH (Engle, 1982; Tsay 2002) and time-space models where least-squares type estimation methods are popular (Aoki, 1990; Geyer et al, 1999), as well as analysis of the tail events, through heavy-tailed distributions (Kelker, 1970) and stable distribution (Rachev et al, 2011). On the other hand, with high dimension data $d \times n$, the analysis can also be conducted in a cross-sectional way. The cross-sectional study refers to a class of research methods that involve observation of a population at one specific point in time. It treats data across different instruments at the same time points as of from the same category. The purpose of cross-sectional analysis often is associated with describing the underneath property of the market and to answer the question that how it behaves during certain time period. Causality and correlation analysis are also conducted with lagged cross-sectional data. The materials and targets for cross-sectional analysis in finance is in general very broad. Merton (1987) reports a positive correlation between idiosyncratic risk and expected return. The type of research similar to Merton, where data is firstly aggregated and few statistics per timestamp is draw, can be generally concluded as exploratory data analysis method. The merit is that since few aggregated statistics are calculated based on usually huge data set, it gives a better signal-to-noise ratio through the large sample property. Another example of this type of cross-sectional research is the periodic phenomenon study of the market (e.g. economic periodic bubble decay model (Zhou et al, 2003). Another branch of cross-sectional analysis tool is based on describing the market by a large set of parameters. Statistical factor model and time series factor model are two examples. The factor loadings are estimated through factor time series or a bunch of observations across different instruments, where cross-sectional factor loadings represent a status of the market. For two different time point, if the factor loadings vary, it means the market has a different covariance structure. While abundant research has examined the time-series relation between volatility of the market and the expected return (Campell and Hentschel, 1992; Glosten et al, 1993), the relationship between cross-sectional volatility (sometimes reffered to as aggregate volatility) and return has received less attention. Andrew Ang et al argue that if the volatility of the market is a systematic risk factor, the arbitrage pricing theory or the factor model predicts that aggregate volatility

should be priced in the cross-section of stocks and therefore stocks with different sensitivities to innovations in aggregate volatility should have different expected returns (2006). Further, they find that stocks with high sensitivities to innovations in aggregate volatility have low average returns.

While using more parameters to describe the market is providing a more detailed measure of the market, the shortcoming on the other hand is that it requires more samples for estimation. Usually the estimation variance is linearly correlated with the inverse of sample size (Aggarwal, 2005). In this thesis, we try to model the market volatility through both low dimension parameterization methods and high dimension parameterization methods.

First we study the periodic phenomenon of the market using low dimension parameters. We treat the cross-sectional returns at every minute as homogenous samples. Further, we aggregate the minute data every other hour, so that we can amplify the signal-to-noise ratio to the largest extend for periodic phenomenon on one hour level.

Among visual patterns evidence in financial time series are periodic patterns exhibited at different time scales. Significant daily patterns can be observed in return volatility, bid-ask spread, and trading volume [5]. Early studies about intraday patterns were based on daily and weekly data (for example French [22], Gibbons and Hess [28], and Keim and Stambaugh [37]). All three studies find that the average market close-to-close return on the New York Stock Exchange (NYSE) was significantly negative on Mondays and significantly positive on Fridays.

With the advent of high-frequency data, which has been recorded since the 1980s, it became possible to explore these patterns in a fine-grained way. Wood *et al.* examined minute-by-minute returns data for a large sample of stocks traded on the NYSE [52]. They found significantly positive returns on average during the opening 30 minutes of trading hours and the last 30 minutes prior to market close. This observation was echoed by Ding and Lau [14] who analyzed market activity with regard to 200 stocks on the Singapore Stock Exchange.

Trading volumes also show evidence of periodicity in intraday data. Specifically, a number of important studies were done by Wood, McInish and Ord [52], Jain and Joh [34], Foster and Viswanathan [21], McInish and Wood [49], and Gerety and Mulherin(1992 [27]), who employed hourly aggregated volume (measured as the number of shares traded) for all NYSE and NASDAQ stocks and observed the intraday trading volumes forming a U-shaped

pattern. In addition, Foster and Viswanathan [19] observed a similar pattern in regard to volume data for individual NYSE stocks. Darrat et al [12] examined the Dow Jones index stocks for 2003, based on which they reported a significant lead-lag relationship between volume and volatility.

Much of the research on high-frequency data focuses on patterns of US stock markets; besides, some research has been conducted on other markets, and this has reported similar patterns. Intraday effects have been documented for the London Stock Exchange (LSE) [54] [38] [1] [47] [17] [18], the Hong Kong Stock Exchange [31] [9], the Tokyo Stock Exchange [30], and the Toronto Stock Exchange [49]. Among previous studies of stocks traded in foreign exchanges, a few studies used large data sample, which included most of trading records on the whole stock exchange. One of these studies is the project carried by the University of Leeds' International Institute of Banking and Financial Services [50] which included a number of exploratory objectives, and documented a wide range of stock market characteristics. Their study used data sampled at a one-minute frequency, and drew on a total of 25 million observations. According to this study, there are major institutional differences between the UK and US stock markets, which could be why the intraday patterns of each are different. For example, their study showed a higher proportion of institutional investors in the UK than in the US [36]. In addition, the mixture of order-driven and quote-driven systems in the UK, introduced in October 1997, affected trading behaviors of this market [5]. Based on samples from a trading time period of 8:30 a.m. to 16:30 p.m. GMT, Abhyankar et al [1] found that the trading volume for the LSE had a two-humped pattern instead of a U-shaped pattern, with highs at 09:00 a.m. and 15:00 p.m.

There are also some important findings in regard to other foreign exchange markets. Although no U-shaped pattern has been observed in the dollar-yen or euro-dollar market in the New York market, it is reported that overlapping business hours enhanced the cross-region transaction activity for the overlapping time period. This perspective could be useful in efforts to explore stock trading behavior as it is relating the human behavior to trading hours [32]. Furthermore, the one-hour shift between regular time and daylight savings time in the US leads to a corresponding shift in the overlapping trading time, which would make trading behaviors during this time of particular interest.

The time-dependent nature of the market's microstructure (e.g. orders submitted at mid-day are executed more slowly than orders submitted around open and close) has been

considered a reason for time-dependent intraday patterns [23]. In addition, the fact that informed traders concentrate their trades at open and close of trading sessions offers an alternative explanation for the U-shape pattern [35]. Atkin and Basu examined public announcements after trading hours using 400 NYSE stocks and found that public information had a significant effect on the U-shaped pattern of volume. From this research, it was theorized that the large volume of trades at the beginning of the day could be the result of the aggregate amount of new information that becomes known overnight.

Inspired by the previous research, we investigate the minute data for finer periodic phenomenon of the market. We find a five minute repetition of spikes both exists for cross-sectional returns and cross-sectional volume. This is the first fine-resolution analysis of market performance that we are aware of, except for a blog entry from a Taiwan Index Futures (TIF) trader who observed a similar 5-minute pattern for the TIF. Based on this finding, we try to identify this phenomenon on individual stock level by applying jump detection method. As cross-sectional phenomenon we further categorize different spike patterns by clustering method and study how it evolves through 1993 to 2009 time period.

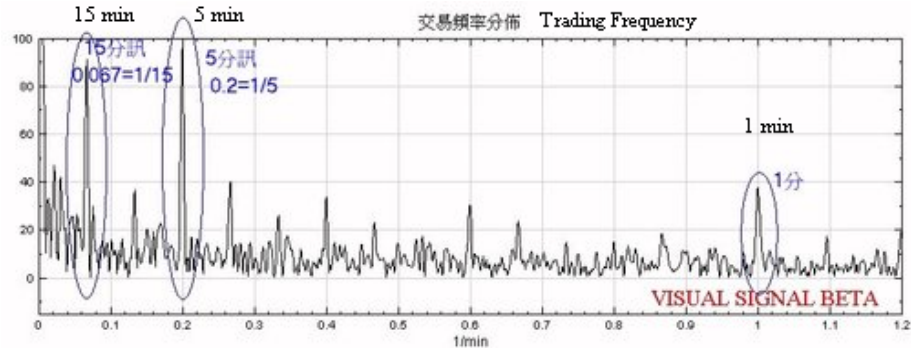


Figure 1. The Taiwan TIF Minute Pattern

Second we use high dimensional parametrization method to study the risk structure of the market. First we analyze the fisher information, which gives an insight in understanding the

0

References

[1] http://tw.myblog.yahoo.com/futurex168/article?mid=78&page=1#1370

estimation variance since the inverse of fisher information is the Cramer-Rao Lower Bound of the estimated parameters. Then we propose a method for improving the factor model under the small sample size by introducing Jeffrey's Prior. The Jeffrey's prior works as a non-informative prior, and brings robustness at a small sample size situation. Based on the Jeffrey's prior, we modify the EM algorithm and it shows that when the sample size is small, it gives a penalty on for betting on the small eigenvalues. Alternatively, we construct an EH (empirical factor and hidden factor) model, which structures the factor loading matrix to include empirical analysis information, and does one step maximum likelihood estimation based on EM framework. The simulation shows EH model has a closer distance to real covariance when the samples is small compared with the statistical factor model.

2 Low Dimension Models and the Cross-sectional Volatility

In finance, high quality data is essential for any meaningful analysis. New York Stock Exchange has been recording high frequency data since 1993 of all the trades from multiple exchange, aggregating in second level. However, previous research on the U.S. stock market in our literature review from 1973 to present does not offer fine-grained exploratory analysis; i.e., it does not include any patterns at the minute-to-minute level. In our view, researchers have refrained from conducting this kind of analysis for one of two reasons: first, even as recently as 2005, it was by no means common for researchers to have a powerful computing system capable of scanning through all the TAQ data from 1993 to the present, which totals approximately 5TB; Second, the high-frequency TAQ data were available only in an unorganized format, so that extracting all the data required copying out approximately 1,000 CD/DVDs manually, and therefore making the extraction of the data is nontrivial because of the longtitude of time period and the multiple times change of data format without publishing notice. The changing formats of the TAQ data (three different formats which has undergone changes from time to time) and data structure chosen for storing this huge data set added extra difficulties for researchers to work with the TAQ data and arrange them in a manner canducive to studying. In sum, a significant time investment was necessary to organize the information before it could be put to any productive research use.

On the other hand, the fact that high-frequency trading comprised more than 70% of the total daily volume of the US stock market up to 2009 [15] made the idea of exploring the high-frequency dynamics at a finer resolution interesting and challenging–both in terms of academic theory and practical applicability.

Extracting the data from TAQ raw data CD/DVDs is nontrivial largely due to the highly chaotic data organization. The TAQ data format consist of three main versions: TAQ1, TAQ2 and TAQ3. The versions differ only slightly but it is important that they be carefully aligned with each other. The TAQ trade records for each day are captured in two files: the index file and the binary file. In binary files, all trades for one ticker are stored sequentially in a data block. The location for each data block is indexed as the start position and the end position, which is recorded in the index file. With the position information from the index file, the extraction process can point to the the binary files and extract trading time, trading price, shares per trade, and other useful information. Therefore the basic routine for extracting data is a two-step process: first, the recorded beginning position and record ending position for each symbol is located from the index file; second, this information is used to extract the price, trading size, and trading time from the binary file.

Field Name	**Layout**	**Description**
SYMBOL	Character 10 bytes	Stock symbol
TDATE	Binary 4 bytes	Transaction date. Format: yyyymmdd
BEGREC	Binary 4 bytes	Start position
ENDREC	Binary 4 bytes	End position

Table 1. CT Index File (*.IDX)

Field Name	**Layout**	**Description**
TTIM	Binary 4 bytes	Trade time
PRICE	Binary 4 bytes	Actual trade price per share.
SIZE	Binary 4 bytes	Number of shares traded.
ENDREC	Binary 4 bytes	End position

Table 2. CT Binary File (*.BIN)

Our extraction process is highly dependent on the memory-mapped file technique. The memory mapping technique greatly reduces the amount of memory necessary to perform

I/O operations. For each day, the binary file can have a size in hundreds of megabytes, This means significant time to be spent engaged in reading and writing processes if the extraction follows the regular I/O operation, because usual I/O operation requires moving data into RAM before doing any calculation needed for the extraction. The memory mapping technique enables the extraction process to avoid copying the whole file into the memory. Instead it uses only a small size of RAM to create pages to locate a small part of the data. After reading the small size of the data into RAM, the extraction process starts to compute with the data available, and meanwhile the memory mapping technique starts to load the next block of the binary file. In this way, the program maintains computing and loading at the same time and therefore becomes much more efficient. In addition, this technique is operated under the help of the virtual memory manager, which is highly optimized for Linux system, and therefore ensures the efficiency of the technique.

The algorithm for extracting the minute data is designed so that for every minute the last trade information is retained and the interval is omitted. In extracting the minute volume information, we sum up all the prices multiplied by shares during one minute and thereby obtain the total amount of money traded during any given minute.

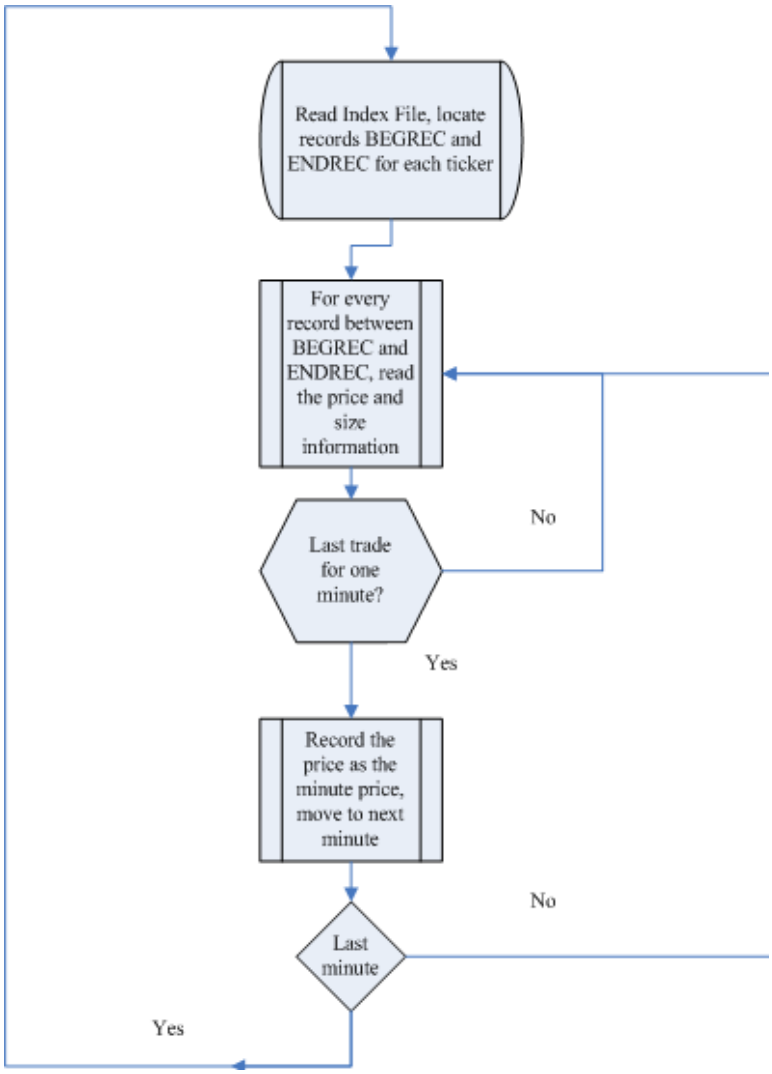


Figure 2. The Algorithm for the Data Extraction.

Based on the tick data from the TAQ data, we first conduct an exploratory data analysis with the initial purpose of identifying patterns in the price process. We start the analysis by considering the most important factors: returns, volatility, and volume. In the beginning of this section, we use the absolute log return in our analysis. We make this decision because (1) the correlation relationship of the absolute return and volume have been studied through [29], and (2) volatility can be measured either by the squared return or by absolute value of the return [8], which we will review in the next sub-section.

2.1 Advantage of Using Cross-Sectional Volatility

The concept of realized volatility (historical volatility) can be traced back as far as in the work by Taylor [48]. It can be seen in some early high-frequency studies such as that by Muller et al. [44]. The concept, however, has been drawn on more frequently since Andersen et al. [3], which showed its power in process modeling.

The realized volatility $v(t_i)$ at time t_i is defined as

$$v(t_i) = v(\Delta t, n, p; t_i) = [\frac{1}{n}\sum_{j=1}^{n} |r(\Delta t; t_{i-n+j})|^p]^{\frac{1}{p}}$$

where r is the regularly spaced return, n is the number of return observations, Δt is the observation time interval, and the size of the total sample is $n\Delta t$. The exponent p is often set to 2.

Notably, the absolute return, which is the realized volatility when p is set to 1, is one of the alternative measures of dispersion [26]. It has some important properties in studying periodic patterns. When there exists a periodic pattern in a time series, the autocorrelation coefficients are significantly higher for time lags that are integer multiples of the periodic interval than for other lags. The general framework for determining seasonal volatility comprises several steps [24]. First, a grid of observation intervals are chosen, with consideration of possible sources of bias, such as five different days of a complete week, business holidays, daylight savings time, days when open or close time of the market changes, etc. For the present research, we use the trading days as the time frame.

A number of papers have proposed different approaches for addressing volatility periodicity. A few of these papers are based on factoring volatility into an essentially deterministic seasonal part and stochastic part, in which the seasonal part is modeled by a set of smooth functions [2]. Gencay et al., for example, used the wavelet multi-resolution method in their study of volatility periodicity [25].

In our research, we use cross-sectional volatility, defined as

$$v^c(t) = std\begin{pmatrix} r_{1,t} \\ \dots \\ r_{d,t} \end{pmatrix}$$

where $r_{i,t}$ is the return for the i-th instrument at time t. In this way, we treat the different instruments at the same time point as samples from the same group and use $v^c(t)$ as one parameter statistics to describe the second moment behaviors of market at time t.

Assuming there are d instruments, each following Gaussian $(0, \sigma_i^2)$ distribution. Denote the total observations in the time interval $[t+1, t+T]$ from d instruments as N, so that for each instrument, there are N/d observations. $x_{i,t}$ is a sample from i-th instrument on time t. Then the variance from maximum likelihood for σ_i^2 is $\hat{\sigma}_i^2 = \frac{\sum_{j=1}^{N/d} x_{i,t+j}^2}{d}$. When calculating the cross-sectional variance, it is treating different instruments as of the same distribution $N(0, \sigma^2)$, and the estimated variance $\hat{\sigma}^2 = \frac{\sum_{i=1}^{d}\sum_{j=1}^{N/d} x_{i,t+j}^2}{N} = \frac{\sum_{i=1}^{d} \hat{\sigma}_i^2}{d}$. For Gaussian distribution $N(0, \sigma^2)$ with N samples, the Cramer-Rao Lower Bound is $\frac{\sigma^2}{2N}$(Cramer, 1946; Rao, 1945). The Cramer-Rao Lower Bound for the individual instrument variance is similarly $\frac{\sigma_i^2}{2(N/d)}$, which means the variance of estimation of cross-sectional volatility is $1/d$ of the individual instrument.

2.2 Calculation of Cross-Sectional Volatility

First, we examine the intraday pattern of all stocks in Russell 3000. We extract every minute's trade price information for every stock from January 3, 2005, to December 28, 2009, excluding the ones evincing price discontinuity and form the n by 391 matrix, where n is the total number of stocks traded multiplied by the total trading days, which vary from year to year. Next, we consider every column of the matrix as comprising all the information for one minute. We draw graphs in order to display general information for every minute's absolute log return information. The graphs in Figure 3 are as follows (from the top one, subgraph 1, to the bottom, subgraph 5): subgraph 1 shows 10%, 25%, 50% and 75% percent level of the absolute log return; subgraph 2 shows a histogram representing the distribution of the absolute log return; subgraph 3 shows the local maximum of the 25% quantile; subgraph 4 shows the standard deviation of the absolute log return together with the changing point; and subgraph 5 shows a spectrum view for the absolute log return. In comparing the quantilegraph with the standard deviation graph, we can see the U-shape pattern identified in previous research. In addition, in subgraph 2 a fat tail is evident for the absolute log return. In subgraph 3 and subgraph 4, we use a local maximum detector to find the 15 points that are most different from their two neighboring points.

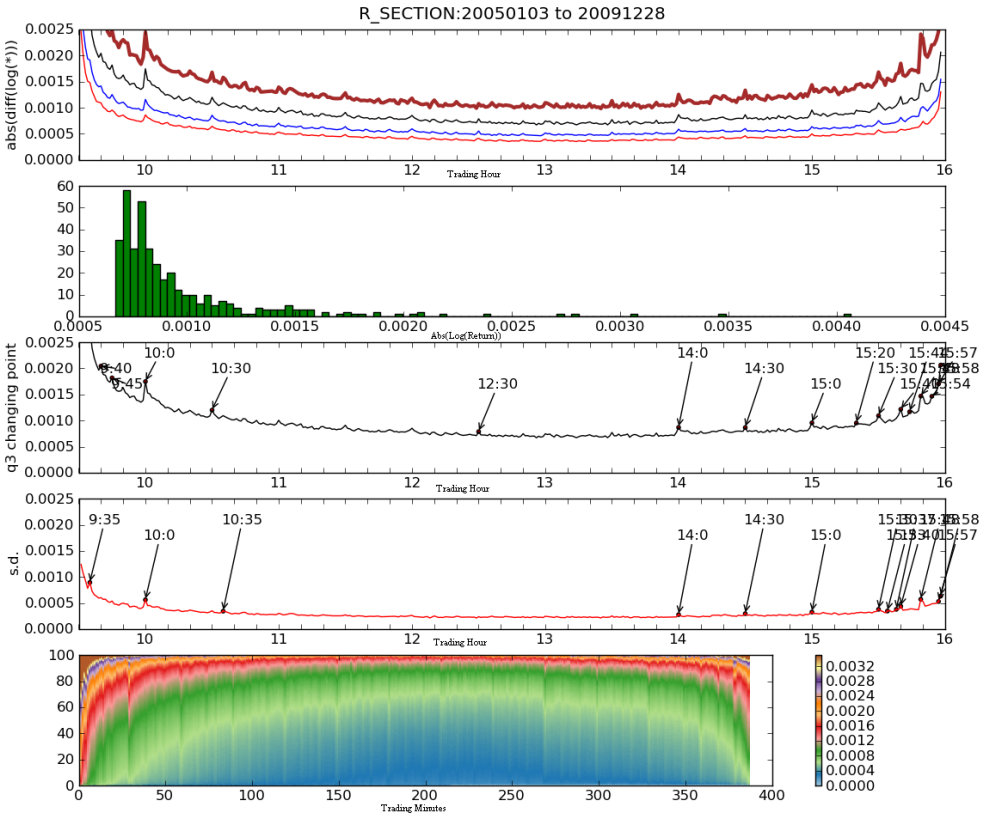


Figure 3. The Russell 3000 Index Information

It is noteworthy that the volatility and the absolute log price increase simultaneously at some time points. The top 25% of the absolute log price as shown in subgraph 3 reaches its local maximum at 10:00 a.m., 14:00 p.m., 14:30 p.m, and 15:00 p.m. The local maximums for the absolute log price are at 10:30 a.m., 15:20 p.m., 15:30 p.m., and 15:40 p.m.. In addition, the local maximums for volatility are at 10:35 a.m., 15:15 a.m. and 15:40 a.m. Therefore, we suspect a repeated local maximum pattern exists at a fixed time interval on both the absolute log returns and the volatility.

In order to improve our knowledge in regard to the details of the volatility curve, we fit the data with the rational function whereby

$$1/y = \frac{ax + b}{cx^2 + dx + e}$$

After fitting this function with the least-square method, we have a fitted inverse series. By doing so we remove the general curve. Next we reset the observation window in order to condense the data. This is necessary because on the 390-minute observation window it is hard to identify trends within the data. As we assume that there is some pattern based on 5-minute periods, our objectivity in the observation is not compromised by using an observation window of 60 minutes. The benefit of using the 60-minute window is that the 5-minute pattern, if indeed there is such an effect, would be easier to observe over this time period. After scaling with the rational function and using the 60-minute observation window, we calculate the "neighbor difference value" d_i for every minute in the 1 to 60 minute range by letting $d_i = |y_i - y_{i-1}||y_i - y_{i+1}|$, where integer $i \in (1, 60)$. If the 5-minute pattern does exist, there would be 12 local maximum points as $[5, ..., 55, 60]$. Therefore, we sort all the d_i and mark out the top points with the highest d_i values, and the result is that we find 14 local maximum points to be $[1, 5, 10, 12, 15, 20, 22, 25, 30, 35, 40, 45, 50, 60]$, during which only the 55 minute data point is not on the list of the local maximum points. In conclusion, the exploratory data analysis here validates our assumption that there is a 5-minute-spike pattern in the intraday pattern. In later discussion, for simplicity, we call the 5-minute-spike pattern as "the 5-minute pattern".

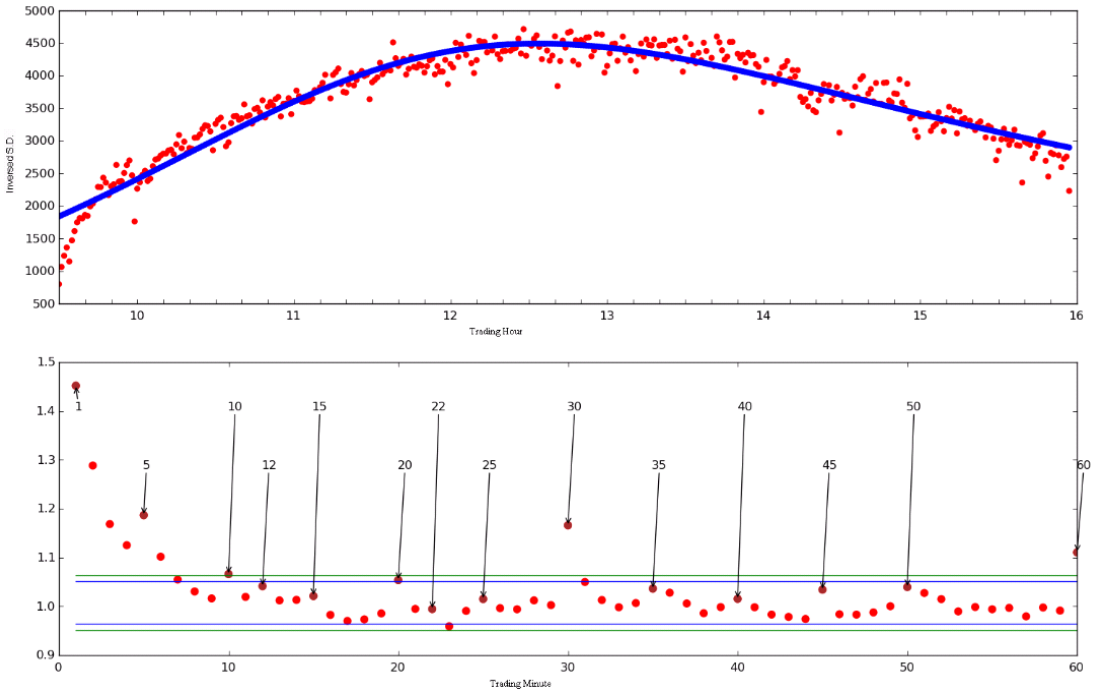


Figure 4. The Scaled Volatility

We want to know the starting time for this pattern, and we also want to determine whether it is consistent at different time periods of the day; therefore, we draw graphs for the individual hours between 10:00 a.m. and 15:00 p.m. We exclude the first half hour of trading, as it has a poor fit with our model, as Figure 4 shows.

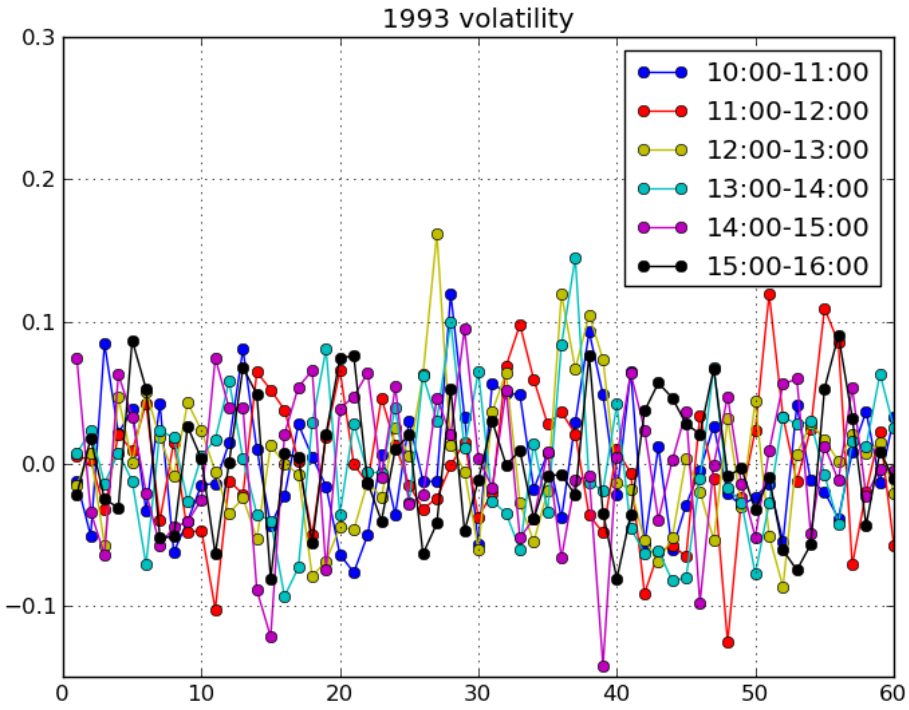


Figure 5. 1993 Volatility

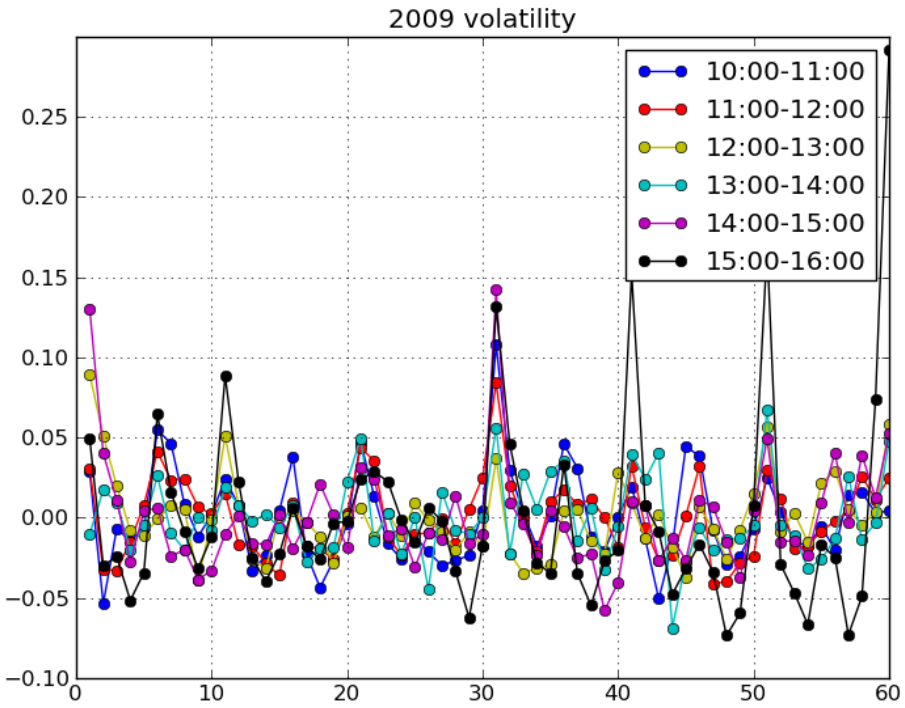


Figure 7. 2009 Volatility

From the volatility graphs for each year, it is evident that this pattern is not observable before year 2000. In contrast, in 2009 the volatility spikes that occur at five-minute intervals are easy to identify; Figure 7 shows spikes at every five minutes throughout the entire trading hour period. For example, looking at the data point on the 31 minute point from the hourly observation window, from the top to bottom, the volatility levels are all larger than the two neighboring points–the 32 minute point and the 30 minute point. This volatility spike is repeated again 5 minutes later, on the 36 minute point, and then again on the 40 minute point, etc.

2.3 Cross-sectional Dollar-Volume Study

Information about volume is also necessary for understanding the intraday dynamics of the stock market. In the previous section, we presented a minute volatility graph. However, for our calculation of minute volatility, we used only the price from the last trade within

every minute. The information during that minute was simply omitted. Therefore, we want to utilize a statistic that comprises information for every trade that take place during the period of any given minute. Moreover, in explaining the formation for a new price, the number of shares traded at previous price is also an important factor. Therefore, it is necessary to study the patterns of dollar-volume, which is a time series formed by the price multiplied with the shares, in order to determine if the total money traded in any given minute creates a notable pattern.

We calculate the money traded for every minute, by multiplying every trade price with every trade size and adding all of the products up; for each stock. Then in a way that is similar to our scaling of the volatility data, we scale this data to fit a rational function in accordance with a 60-minute observation window. We find a similar 5-minute pattern in this context. It should be noted here that the 5-minute pattern for the total money traded, unlike that for the volatility graph, is more of a wave shape. Besides the 5-minute pattern, we can clearly identify a 30-minute pattern, which comes much larger in the magnitude than the 5-minute peaks.

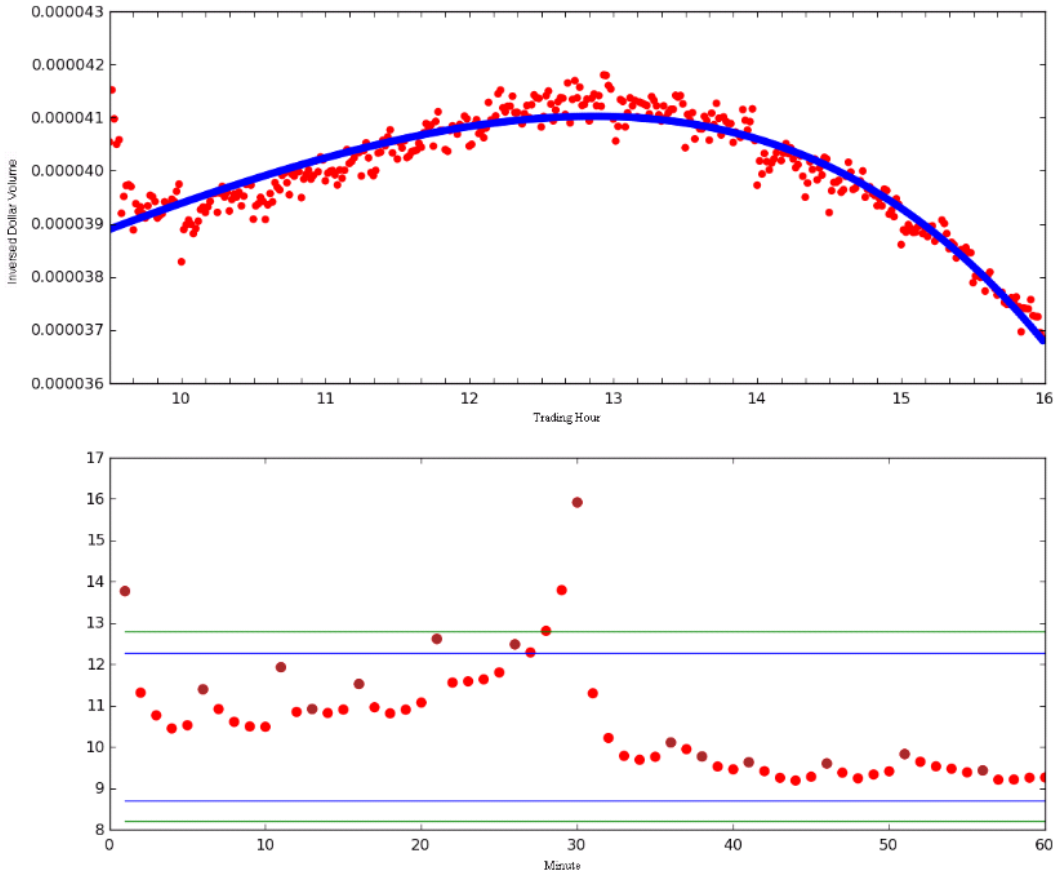


Figure 5. The Scaled Volume

In addition to observing a fixed 60-minute observation window for the overall 2009 dollar-volume data, we conduct a further exploratory data analysis using different observation windows lasting 5 minutes, 15 minutes, 60 minutes, and 120 minutes together with datasets of 1 day, 30 days, 250 days, etc. We perform a series of analysis whereby we either exclude or include the opening and closing 30 minutes of the trading time for each day as well as exclude or include opening or closing 60 minutes of each day's trading time. We also compared 2009 graph to similar data plots from different years (2005 - 2008), the conclusion is similar and therefore we only list two figures below. Based on this set of analysis, we are able to offer evidence in support of the idea that different trading patterns have different 5-minute pattern shapes. Our conclusions are as follows:

1. The peak at 30 minutes is not due to an opening or closing effect. We exclude

the opening and closing time by only considering 10:00 a.m.–3:30 p.m. period, and the 30-minute peak is still apparent.

2. Peaks occur every 5 minutes, but their strength varies. Using a 10-minute observation window, comparing a graph starting at 10:00 a.m. with a graph starting at 9:55 a.m., we can see that the peak in the latter graph has shifted from the beginning to the middle of the observation window. This indicates that the 5-minute peaks vary in strength. In addition, we should note that varying strength of the peaks does follow a pattern; that is, a weaker 5-minute peak comes after a stronger 5-minute peak . We also test to determine the spikes at every 10-minutes do not always have the same strength.

3. Despite the observation that the 5-minute peaks vary in strength, we observe that the peaks at the 30-minute intervals all share a similar strength level.

4. The 30-minute effect is observable consistently throughout the entire day based on our results for the 360-minute observation window.

The conclusions above are also consistent with the volatility graphs, which are based on the same test principles.

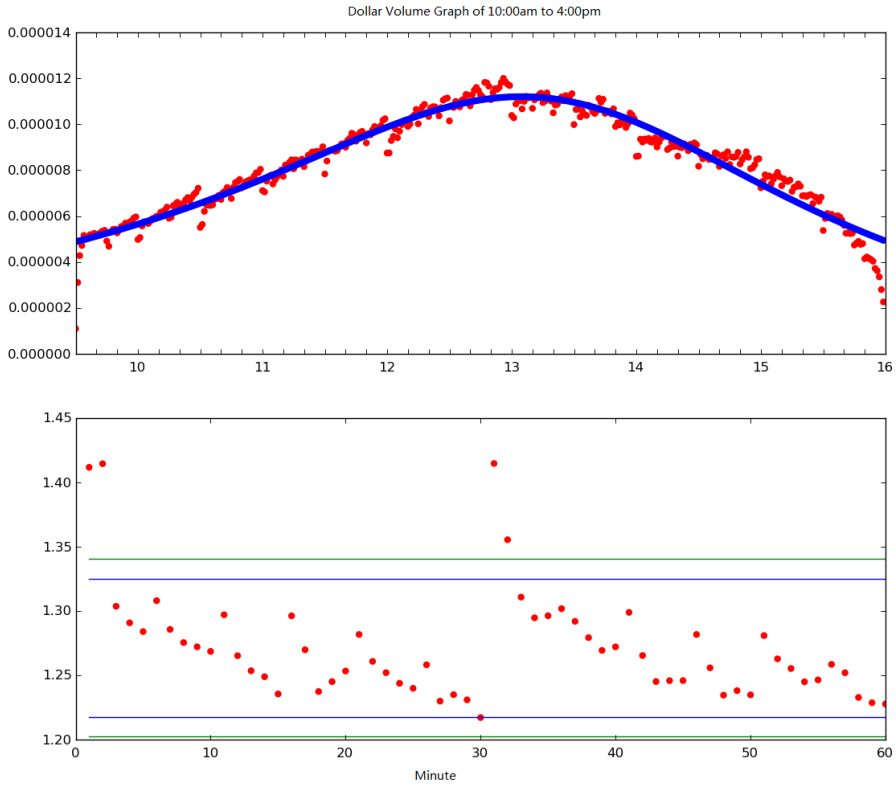


Figure 6. Dollar-Volume Graph for 10:00 a.m. to 3:30 p.m.

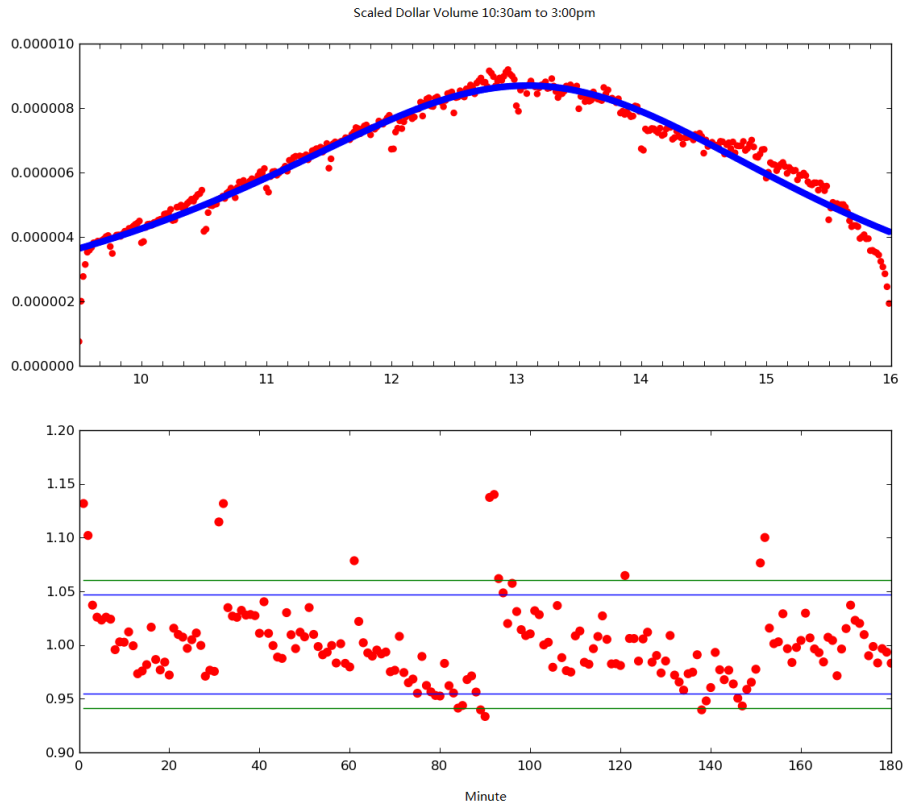


Figure 7. Dollar-Volume Graph with a 180-Minute Observation Window

2.4 Spike Phenomenon in Different Markets

We are interested in determining if there are differences in the way the dollar-volume behaves in the NYSE and the NASDAQ (NASD) markets. Figure 8 shows that the markets have similar general patterns. To compare the difference between dollar-volume patterns and volatility patterns, we add the volatility curve. We aslo achieve a more comprehensive view of the three curves by scaling them according to the individual sample mean value of each. Based on this scaling, Figure 8 shows the scaled dollar–volume and the volatility graph for the year 2009. It is interesting to see that though the dollar-volume and volatility patterns both have a general U shape, they are different at either end: the volatility curve is high at the beginning and low at the end whereas the dollar-volume graph is high at the end and

low at the beginning.

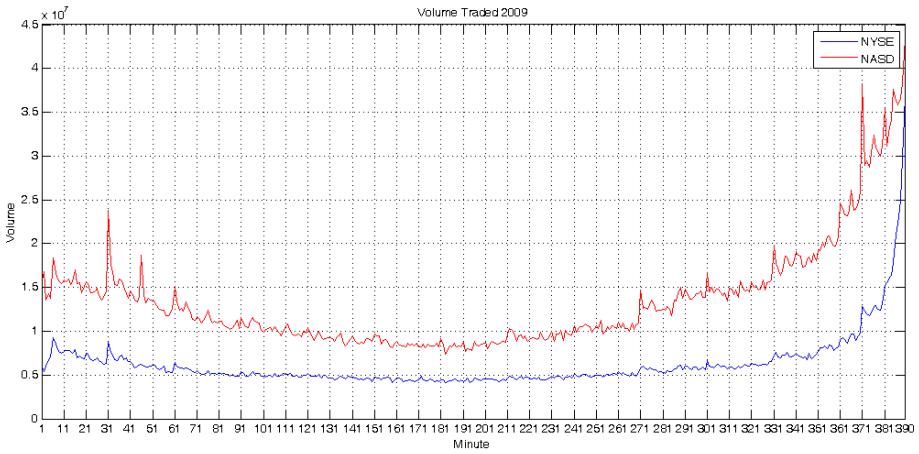


Figure 8. The NYSE and the NASD Dollar-Volume Graph (2009)

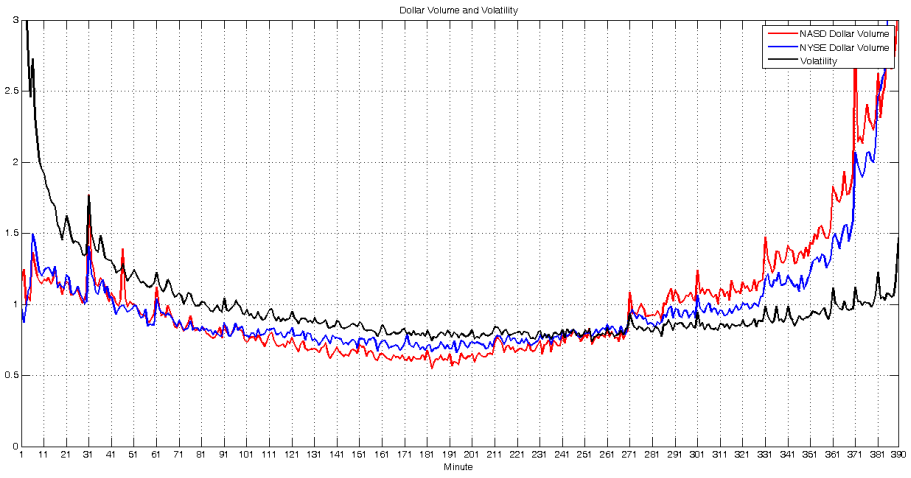


Figure 9. The Scaled Dollar-Volume and Volatility Graph(2009)

After fitting and adapting the 60-minute observation window, we are able to see the 5-minute patterns quite consistently in both curves, despite the differences at the end between the dollar-volume curves and the volatility curve, as Figure 9 shows.

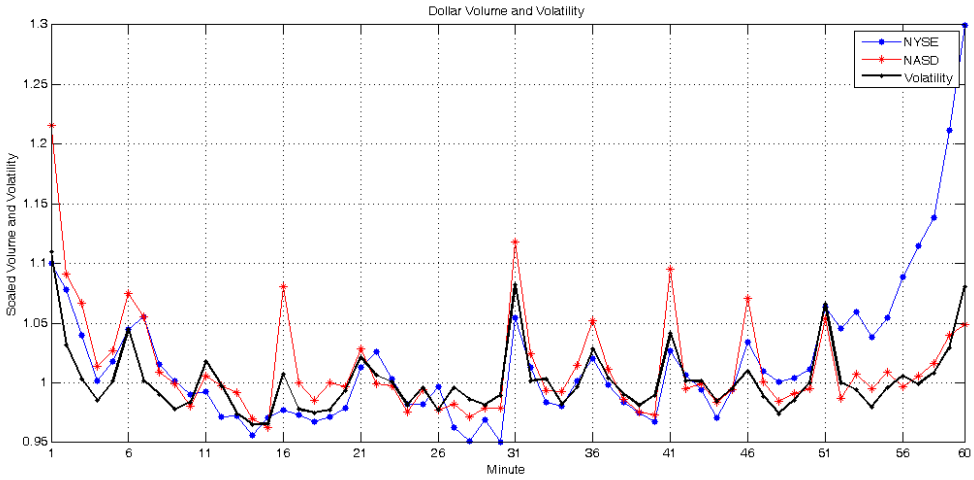


Figure 10. The 5-Minute Patterns in the Dollar-Volume and Volatility Graph (2009)

As a conclusion for the previous three subsections, we identify a 5-min periodic phenomenon in both cross-sectional volatility and volume sense, and this phenomenon progressively becomes more observable in time.

2.5 Jump Diffusion Detector Approach on Individual Stock Level

2.5.1 Review of Lee-Mykland Jump Detector

The 5-minute effect based on a large data sample consisting of thousands of stocks is clearly presented in the previous section; yet we have not found a similar effect for individual stock. The question is whether this effect simply does not exist for individual stocks, or it is difficult to identify because of an insufficient number of samples. In order to answer this question, we turn to the jump diffusion theory, which is well suited to identifying small jumps in data. Specifically, we use the statistical jump detector in order to detect discontinuities in the time series of prices. In regard to the discontinuities of the financial markets, researchers have found jumps to be difficult to identify empirically because only discrete data are available from continuous-time models and the jump detector is designed to identify jumps that are associated with unstable volatility. We apply the Lee-Mykland nonparametric detector using 1-second high-frequency data from the U.S. equity market [40]. Our test construction, which is described in the next section, is based on Lee-Mykland's work.

We follow Lee and Mykland's paper to construct our test:

step1: Make the basic assumption(s)

step2: Construct statistics L

step3: Construct estimation $\widehat{L}$ for L

step4: Find asymptotic distribution for $f(\widehat{L})$, and construct the confidence interval value as β

step5: Perform the test, and check the probability of misclassification

We consider a one-dimensional asset return process with a fixed complete probability space (Ω, F_t, P). The continuously compounded return is written as $d\log S(t)$.

When there are no jumps in the market, S(t) is represented as

$$d\log S(t) = \mu(t)\,dt + \sigma(t)dW(t)$$

When there are jumps, S(t) is given by

$$d\log S(t) = \mu(t)dt + \sigma(t)dW(t) + Y(t)dJ(t)$$

where J(t) is a counting process independent of $W(t)$ while $Y(t)$ is the jump size.

Discrete observations are made based on the return process. We assume there are n observations of log price information during a [0,T] time period. For simplicity, we assume that observation times are equally spaced: $\Delta t = t_i - t_{i-1} = T/n$.

Assumption. for any $\varepsilon > 0$,

$$\begin{aligned}
\sup_i \sup_{t_i \leqslant u \leqslant t_{i+1}} |\mu(u) - \mu(t_i)| &= O_p\left(\Delta t^{1/2-\in}\right) \\
\sup_i \sup_{t_i \leqslant u \leqslant t_{i+1}} |\sigma(u) - \sigma(t_i)| &= O_p\left(\Delta t^{1/2-\in}\right)
\end{aligned}$$

This assumption ensures that the drift and diffusion coefficients are not changing dramatically over a short time interval. Therefore, we can replace $\mu(u)$ and $\sigma(u)$ with $\mu(t_i)$ and $\sigma(t_i)$ respectively with some restrictions.

The statistic $\mathcal{L}$ is defined to detect the jump at time t_i. We want to test that if the realized asset return at that time is much greater than the usual continuous innovations.

In order to determine so, we test a jump on t_i by using the realized log return over an instantaneous volatility, which is formularized as

$$\mathcal{L}(i) \triangleq \frac{(\log S(t_i) - \log S(t_{i-1}))}{\widehat{\sigma\,(t_i)}}$$

The reason for dividing by the instantaneous volatility is to avoid the impact of the volatility change. To estimate the instantaneous volatility, we use the previous K-1 data window, which contained a Δt interval, by

$$\widehat{\sigma\,(t_i)} \triangleq \frac{\sum_{j=i-K+2}^{i-1} |\log S(t_i) - \log S(t_{i-1})||\log S(t_{i-1}) - \log S(t_{i-2})|}{K-2}$$

In our study, the drift (of order dt) is negligible compared to the diffusion (of order $\sqrt{dt}$).

Select a window size $K = Op(\Delta t^{\alpha})$, where $-1 < \alpha < -0.5$.

For window i, if there is no jump in $(t_{i-1}, t_i]$, then as $\Delta t \to 0$,

$$|\mathcal{L}(i) - \widehat{\mathcal{L}(i)}| = Op(\Delta t^{\frac{3}{2}-\delta+\alpha-\varepsilon}),$$

where $0 < \delta < \frac{3}{2} + \alpha$ and

$$\widehat{L(i)} = \frac{U_i}{c}$$

where $U_i = \frac{1}{\sqrt{\Delta t}}(W_{t_i} - W_t) \backsim N(0,1)$ and $c = E|U_i| = \sqrt{2}/\sqrt{\pi}$

The above statements indicates that by using $\widehat{\mathcal{L}(i)}$, we can approximate the $\mathcal{L}(i)$'s asymptotical distribution with mean 0 and variance $\frac{1}{c^2}$.

The following statements show how the jump test reacts to the arrival of jumps: when Δt goes to 0, the test statistic becomes so large that we are able to detect the jump arrival at time $t_{i.}$

If there is a jump at any τ time in $(t_{i-1}, t_i]$, then

$$L\,(i) \approx \frac{U_i}{c} + \frac{Y(\tau)}{c\sigma\sqrt{\Delta t}}$$

where $Y(\tau)$ is the actual jump size at actual jump time τ. Therefore, $L(i) \to \infty$, as $\Delta t \to 0$, and $U_i = \frac{1}{\sqrt{\Delta t}}(W_{t_i} - W_t) \backsim N(0,1)$ and $c = E|U_i| = \sqrt{2}/\sqrt{\pi}$.

Next, we need to determine a proper window size K. The window size K must be large enough for the effect of jumps on estimating instantaneous volatility to disappear, but it must also be smaller than the number of observations n. Because $K = Op(\Delta t^{\alpha})$, with $-1 < \alpha < -0.5$, therefore $\sqrt{days \times nobs} < K < days \times nobs$ would fit, where nobs is the observations per day. In the original papers, optimal K for one-hour, 30-minute, 15-minute and 5-minute data was suggested as 78,110,156 and 270, respectively.

When there is a jump, the test statistic tends to be infinity as Δt goes to 0, therefore is generally larger than when there is no jump. In order to determine a rejection region in accord with the hypothesis that there is no jump, the key question is how large the test statistic will be when there is no jump. The study of the asymptotic distribution of maximums of the test $\mathcal{L}$(i) under the absence of jumps at any time in $(t_{i-1}, t_i]$ lead to that, as $\Delta t \to 0$,

$$\frac{\max |\mathcal{L}(i)| - C}{S_n} \to \xi$$

where ξ follows that $P(\xi \leq x) = \exp(-e^{-x})$, $C_n = (2\log n)^{1/2}/c - \frac{\log \pi + \log(\log n)}{2c(2\log n)^{1/2}}$, and $S_n = \frac{1}{c(2\log n)^{1/2}}$, where n is the number of observations.

Therefore, with a significance level of α and let the threshold for $\frac{|L(i)-C_n}{S_n}$is β, then $\beta = -\log(-\log(0.99))$. Therefore, if $\frac{|L(i)-C_n}{S_n} < \beta$, we reject the hypothesis of no jump at t_i.

The misclassification can be categorized as two cases: (1). the failure to detect an existing jump (FTD_i) at t_i; (2). the mistake of wrongly including a jump which doesn't exist. Original paper shows if $\widehat{A(T)}$ is the estimator of the jumps in [0,T] from the test, and $A(T)$ refers to the real jumps, then the probability of global misclassification is

$$P(\widehat{A(T)} \neq A(T)) = \frac{2}{\sqrt{2\pi}} y_n N + \exp(-\beta_n) + o(\exp(-\beta_n))$$

where $\beta_n = 1 - \alpha$ and $y_n = (\beta_n S_n + C_n)c\sigma\sqrt{\Delta t}$.

2.5.2 Jump Detection Experiment on Individual Stocks

From our simulation work, we can confirm that the jump detector has a stable and efficient ability to detect jumps from random Brownian Motion. The testing error is small with

K=2500 observation window using second tick data. To detect the movement of capital on the stock market, we apply the jump detector to the TAQ high-frequency data with a K=2500, alpha=0.01 significance level. A 10-minute-spike pattern in the jump probability distribution is observed in individual stocks.

We record the probability that a jump would happen in an individual minute for all the trading days in the following way: first, the jump detector is applied to detecting jumps for IBM and SPY individually, and if there is a jump on a certain minute, say t_i, where $0 < t_i < 391$ (9:30 to 16:00), then $P[t_i]$ increases by one. Next, the detector goes through the whole trading days from 2005 to 2009 in this way, and after normalization, it results in the jump probability distribution graphs for IBM and SPY.

Unlike the 5-minute patterns that we observed in previous sections based on thousands of stocks, it is observable that every ten minutes the jump probability shows a sudden increase in individual stocks.

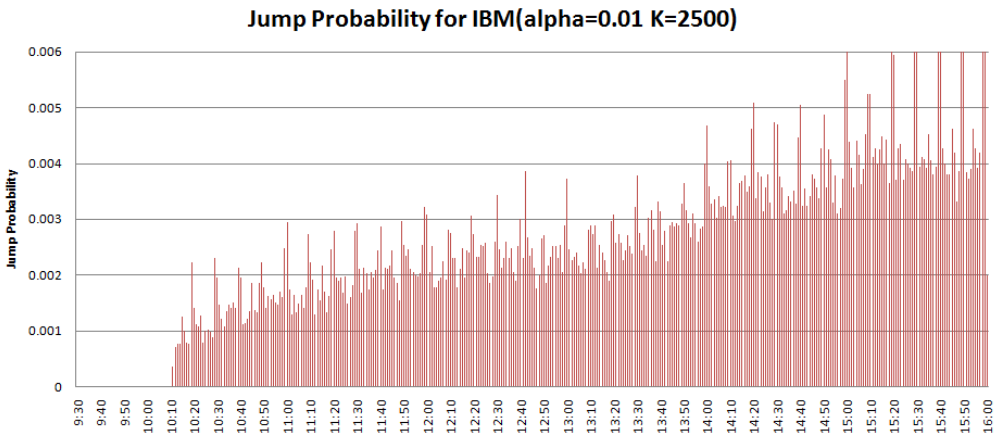


Figure 11. Jump Probability for IBM

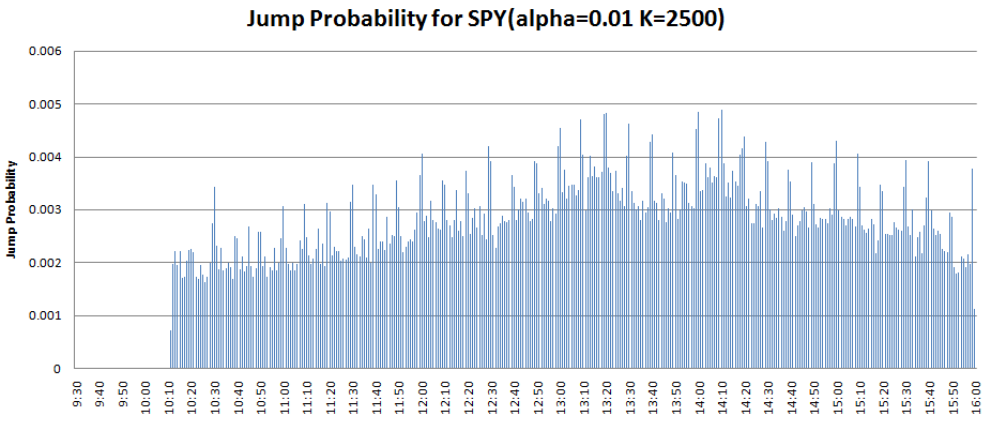


Figure 12. Jump Probability for SPY

To verify this 10-minute observation, multiple price movements on other stocks are tested. We apply the jump detector to FNM, AAPL, GE, INTC, and QQQQ in the same way for general days, and record the probability that jumps would occur for individual minutes from 9:30 a.m. to 16:00 p.m. We mark every ten minutes starting from 9:30 a.m. in red, and it is evincing that these red lines mostly show a spike.

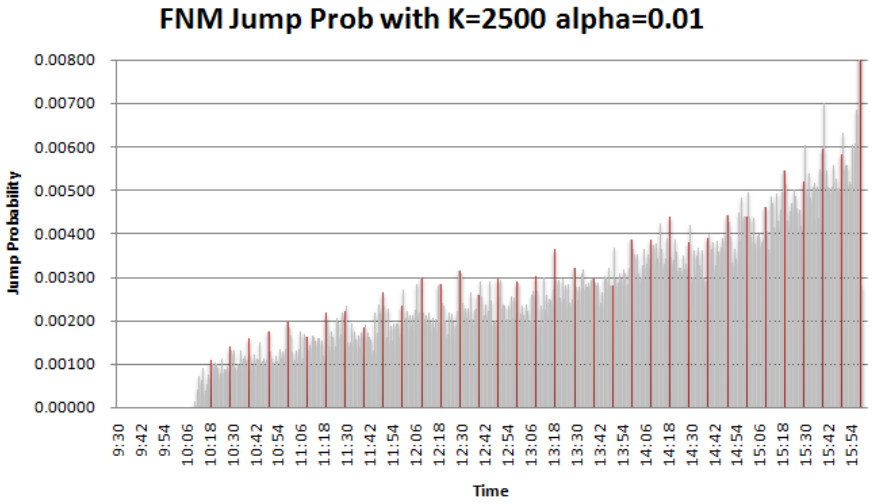


Figure 13. FNM Jump Probability

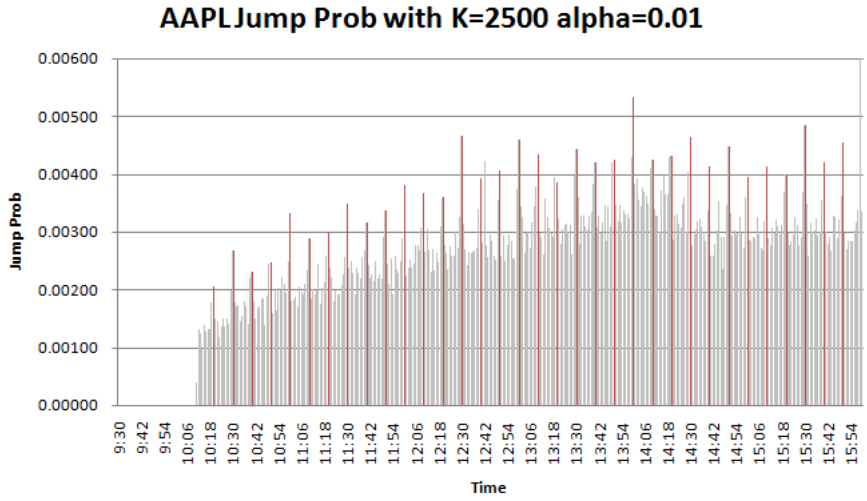


Figure 14. AAPL Jump Probability

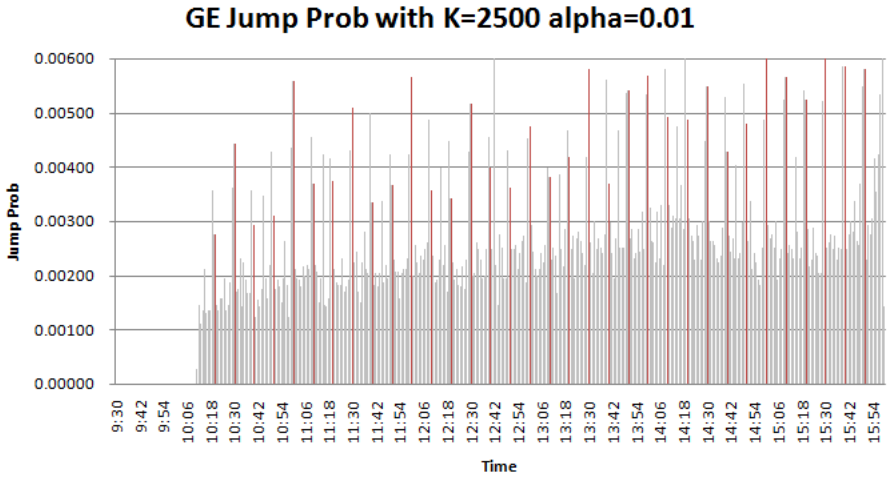


Figure 15. GE Jump Probability

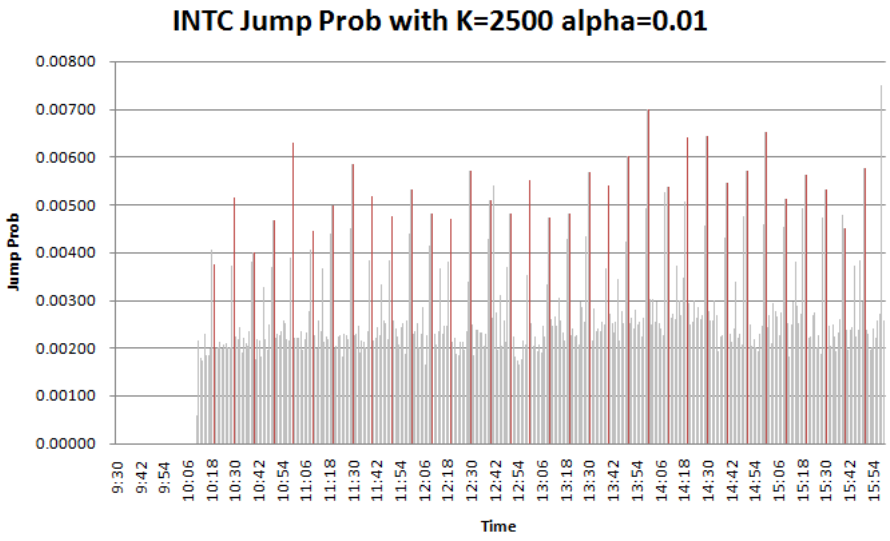


Figure 16. INTC Jump Probability

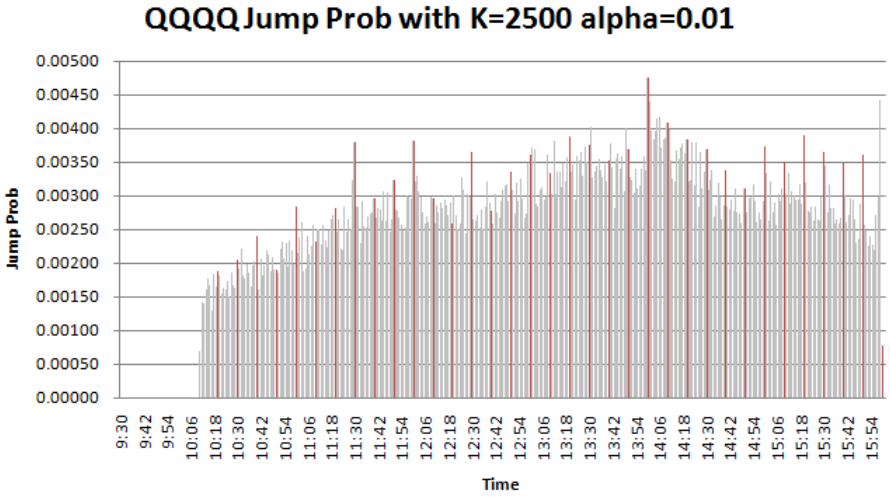


Figure 17. QQQQ Jump Probability

The jump probability spike in every ten minutes explains the source of the spike in cross-sectional volatility, that the later is a result of individual stocks' price jump behaviors. However, we don't know if the jump is due to factors or exists in idiosyncratic volatility in each instrument. For example, assuming the factor model that

$$y_i = Vx_i + \varepsilon_i$$

We do not know if the volatility spike exists in V, such that $V = V_t$ is periodic or if $D = D_t$ such that the idiosyncratic volatility change on every ten minute level.

2.6 A Historical Evolving Path Study on the Spike Phenomenon

A direct observation gives an impression that the 5-minute phenomenon can be further grouped according to the magnitude on each spike. First we labels the spikes according the minute on which the spike occurs during one hour. We label 12 time points: T+1, T+6, T+11, T+16, T+21, T+26, T+31, T+36, T+41, T+46, T+51, T+56. Here T refelcts the fact that we assume every other hour the phenomenon is under same periodic patterns, and T+i means the i-th minute during that hour. We construct spike ratio statistics and build up the null hypothesis that the market is uniform and spike phenomenon doesn't exist, so that we can set up the boundary where the null hypothesis is rejected.

2.6.1 Cluster of Spike Ratio Statistics

From the top 100 stocks which have the top Lee-mykland spike ratio, a direct impression that the spike on 1-min and 31-min may come from the same category, while the spikes on other 5 minute interval come from other categories. We build the spike ratio number to test this thought.

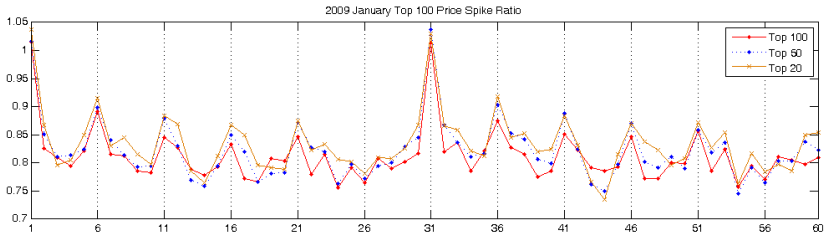


Figure. 18. The Similarity and Sub-groups of Different Spikes.

For a series $x = \{x_t\}$, we define the spike ratio statistics

$$s_{t,m} = \begin{cases} 1 & x_{t-1} \leq x_t \leq x_{t+1} \\ -\frac{1}{3} & otherwise \end{cases}$$

The 1 and $-\frac{1}{3}$ number is determined so that under neutral assumption that x_t has only $\frac{1}{3}$ chance to be the maximum among its neighbouring points, $s_{t,m}$ should have a zero mean.

For each month's return time series, we first remove U shape phenomenon and normalize the data by rational fit. Then for each minute we calculate the standard deviation as $x_{t,m}$. From $x_{t,m}$ we further obtain the $s_{t,m}$ series by checking it's neighbouring data points. First k-means clustering is applied using 2 groups, with the result recorded in the following table:

Group 1(min)	6 16 26 36 46 56
Group 2(min)	1 11 21 31 41 51

Table 3. Clustering of Spike Ratio

Next we draw the cumsum plot for the $s_{t,m}$ statistic. By assuming a neutral probability of a spike striking on the t minute, a flat plot is expected; otherwise, the cumsum plot is able to show a tendency how the spike statistics evolves in time. Each plot is normalized by the total number of months from 1993 to 2009, so that the maximum possible spike ratio would equal to 1.

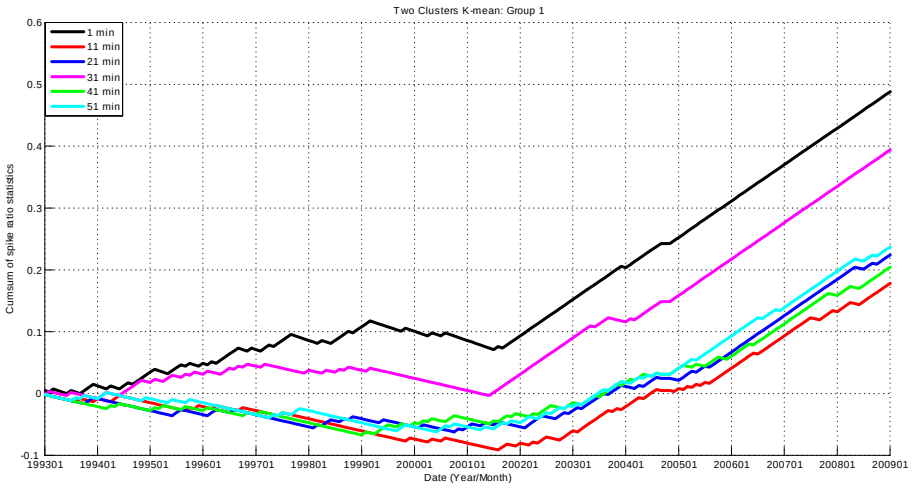


Figure 19. The Cumsum Plot of Spike Ratio Group 1.

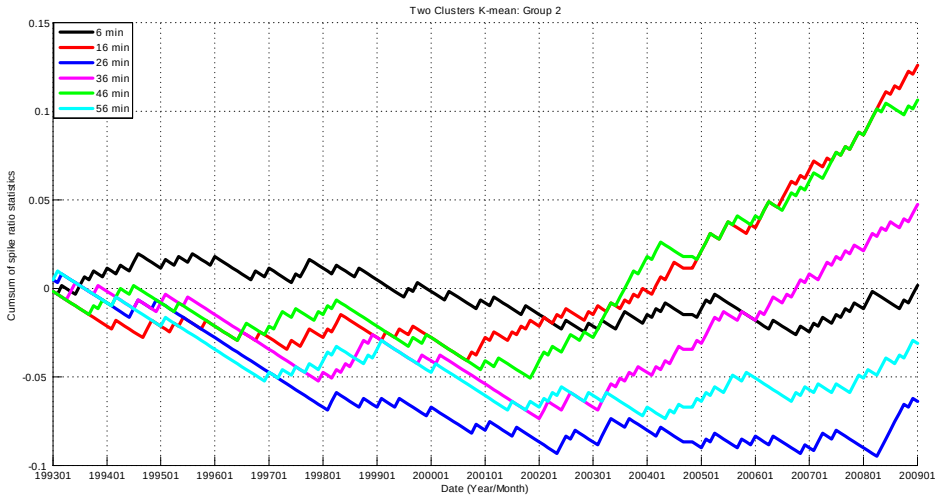


Figure 20. The Cumsum Plot of Spike Ratio Group 2.

As shown in Figure 19 and Figure 20, the spike occuring at different minute shows a different evolving path. The 1-min group and 31-min are following the same path : they show a positive cumsum spike ratio between 1993 and 1999 time period, which is followed by a decrease in 2000 to 2001, and an amplified increase since middle 2001. Before 1999, the cumsum spike ratio statistic is 0.13 for 1-min group and 0.05 for 31-min group. Between 1999 January and 2001 June, the cumsum spike ratio is -0.02 for 1-min group and -0.05 for 31-min group. After 2001, the cumsum spike ratio statistic is 0.43 for 1-min group and 0.40 for 31-min group. The spike ratio is therefore very high for 1-min and 31-min group since the maximum possible spike ratio statistic is 1.

The 11-min, 21-min, 31-min and 41-min are similar in the sense that they only start to show a positive cumsum since middle 2001. The average cumsum statitstic ratio for this four groups is approximately 0.25.

Compare with cluster one, the second cluster has generally weaker spike ratio performance. 16-min group and 36-min group have a similar cumsum spike ratio number: 0.13 versus 0.11. The other groups do not show strong spike performance.

The conclusion can also be reached from the table of average spike ratio statistics in different time period. Generally, the time period can be divided into three parts: 1993-1999, 1999-2001.06 and 2001.07-2009. Since 2001.07, the spike phenomenon becomes prominent.

	1993-1999	1999-2001.06	2001.07-2009	1993-2009
1-min	0.4301	-0.0729	0.9444	0.5915
6-min	0.0215	-0.1146	0.1078	0.0425
11-min	-0.1935	-0.1979	0.6046	0.2124
16-min	-0.086	0.0104	0.3562	0.1536
21-min	-0.129	-0.0729	0.6569	0.2778
26-min	-0.2151	-0.1146	0.0425	-0.0621
31-min	0.129	-0.2396	0.8922	0.4477
36-min	-0.086	-0.2396	0.317	0.0948
41-min	-0.2151	0.1354	0.5654	0.2386
46-min	-0.0645	-0.1563	0.3301	0.1209
51-min	-0.1505	-0.0729	0.6961	0.2908
56-min	-0.129	-0.1563	0.0686	-0.0294

Table 4. Average spike ratio statistics in different years

2.6.2 Cluster of Modified Spike Ratio Statistics and the Confident Interval

To construct the confident interval, we construct the modified spike ratio statistic:

$$s^1_{t,m} = \begin{cases} 1 & x_{t-1} \le x_t \le x_{t+1} \\ 0 & otherwise \end{cases}$$

with

$$p(s^1_{t,m} = 1) = \frac{1}{3}$$

and

$$p(s^1_{t,m} = 0) = \frac{2}{3}$$

which becomes a binomial distribution $B(n, \frac{1}{3})$. Assuming n is sufficiently large, and $S_N = \sum_{m=1}^{N} s^1_{t,m}$ has a asymptotical normal distribution that $S_N \sim N(\frac{1}{3}N, \frac{2}{9}N)$. We construct the null hypothesis that S_N follows binomial distribution, that there is no spike phenomenon.

With 5% and 95% interval drawn, as shown in figure and figure, the groups in first cluster in the 2001 to 2009 all reject the null hypothesis. As a contrast, only 16-min and 46-min groups in the second cluster reject the null hypothesis (with 36-min group is on the boundary).

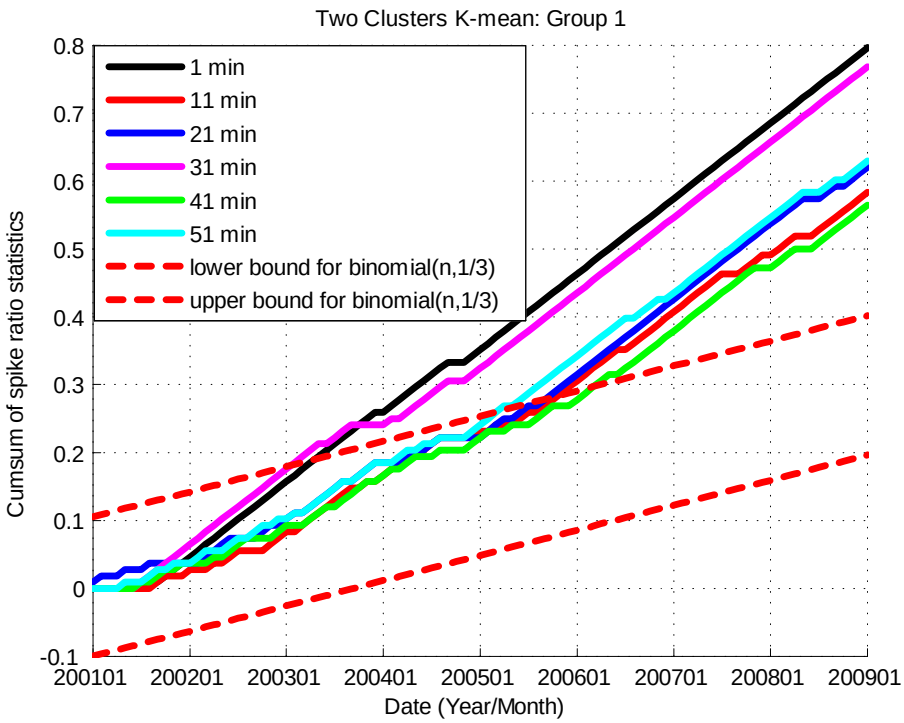


Figure 21. The Confidence Interval for Group 1.

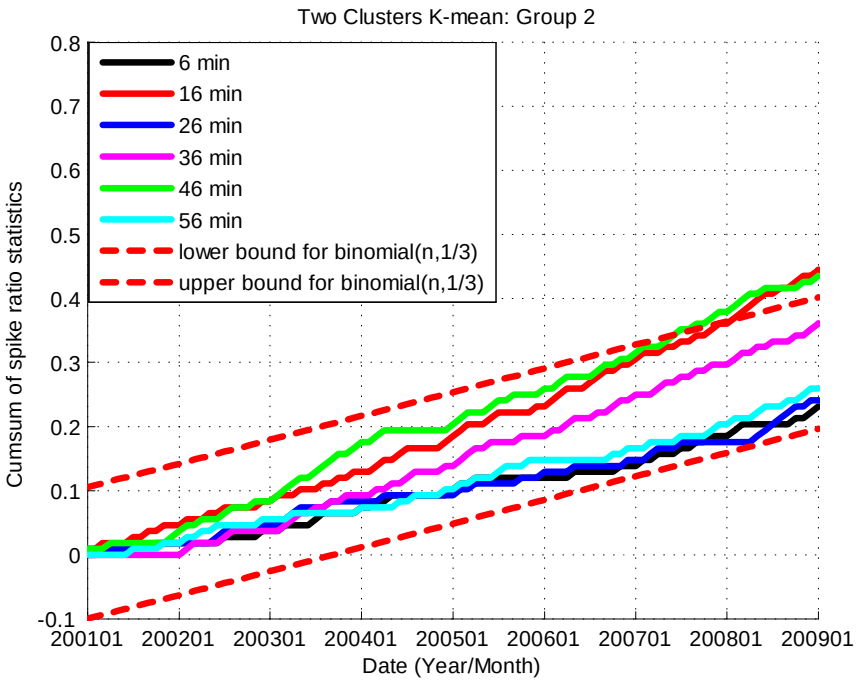


Figure 22. The Confidence Interval for Group 2.

It is also reflected in the Table 5. The boundary for average 5% and 95% interval is $[0.2, 0.4]$, and from the Table. it can be seen that in 1993-1998 time period, only 1-min group rejects the null hypothesis (31-min group is close to the boundary with 0.39 value); after 2001.07, all the $10n+1$ min groups reject the hypothesis, and the 16-min, 36-min and 46-min reject the hypothesis.

	1993-1998	1999-2001.06	2001.07-2009	1993-2009
1-min	0.6129	0.1875	0.951	0.6912
6-min	0.3065	0.1563	0.3235	0.2794
11-min	0.1452	0.0938	0.6961	0.4069
16-min	0.2258	0.25	0.5098	0.3627
21-min	0.1935	0.1875	0.7353	0.4559
26-min	0.129	0.1563	0.2745	0.201
31-min	0.3871	0.0625	0.9118	0.5833
36-min	0.2258	0.0625	0.4804	0.3186
41-min	0.129	0.3438	0.6667	0.4265
46-min	0.2419	0.125	0.4902	0.3382
51-min	0.1774	0.1875	0.7647	0.4657
56-min	0.1935	0.125	0.2941	0.2255

Table 5. The average modified spike ratio for three time period.

As the conclusion in this section, in the two clusters between $10n+1$ minute and $10n+6$ minute groups, we find that the spike phenomenon has separated evolving path. The 1-min and 31-min are starting early as back to 1993. While the 11-min, 21-min, 41-min and 51-min groups are starting after 2011. Also the difference exists among $10n+6$ min group, that only 16-min, 36-min and 46-min groups show a significant spike phenomenon under 5% and 95% interval.

2.7 Conclusion and Discussion

In this section, we identify a 5-minute spike phenomenon for both Nasdaq and NYSE market during the time period 1993 to 2009. We find this phenomenon observable on individual stock level. We further find this phenomenon can be subgrouped that the T+1 and T+31 minute groups are similar and have an early starting time back to 1993, while the T+11, T+21, T+41 and T+51 group starts to be significant since 2001.

A possible explanation for this phenomenon is the usage of algorithm trading. The algorithm trading has become popular since 1990, and currently count for 70% of the U.S stock market volume and also occupy a high trading volume in foreign exchange markets (Boehmer et al, 2012; Chaboud et al, 2009; HenderShott et al, 2011). In algorithm trading, the trading system includes generally three parts: signal estimation, risk estimation and portfolio selection. The risk estimation usually requires high computation resource and is

done in a fixed time circle. When the risk estimation is re-evaluated, the portfolio needs to be rebalanced, where the trading occurs. The fixed rebalancing time window can be a causing reason for the 5-minute spike phenomenon. Nevertheless, this hypothesis needs more detailed research and may be a result of more complicated and multiple causes.

3 High Dimension Models and the Small Sample Problem

3.1 Factor Model

In this section, we study a case where high dimension parametrization is used through factor model. Typically, there are two different approaches when doing factor analysis. The first type is to apply exploratory analysis by building up empirical factors(therefore we call this type of method the empirical factor analysis) and test how much variance is explained through empirical factors [20]. For example, in the Fama-French factor model, risk-free rate, market capitalization and book-to-market ratio are found to be three factors explained 90% of the diversified portfoflio returns. The second type of factor analysis is through hidden factor analysis, in which the whole factor structure is considered as unknown, and EM algorithm is derived to iteratively estimate the factor loadings.

These two methods are exclusive between each other, and have their own merits. The hidden factor model is unbiased and purely data driven, which theoretically would converge to the true underlying market structure. However as discussed above for most of the case the data is still insufficient to reduce the estimation variance. The empirical factor analysis, on the other hand, has the merit of using prior information to further reduce the number of parameters to estimate, and therefore decrease the variance. However, the empirical construction of factors which are impacting the market would be hard and almost guaranteed to be incomplete.

As the common problems encountered in high dimension parametrization, the sample size is often less than sufficiently large, so that the models can be ill-conditioned. Assuming under the scenario that the portfolio manager needs to run a portfolio optimization based on Russell 3000 index. To have a none degenerated sample covariance matrix, it requires to have more than 3000 trading days' data, that is approximately 12 years. This requirement

leads to two potential problems. First of all, a lot of instruments would not have a long enough history. For example, the ETF GLD(SPDR Gold Shares) is only available since 2004. Also, it is questionable how good that old information can help with the estimation accuracy.

To solve this problem, one way is to reduce the parameter that requires to be estimated. The factor structure is a good candidate covariance structure. It states that the return of any asset y_i satisfies

$$y_i = Vx_i + \varepsilon$$

where V is the factor loading of dimension $d \times k$, which is far less than the number of natural covariance parameters. The factor model has been widely applied and studied [20].

In the following sections, we first review the sample size effect during the EM algorithm estimation on the statistical factor model, and we propose two approaches to solve this issue: using Jeffrey's prior and developing a so-called the EH factor model.

3.2 Sample Size Effect

In real practice, the quality of estimated factor covariance depends on sample size, which, as discussed in the introduction part, can in certain situations be not enough. Although the factor structure highly reduces the parameters that is required to be estimated, the updating formula requires to calculate the inverse of sample covariance, therefore it requires at least $N = d$ samples to insure the sample covariance to be positive definite. Even if $N \geq d$, the sample covariance may not be close to the "true" market covariance Σ, as shown in the following simulation.

In the simulation, we set up the "true" market as a factor structure, that

$y = Vx + \varepsilon$ and the covariance is $\Sigma = VV^T + D$. In this case, Σ is 20×20, while V is 20×10. From sample covariance C by forming random factor x_i where $i = 1, ..., N$ and transform x_i by V to get y_i. Then we estimated factor structure $\bar{V}$ and $\bar{D}$. In this way we obtain the estimator for Σ, denoted as $\bar{\Sigma}$. We use three different ways to measure the distance between Σ and C, and between Σ and $\bar{\Sigma}$.

First we measure the Frobenius norm between $\bar{\Sigma}$ and C, that is $\left\|\bar{\Sigma} - C\right\|_F$ and compared with $\left\|\bar{\Sigma} - \Sigma\right\|_F$. We want to compare both cases, because $\left\|\bar{\Sigma} - C\right\|_F$ is the target to minimize for maximum likelihood estimation, while $\left\|\bar{\Sigma} - \Sigma\right\|_F$ is the target we want to

minimize, which reflects the distance between the estimated covariance the the true market covariance.

Second we calculate the Hellinger distance between estimated covariance and sample covariance, and between true covariance. Hellinger distance is calculated by assuming two covariance Σ_1 and Σ_2 are from two different distribution p and q, and the Hellinger distance between p and q are defined as

$$\begin{aligned} H(p,q) &= \frac{1}{2}\int (p^{\frac{1}{2}} - q^{\frac{1}{2}})^2 dx \\ &= 1 - \int p^{\frac{1}{2}} q^{\frac{1}{2}} dx \end{aligned} \tag{1}$$

Since here p and q are Gaussian distribution,

$$\begin{aligned} p(x|\Sigma) &= (2\pi)^{-\frac{p}{2}} |\Sigma_1|^{-\frac{1}{2}} \exp(-\frac{x^T \Sigma_1^{-1} x}{2}) \\ q(x|\Sigma) &= (2\pi)^{-\frac{p}{2}} |\Sigma_2|^{-\frac{1}{2}} \exp(-\frac{x^T \Sigma_2^{-1} x}{2}) \\ \int p^{\frac{1}{2}} q^{\frac{1}{2}} dx &= \int (2\pi)^{-\frac{p}{2}} |\Sigma_1|^{-\frac{1}{4}} |\Sigma_2|^{-\frac{1}{4}} \exp(-\frac{1}{4} x^T (\Sigma_1^{-1} + \Sigma_2^{-1}) x) dx \\ &= |\Sigma_1|^{\frac{1}{4}} |\Sigma_2|^{\frac{1}{4}} |\frac{\Sigma_1 + \Sigma_2}{2}|^{-\frac{1}{2}} \\ H(\Sigma_1, \Sigma_2) &= 1 - |\Sigma_1|^{\frac{1}{4}} |\Sigma_2|^{\frac{1}{4}} |\frac{\Sigma_1 + \Sigma_2}{2}|^{-\frac{1}{2}} \end{aligned} \tag{2}$$

Also we evaluate the log likelihood $l(V, D|data)$. We use $l(V, D|C)$ as in 5 and also compared with $l(V, D|\Sigma)$. The later function reflects the log likelihood value of the parameter when having a perfect sample covariance, and is an increasing function as $VV^T + D \to \Sigma$.

In our experiment, we use $V_{20\times 10}$ and $D_{20\times 20}$ as the true market. With each $N = 20, ..., 200$, we run the factor model for 100 times and each time with 1000 iterations. Since the true market structure is unobservable where the decision on numbers of factor is not guarantee to be optimal, we use $\bar{V}_{20\times 5}$ as the model factor loadings. In this way, the sanity of the model would not depend on a miracle factor number decision and the distance between the covariance estimation and the true covariance would not be 0 at the best case (where we have a "perfect" sample covariance). Comparing the distance between $\bar{\Sigma}$ and C, it can be seen that Frobenius norm is not a good candidate to show the sample size effect, while under Hellinger distance $H(\bar{\Sigma}, C)$ shows that when N goes to 80, which is four times of the dimension d, the distance . As can be seen in Figure 1, while $\left\|\bar{\Sigma} - C\right\|_F$ is not sensitive to the sample size. $\left\|\bar{\Sigma} - \Sigma\right\|_F$ sees a continuous drop as the sample size N. There is an

interesting behavior that $l(V, D|C)$ is a decreasing function while $l(V, D|\Sigma)$ increases as the sample size increases. The reason is that the model space by $\{\bar{V}_{20\times5}, \bar{D}_{20\times20}\}$ is smaller than the true covariance space $\{V_{20\times5}, D_{20\times20}\}$ and therefore the more samples it draws, the less chance the sample covariance can be explained by the model. Meanwhile, when measuring the likelihood by $l(V, D|\Sigma)$, the sample space is fixed by Σ, therefore the gap between model space and true space is a constant piece, and therefore only sample size effects the likelihood value.

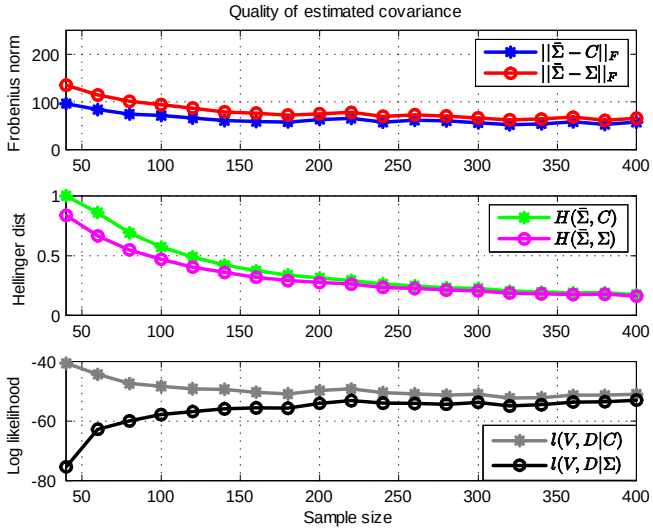


Figure 23. The Quality of Estimated Covariance.

3.3 The Hidden Factor EM Algorithm

3.3.1 Review of EM algorithm

Rubin proposed the following factor analysis method [46]:

The basic model is

$$y = Vx + \varepsilon \tag{3}$$

where y, the i-th observation, is a vector of order d; x, vector of order $p < d$, is the hidden factor variable following normal distribution $N(0, 1)$; ε is the innovation that follows normal distribution $N(0, D)$; V is the factor loading matrix of the size $d \times p$. Because of the property

of linearity, y is a normal distribution with zero mean value and has the variance-covariance matrix, $\Sigma = E[yy^T]$, is

$$\Sigma = VV^T + D \tag{4}$$

The elements of V and D are parameters to be estimated from the data. Suppose that from a random sample of N observations of y we find the sample covariance C, whose elements are the usual estimates of variances and covariances of the components of y and follow a Wishart distribution with d degree of freedom. The log likelihood function is given by

$$\begin{aligned} L(V,D) &= \log \prod_{i=1}^{N} p(y_i) \\ &= -\frac{Np}{2}\log(2\pi) - \frac{N}{2}\log\det(VV^T + D) - \frac{1}{2}\sum_{i=1}^{N}\{Tr(VV^T + D)^{-1} y_i y_i^T\} \\ &= -\frac{Np}{2}\log(2\pi) - \frac{N}{2}\log\det(VV^T + D) - \frac{N}{2}Tr(VV^T + D)^{-1}C \end{aligned} \tag{5}$$

Another useful way to look at the log likelihood function is to consider the conditional distribution of the observation given hidden factors, that

$$\begin{aligned} L &= \log \prod_{i=1}^{N} p(y_i; x_i) \\ &= \sum_{i=1}^{N} \lg\{p(y_i|x_i)p(x_i)\} \\ &= \sum_{i=1}^{N} \lg p(y_i|x_i) + \sum_{i=1}^{N} \lg p(x_i) \end{aligned} \tag{6}$$

The EM algorithm proposed by Rubin iteratively follows two steps: evaluate $E[L]$ (E-step) based on current parameter V, D and find a new parameter V_1,D_1 which maximize the $E[L]$ (M-step).

Notice that to maximize the value of L, which is a function of V, D and $\{y_i | i \in [1, N]\}$, it is the same to maximize only $\Sigma_{i=1}^{N} \lg p(y_i|x_i)$ since x is independent of V, D and $\{y_i\}$, and therefore

$$
\begin{aligned}
\{V_1, D_1\} &= \arg\max E[L] \\
&= \arg\max\{E[\sum_{i=1}^{N} \lg p(y_i|x_i)]\} \\
&= \arg\max\{E[-\frac{Np}{2}\lg 2\pi - \frac{N}{2}\lg\det D - \sum_{i=1}^{N}\frac{(y_i - Vx_i)^T D^{-1}(y_i - Vx_i)}{2}]\}\} \\
&= \arg\max\{-\frac{N}{2}\lg\det D - \frac{1}{2}\sum_{i=1}^{N} y_i^T D^{-1} y_i + \sum_{i=1}^{N} E[x_i|y_i]^T V^T D^{-1} y_i - \frac{1}{2}\sum_{i=1}^{N} Tr\{V^T D^{-1} V E[x_i x_i^T|y_i]\}\} \quad (7)
\end{aligned}
$$

3.3.2 E-step

In E-step, it find the explicit formula for 7 where $E[x_i|y_i]$ and $E[x_i x_i^T|y_i]$ is expressed as a function of current parameter $\{V, D\}$ and the data y_i, where $y_i \in [1, .., N]$. The calculation of these sufficient statistics are based on treating the y_i and x_i as joint normal distribution and applying the formula for conditional expectation of first and second moment, that since $y_i = Vx_i + \varepsilon_i$,

$$
\begin{bmatrix} x_i \\ y_i \end{bmatrix} \tilde{} N(0, \begin{bmatrix} I & V^T \\ V & D + VV^T \end{bmatrix})
$$

By conditional expectation formula for multivariate normal distribution,

$$
\begin{aligned}
E[x_i|y_i] &= V^T(D + VV^T)^{-1} y_i \quad (8) \\
&\triangleq \delta y_i
\end{aligned}
$$

where

$$
\delta = V^T(D + VV^T)^{-1} \quad (9)
$$

so that δ reflects the inverse transformation from the observation data to the conditional expectation of the hidden factors.

Notes that δ can also be writtenas :

$$\begin{aligned}
\delta &= V^T(D+VV^T)^{-1} \qquad (10)\\
&= V^T(D^{-1}-D^{-1}V(I+V^TD^{-1}V)^{-1}V^TD^{-1})\\
&= \{I-V^TD^{-1}V(I+V^TD^{-1}V)^{-1}\}V^TD^{-1}\\
&= \{(I+V^TD^{-1}V)(I+V^TD^{-1}V)^{-1}-V^TD^{-1}V(I+V^TD^{-1}V)^{-1}\}V^TD^{-1}\\
&= \{(I+V^TD^{-1}V)-V^TD^{-1}V\}(I+V^TD^{-1}V)^{-1}V^TD^{-1}\\
&= (I+V^TD^{-1}V)^{-1}V^TD^{-1}
\end{aligned}$$

Similarly, the conditional expectation of the second moment of the factor, $E[x_ix_i^T|y_i]$, is calculated directly through the conditional expectation formula for multivariate normal distribution, that

$$\begin{aligned}
E[x_ix_i^T|y_i] &= Var(x_ix_i^T|y_i)+E[x_i|y_i]E[x_i|y_i]^T\\
&= I-V^T(D+VV^T)^{-1}V+\delta y_iy_i^T\delta^T\\
&\triangleq \Delta+\delta y_iy_i^T\delta^T \qquad (11)
\end{aligned}$$

where

$$\begin{aligned}
\Delta &\triangleq I-V^T(D+VV^T)^{-1}V\\
&= I-\delta V \qquad (12)
\end{aligned}$$

Therefore, the expectation of log likelihood function can be written as a combination of current V, D parameter and the sample covariance, and the next step updating formula for V, D would be

$$\begin{aligned}
\{V_1,D_1\} &= \arg\max\{-\frac{N}{2}\lg\det D-\frac{1}{2}\sum_{i=1}^{N}y_i^TD^{-1}y_i+\sum_{i=1}^{N}y_i^T\delta^TV^TD^{-1}y_i-\frac{1}{2}\sum_{i=1}^{N}Tr\{V^TD^{-1}V(\Delta+\delta y_iy_i^T\delta^T)\}\}\\
&= \arg\max\{-\frac{N}{2}\lg\det D-\frac{N}{2}Tr(D^{-1}C)+NTr(\delta^TV^TD^{-1}C)-\frac{N}{2}Tr\{V^TD^{-1}V(\Delta+\delta C\delta^T)\}\} \qquad (13)
\end{aligned}$$

3.3.3 M-step

In M-step, the update formula for V is obtained by taking derivative of 13 and set it to zero. The detail of the deduction as shown below,

$$\begin{aligned}
\frac{\partial E[L]}{\partial V_{r,s}} &= \frac{\partial\{-\frac{N}{2}\lg\det D - \frac{N}{2}Tr(D^{-1}C) + NTr(\delta^T V^T D^{-1}C) - \frac{N}{2}Tr\{V^T D^{-1}V(\Delta+\delta C\delta^T)\}}{\partial V} \\
&= N\frac{\partial Tr(\delta^T V^T D^{-1}C)}{\partial V_{r,s}} - \frac{N}{2}\frac{\partial Tr\{V^T D^{-1}V(\Delta+\delta C\delta^T)}{\partial V_{r,s}} \\
&= NTr(\delta^T E_{r,s}^T D^{-1}C) - \frac{N}{2}Tr\{E_{r,s}^T D^{-1}V(\Delta+\delta C\delta^T) + V^T D^{-1}E_{r,s}(\Delta+\delta C\delta^T)\} \\
&= N(D^{-1}C\delta^T)_{s,r} - N(D^{-1}V(\Delta+\delta C\delta^T))_{s,r}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\partial E[L]}{\partial V} &= N(D^{-1}C\delta^T)^T - N(D^{-1}V(\Delta+\delta C\delta^T))^T = 0 \\
D^{-1}C\delta^T &= D^{-1}V(\Delta+\delta C\delta^T) \\
V &= C\delta^T(\Delta+\delta C\delta^T)^{-1} \qquad (14)
\end{aligned}$$

And similarly

$$\begin{aligned}
\frac{\partial E[L]}{\partial (D^{-1})_{i,i}} &= \frac{\partial\{-\frac{N}{2}\lg\det D - \frac{N}{2}Tr(D^{-1}C) + NTr(\delta^T V^T D^{-1}C) - \frac{N}{2}Tr\{V^T D^{-1}V(\Delta+\delta C\delta^T)\}}{\partial D_{i,i}} \\
&= -\frac{N}{2}[D]_{i,i} - \frac{N}{2}C_{i,i} + N(C\delta^T V^T)_{i,i} - \frac{N}{2}\{V(\Delta+\delta C\delta^T)V^T\}_{i,i} \\
D &= diag(C - 2V\delta C + V(\Delta+\delta C\delta^T)V^T) \qquad (15)
\end{aligned}$$

We obtain this by solving the equation of setting the derivative of 7 to zero. Given a random point as initial value, EM iteration starts from a random point and convergence is guaranteed to a local maximum[Appendix B], and V is unique up to unitary transformation, that

$$VV^T + D = (VQ)(VQ)^T + D \qquad (16)$$

where Q is arbitrary unitary transformation. That is

$$\Sigma = VV^T + D \qquad (17)$$

3.4 The Jeffrey's Prior for the Factor Model Estimation.

3.4.1 The Fisher Information for Factor Model

Th EM algorithm has become a popular numerical method for its low complexity of implementation and robustness (McLachlan et al, 2007). The main drawback of the EM algorithm is its speed of convergence. Dempster, Laird, and Rubin (1977) showed that the EM algorithm exhibits a linear speed of convergence, with a rate of convergence obtained from the information matrices associated to the missing and complete data sets, that

$$e^{(n+1)}(\Psi) = C(\Psi)e^{(n)}(\Psi)$$

and

$$C(\Psi) = I_c^{-1}(\Psi)I_m(\Psi)$$

where $e^{(n)}(\Psi) = |b^{(n)}(\Psi) - b_{ML}(\Psi)|$ is the error term and $C(\Psi)$ is the rate of convergence of the EM algorithm, $\mathcal{I}_c(\Psi)$ is the information matrix associated to the complete data, $\mathcal{I}_m(\Psi)$ is the information matrix associated with the missing data. This relationship is the fundamental for analyzing the convergence property of EM algorithm. In next section, we will give detailed

The Cramer-Rao bound is a lower bound on the error variance of any unbiased estimate, and as such serves as a useful benchmark for practical estimators[H.L.Van Trees, Detection, Estimation and Modulation Theory].

Let y be the observed data, x be the missing data, and $z = [y, x]$ is the complete data.

The fisher information matrix is defined as

$$\mathcal{I}_c(\Psi) \triangleq E_z\left[-\frac{\partial^2 L_c(\Psi)}{\partial\Psi\partial\Psi^T}\right]$$

Similary, we can define the fisher information when given y is observed, that is to take expectation over the conditional distribution $p(x|y)$:

$$\mathcal{I}_c(\Psi; y) \triangleq E_x\left[-\frac{\partial^2 L_c(\Psi)}{\partial\Psi\partial\Psi^T}\right]$$

Note we use E_z to treat all variables as random variables, while E_x means only x is the random variable.

Here the symbol $\mathcal{I}_c(\Psi; y)$ means y is treated as parameters.

3.4.2 Information Matrics and their Relationships

Fisher information matrix, as a metric, naturally arises in the maximum likelihood estimation as a measure between estimated parameters.

Definition 1. *Assuming y is the observable random variable, x is unobservable random variable, and $z = [x, y]$ as the complete data. Define $L_o(\Psi)$ be the likelihood for observed data, that*

$$L_o(\Psi) = p(y; \Psi)$$

Define L_c

Let $L_o(\Psi)$ be the likelihood function for observed data, then

$$\begin{aligned} p(y;\Psi) &= \frac{p(x,y;\Psi)}{p(x|y;\Psi)} \\ \log L_o(\Psi) &= \log L_c(\Psi) - \log p(x|y;\Psi) \end{aligned} \tag{18}$$

We let

$$I_o(\Psi) = -\frac{\partial^2 \log L_o(\Psi)}{\partial\Psi\partial\Psi^T}$$

With respect to the complete-data log likelihood, we let

$$I_c(\Psi) = -\frac{\partial^2 \log L_c(\Psi)}{\partial\Psi\partial\Psi^T}$$

Taking the second derivative of both sides of negative 18 with respect to Ψ,

$$I_o(\Psi) = I_c(\Psi) + \frac{\partial^2 \log p(x|y;\Psi)}{\partial\Psi\partial\Psi^T} \tag{19}$$

Take expectation of 19 over the conditional distribution of x,

$$\begin{aligned} E_x\left[I(\Psi)\right] &= I_o(\Psi) \\ E_x[I_c(\Psi)] &= \mathcal{I}_c(\Psi; y) \\ \mathcal{I}_m(\Psi; y) &\triangleq E\left[-\frac{\partial^2 \log p(x|y;\Psi)}{\partial\Psi\partial\Psi^T}\right] \end{aligned}$$

Therefore

$$I_o(\Psi) = \mathcal{I}_c(\Psi; y) - \mathcal{I}_m(\Psi; y) \tag{20}$$

3.4.3 The Complete Information With Respect to z

For the factor model,

$$y = Vx + \varepsilon$$

we use the same assumptions for the factor model as previous section, that $x \sim N(0_{k\times 1}, I_{k\times k})$, $\varepsilon \sim N(0_{d\times 1}, D_{d\times d})$.

The complete data is defined as

$$z \triangleq \begin{bmatrix} y \\ x \end{bmatrix} \sim N\left(0_{(d+k)\times 1}, \begin{bmatrix} D + VV^T & V \\ V^T & I \end{bmatrix}\right)$$

Define the covariance of z as

$$\Omega \triangleq \begin{bmatrix} D + VV^T & V \\ V^T & I \end{bmatrix}$$

The parameter is

$$\Psi = \begin{bmatrix} vec(V) \\ diag(D) \end{bmatrix}^T$$

The complete data likelihood is therefore

$$\begin{aligned} L_c(\Psi) &= \prod_{j=1}^{n} p(z_j; \Psi) \\ &= \prod_{j=1}^{n} \frac{1}{2\pi^{d+k}|\Omega|} e^{-\frac{1}{2} z_j^T \Omega^{-1} z_j} \end{aligned}$$

and

$$\begin{aligned} \log L_c(\Psi) &= \sum_{j=1}^{n} \left\{ -\frac{(d+k)}{2} \log 2\pi - \frac{1}{2} \log|\Omega| - \frac{1}{2} \sum_{\iota=1}^{n} z_j^T \Omega^{-1} z_j \right\} \\ &= -\frac{n(d+k)}{2} \log 2\pi - \frac{n}{2} \log|\Omega| - \frac{1}{2} \sum_{j=1}^{n} z_j^T \Omega^{-1} z_j \end{aligned} \tag{21}$$

To further simplify 21, using the property of schur complement, that

$$\begin{aligned} -\frac{n}{2} \log|\Omega| &= -\frac{n}{2} \log\det\left(\begin{bmatrix} D + VV^T & V \\ V^T & I \end{bmatrix}\right) \\ &= -\frac{n}{2} \log\det(I) \det(D + VV^T - VI^{-1}V^T) \\ &= -\frac{n}{2} \log|D| \end{aligned}$$

and the block matrix inversion formula, the inverse of Ω is

$$\Omega^{-1} = \begin{bmatrix} D^{-1} & -D^{-1}V \\ -V^T D^{-1} & I + V^T D^{-1} V \end{bmatrix}$$

Expanding $z_j^T \Omega^{-1} z_j$ terms,

$$\begin{aligned} z_j^T \Omega^{-1} z_j &= \begin{bmatrix} y_j^T & x_j^T \end{bmatrix} \begin{bmatrix} D^{-1} & -D^{-1}V \\ -V^T D^{-1} & I + V^T D^{-1} V \end{bmatrix} \begin{bmatrix} y_j \\ x_j \end{bmatrix} \\ &= y_j^T D^{-1} y_j - 2 x_j^T V^T D^{-1} y_j + x_j^T (I + V^T D^{-1} V) x_j \\ &= (y_j^T - x_j^T V^T) D^{-1} (y_j - V x_j) + x_j^T x_j \end{aligned}$$

and therefore a more concise form for $L_c(\Psi)$ is obtained, that

$$\log L_c(\Psi) = -\frac{n(d+k)}{2} \log 2\pi - \frac{n}{2} \log |D| - \frac{1}{2} \sum_{j=1}^{n} \left\{ (y_j - V x_j)^T D^{-1} (y_j - V x_j) + {x_j}^T x_j \right\}$$

Review two matrix calculus equations::

$$\partial \ln \det X = Tr(X^{-1} \partial X)$$

and

$$\partial (X^{-1}) = -X^{-1} (\partial X) X^{-1}$$

Therefore denoting $d_i \triangleq D_{i,i}$,

$$\begin{aligned} \frac{\partial \log \det D}{\partial d_i} &= D_{i,i}^{-1} \\ &= d_i^{-1} \end{aligned}$$

and

$$\frac{\partial D^{-1}}{\partial d_i} = -D^{-1} e_i e_i^T D^{-1}$$

Plugging the two equations above, we get

$$\begin{aligned} \frac{\partial \log L_c(\Psi)}{\partial d_i} &= -\frac{n}{2} d_i^{-1} - \frac{1}{2} \sum_{j=1}^{n} \left\{ -(y_j - V x_j)^T D^{-1} e_i e_i^T D^{-1} (y_j - V x_j) \right\} \\ &= -\frac{n}{2} d_i^{-1} - \frac{1}{2} \sum_{j=1}^{n} \left\{ -\left((y_j - V x_j)^T D^{-1} e_i \right)^2 \right\} \\ &= -\frac{n}{2d_i} + \frac{\sum_{j=1}^{n} (y_j - V x_j)_i^2}{2 d_i^2} \end{aligned} \tag{22}$$

Similarly,

$$\begin{aligned}\frac{\partial \log L_c(\Psi)}{\partial V_{r_1,r_2}} &= \frac{1}{2}\sum_{j=1}^{n}\left\{2x_j^T E_{r_1,r_2}^T D^{-1}(y_j - Vx_j)\right\} \\ &= \sum_{j=1}^{n} Tr\left\{x_j^T e_{r_2} e_{r_1}^T D^{-1}(y_j - Vx_j)\right\} \\ &= \sum_{j=1}^{n}\frac{(y_j - Vx_j)_{r_1}(x_j)_{r_2}}{d_{r_1}}\end{aligned}$$

It can be validated that $E_z\left[\frac{\partial \log L_c(\Psi)}{\partial d_i}\right] = 0$, since $E_z\left[(y_j - Vx_j)_i^2\right] = E\left[\varepsilon_i^2\right] = d_i$, and $E_z\left[\frac{\partial \log L_c(\Psi)}{\partial V_{r_1,r_2}}\right] = 0$.

From 22, $\frac{\partial^2 \log L_c(\Psi)}{\partial d_i \partial d_j} = 0$ for $i \neq j$.

When $i = j$,

$$\frac{\partial^2 \log L_c(\Psi)}{\partial^2 d_i} = \frac{n}{2d_i^2} - \sum_{j=1}^{n}\frac{(y_j - Vx_j)_i^2}{d_i^3}$$

Therefore,

$$\begin{aligned}E_z\left[\frac{\partial^2 \log L_c(\Psi)}{\partial^2 d_i}\right] &= \frac{n}{2d_i^2} - \sum_{j=1}^{n}\frac{E_z\left[(y_j - Vx_j)_i^2\right]}{d_i^3} \\ &= \frac{n}{2d_{i2}} - \frac{nd_i}{d_i^3} \\ &= -\frac{n}{2d_i^2} \qquad (23)\end{aligned}$$

The cross term is

$$\begin{aligned}\frac{\log L_c(\Psi)}{\partial d_i \partial V_{r_1,r_2}} &= -\sum_{j=1}^{n}\frac{\partial\left((y_j - Vx_j)^T e_i\right)^2}{2d_i^2 \partial V_{r_1,r_2}} \\ &= -\frac{1}{2d_i^2}\sum_{j=1}^{n}\left(-2(y_j - Vx_j)_i x^T E_{r_1,r_2}^T e_i^T\right) \\ &= \frac{1}{d_i^2}\sum_{j=1}^{n}\left((y_j - Vx_j)_i x_j^T e_{r_2} e_{r_1} e_i^T\right) \qquad (24)\end{aligned}$$

The expectation of 24 has the value zero for any i, r_1, r_2, because $(y_j - Vx_j)_i = \varepsilon_i$ and $x^T e_{r_2} = x_{r_2}$ while according to assumption $E_z[e_i x_{r_2}] = 0$ for any i and r_2.

Therefore according to symmetry,

$$\begin{aligned}E_z\left[\frac{\partial^2 \log L_c(\Psi)}{\partial d_i \partial V_{r_1,r_2}}\right] &= 0 \\ E_z\left[\frac{\partial^2 \log L_c(\Psi)}{\partial V_{r_1,r_2} \partial d_i}\right] &= 0\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{\partial^2 \log L_c(\Psi)}{\partial V_{r_1,r_2}\partial V_{r_3,r_4}} &= \frac{1}{d_{r_1}}\sum_{j=1}^{n}\frac{\partial (y_j - Vx_j)^T e_{r_1}}{\partial V_{r_3,r_4}} x_{r2} \\
&= -\frac{1}{d_{r_1}}\sum_{j=1}^{n}\left(e_{r_1}^T\left(\frac{\partial V}{\partial V_{r_3,r_4}})x_j\right)(x_j)_{r_2}\right) \\
&= -\frac{1}{d_{r_1}}\sum_{j=1}^{n} e_{r_1}^T e_{r_3} e_{r_4}^T x_j (x_j)_{r_2} \\
&= \frac{1}{d_{r_1}}\sum_{j=1}^{n} e_{r_1}^T e_{r_3} (x_j)_{r_2}(x_j)_{r_4} \\
&= \begin{cases} -\frac{\sum_{j=1}^{n}(x_j)_{r_2}(x_j)_{r_4}}{d_{r_1}} & r_1 = r_3 \\ 0 & otherwise \end{cases}
\end{aligned}$$

Therefore,

$$E_z\left[\frac{\partial^2 \log L_c(\Psi)}{\partial V_{r_1,r_2}\partial V_{r_3,r_4}}\right] = \begin{cases} -\frac{n}{d_{r_1}} & r_1 = r_3 \;\; r_2 = r_4 \\ 0 & otherwise \end{cases}$$

We summarize the formulas above, that

$$\begin{aligned}
\frac{\partial^2 \log L_c(\Psi)}{\partial^2 d_i} &= \frac{n}{2d_i^2} - \sum_{j=1}^{n}\frac{(y_j - Vx_j)_i^2}{d_i^3} \\
\frac{l^2 \log L_c(\Psi)}{\partial d_i \partial V_{r_1,r_2}} &= \frac{1}{d_i^2}\sum_{j=1}^{n}\left((y_j - Vx_j)_i x_j^T e_{r_2} e_{r_1} e_i^T\right) \\
\frac{\partial^2 \log L_c(\Psi)}{\partial V_{r_1,r_2}\partial V_{r_3,r_4}} &= \frac{1}{d_{r_1}}\sum_{j=1}^{n} e_{r_1}^T e_{r_3}(x_j)_{r_2}(x_j)_{r_4} \qquad (25)
\end{aligned}$$

and

$$\begin{aligned}
E_z\left[\frac{\partial^2 \log L_c(\Psi)}{\partial^2 d_i}\right] &= -\frac{n}{2d_i^2} \\
E_z\left[\frac{\partial^2 \log L_c(\Psi)}{\partial d_i \partial V_{r_1,r_2}}\right] &= 0 \\
E_z\left[\frac{\partial^2 \log L_c(\Psi)}{\partial V_{r_1,r_2}\partial V_{r_3,r_4}}\right] &= -\frac{n}{d_{r_1}} I_{r_1=r_3} I_{r_2=r_4}
\end{aligned}$$

Therefore, the complete information matrix is

$$\mathcal{I}_c(\Psi) = \begin{bmatrix} \frac{n}{2d_i^2} & 0 \\ 0^T & diag\left(\frac{n}{d_{r_1}}\right) \end{bmatrix}$$

3.4.4 The Complete Information With Respect to x

By taking the integration over the conditional distribution of $x|y$, the complete information with respect to x is also obtained based on 25:

$$\begin{aligned} E_x\left[\frac{\partial^2 \log L_c(\Psi)}{\partial^2 d_i}\right] &= \frac{n}{2d_i^2} - \sum_{j=1}^{n} \frac{E\left[(y_j - Vx_j)_i^2\right]}{d_i^3} \\ &= \frac{n}{2d_i^2} - \sum_{j=1}^{n} \frac{E\left[e_i^T (y_j - Vx_j)(y_j - Vx_j)^T e_i\right]}{d_i^3} \\ &= \frac{n}{2d_i^2} - \frac{e_i^T \left\{\sum_{j=1}^{n} y_j y_j^T + VV^T\right\} e_i}{d_i^3} \\ &= \frac{n}{2d_i^2} - \frac{n\left(C + VV^T\right)_{i,i}}{d_i^3} \end{aligned}$$

Similarly,

$$\begin{aligned} E_x\left[\frac{\partial^2 \log L_c(\Psi)}{\partial d_i \partial V_{r_1,r_2}}\right] &= \frac{1}{d_i^2} \sum_{j=1}^{n} E\left[(y_j - Vx_j)_i x_j^T e_{r_2} e_{r_1}^T e_i\right] \\ &= \frac{1}{d_i^2} e_i^T \sum_{j=1}^{n} E\left[(y_j - Vx_j) x_j^T\right] e_{r_2} e_{r_1}^T e_i \\ &= -\frac{n}{d_i^2} e_i^T V e_{r_2} e_{r_1}^T e_i \\ &= -\frac{n}{d_i^2} V_{ir_2} I_{r_1=i} \end{aligned} \tag{26}$$

and

$$E_x\left[\frac{\partial^2 \log L_c(\Psi)}{\partial V_{r_1,r_2} \partial V_{r_3,r_4}}\right] = -\frac{n}{d_{r_1}} I_{r_1=r_3} I_{r_2=r_4}$$

Summarize the three formula above, and we get the complete data information matrix with respect to x

$$\begin{aligned} E_x\left[\frac{\partial^2 \log L_c(\Psi)}{\partial^2 d_i}\right] &= \frac{n}{2d_i^2} - \frac{n\left(C + VV^T\right)_{i,i}}{d_i^3} \\ E_x\left[\frac{\partial^2 \log L_c(\Psi)}{\partial d_i \partial V_{r_1,r_2}}\right] &= -\frac{n}{d_i^2} V_{ir_2} I_{r_1=i} \\ E_x\left[\frac{\partial^2 \log L_c(\Psi)}{\partial V_{r_1,r_2} \partial V_{r_3,r_4}}\right] &= -\frac{n}{d_{r_1}} I_{r_1=r_3} I_{r_2=r_4} \end{aligned} \tag{27}$$

Putting 27 in matrix form as

$$\mathcal{I}_c(\Psi, x) = \begin{bmatrix} \frac{n\left(C+VV^T\right)_{i,i}}{d_i^3} - \frac{n}{2d_i^2} & -\frac{n}{d_i^2}V_{ir_2}I_{r_1=i} \\ -\frac{n}{d_i^2}V_{ir_2}I_{r_1=i} & -\frac{n}{d_{r_1}}I_{r_1=r_3}I_{r_2=r_4} \end{bmatrix}$$

3.4.5 The Missing Information With Respect to x

By multivariate conditional normal distribution property, when

$$\begin{bmatrix} y \\ x \end{bmatrix} \sim N\left(0_{(d+k)\times 1}, \begin{bmatrix} D+VV^T & V \\ V^T & I \end{bmatrix}\right)$$

, it has the property that

$$x|y \sim N\left(V^T(D+VV^T)^{-1}y, I - V^T\left(D+VV^T\right)^{-1}V\right)$$

We simplify the conditional distribution of $x|y$ as

$$\begin{aligned} x|y &\sim N(\delta y, \Delta) \\ \delta &\triangleq V^T(D+VV^T)^{-1} \\ \Delta &\triangleq I - V^T(D+VV^T)^{-1}V \end{aligned}$$

Therefore,

$$\begin{aligned} L_m &\triangleq \prod_{j=1}^{n} p(x_i|y_i;\Psi), \\ \log L_m &= -\frac{nk}{2}\log 2\pi - \frac{1}{2}\sum_{j=1}^{n}(x_j-\delta y_j)^T\Delta^{-1}(x_j-\delta y_j) \end{aligned}$$

from 20,

$$\mathcal{I}_m(\Psi; y) = \mathcal{I}_c(\Psi; y) - I_o(\Psi)$$

we just need to calculate

$$\begin{aligned} I_o(\Psi) &= -\frac{\partial^2 \log L_o(\Psi)}{\partial\Psi\partial\Psi^T} \\ &= -\frac{\partial^2 \log p(y;\Psi)}{\partial\Psi\partial\Psi^T} \end{aligned}$$

Since

$$y \sim N(0, D+VV^T)$$

the likelihood function for y is

$$\begin{aligned}\log L_o(\Psi) &= \sum_{j=1}^{n} \log p(y_j;\Psi) \\ &= -\frac{nd}{2}\log 2\pi - \frac{n}{2}\log\det\left(D+VV^T\right) - \frac{1}{2}\sum_{j=1}^{n} y_j^T(D+VV^T)^{-1}y_j\end{aligned}$$

Applying log determinant derivative trick,

$$\begin{aligned}\frac{\partial \log\det(D+VV^T)}{\partial d_i} &= Tr\left\{(D+VV^T)^{-1}\frac{\partial(D+VV^T)}{\partial d_i}\right\} \\ &= \left(D+VV^T\right)^{-1}_{i,i} \\ \frac{\partial \log\det(D+VV^T)}{\partial V_{r_1,r_2}} &= Tr\left\{(D+VV^T)^{-1}\frac{\partial(D+VV^T)}{\partial V_{r_1,r_2}}\right\} \\ &= 2\left\{\left(D+VV^T\right)^{-1}V\right\}_{r_1,r_2}\end{aligned}$$

Using $S \triangleq VV^T + D$,

$$\begin{aligned}\frac{\partial^2 \log L_0(\Psi)}{\partial d_i^2} &= \frac{n}{2}e_i^T S^{-1}e_i e_i^T S^{-1}e_i - \sum_{j=1}^{n}(e_i^T S^{-1}y_j)(e_i^T S^{-1}e_i e_i^T S^{-1}y_j) \\ &= \frac{n}{2}S^{-2}_{i,i} - \sum_{j=1}^{n} S^{-1}_{i,i}(S^{-1}y_j)_i^2\end{aligned}$$

and similarly

$$\frac{\partial^2 \log L_0(\Psi)}{\partial d_i \partial V_{r_1,r_2}} = n(S^{-1}V)_{i,r_2} - \sum_{j=1}^{n}(S^{-1}y_j)_i\left\{(S^{-1}_{i,r_1})(V^TS^{-1}y_j)_{r_2} + (S^{-1}V)_{i,r_2}(S^{-1}y_j)_{r_1}\right\}$$

and

$$\begin{aligned}\frac{\partial^2 \log L_0(\Psi)}{\partial V_{r_1,r_2}\partial V_{r_3,r_4}} &= -n\left\{S^{-1}_{r_1,r_3}I_{r_2=r_4}\right\} + n\left(S^{-1}\right)_{r_1,r_3}\left(V^TS^{-1}V\right)_{r_2,r_4} + n\left(S^{-1}V\right)_{r_1,r_4}\left(S^{-1}V\right)_{r_2,r_3} \\ &\quad -\sum_{j=1}^{n}\left\{\left(y_i^TS^{-1}\right)_{r_3}\left(V^TS^{-1}\right)_{r_1,r_4} + \left(y_i^TS^{-1}V\right)_{r_4}\left(S^{-1}\right)_{r_1,r_3}\right\}\cdot \\ &\quad \left\{\left(y_j^TS^{-1}\right)_{r_3}I_{r_2=r_4} + \left(y_j^TS^{-1}\right)_{r_3}\left(V^TS^{-1}V\right)_{r_2,r_4} + \left(y_j^TS^{-1}V\right)_{r_4}\left(S^{-1}V\right)_{r_2,r_3}\right\}\end{aligned}$$

Denote $Q \triangleq S^{-1}V$ and $p \triangleq S^{-1}y_j$

$$\begin{aligned}
\frac{\partial^2 \log L_0(\Psi)}{\partial d_i^2} &= \frac{n}{2}S_{i,i}^{-2} - \sum_{j=1}^{n} S_{i,i}^{-1}p_i^2 \\
\frac{\partial^2 \log L_0(\Psi)}{\partial d_i \partial V_{r_1,r_2}} &= n(S^{-1}V)_{i,r_2} - \sum_{j=1}^{n} p_i \left\{(S_{i,r_1}^{-1})(V^T p)_{r_2} + Q_{i,r_2}p_{r_1}\right\} \\
\frac{\partial^2 \log L_0(\Psi)}{\partial V_{r_1,r_2}\partial V_{r_3,r_4}} &= -n\left\{S_{r_1,r_3}^{-1}I_{r_2=r_4}\right\} + n\left(S^{-1}\right)_{r_1,r_3}\left(V^T S^{-1}V\right)_{r_2,r_4} + nQ_{r_1,r_4}Q_{r_2,r_3} \\
&\quad -\sum_{j=1}^{n}\left\{p_{r_3}Q_{r_4,r_1} + \left(y_i^T S^{-1}V\right)_{r_4}\left(S^{-1}\right)_{r_1,r_3}\right\}
\end{aligned}$$

and

$$I_0(\Psi) = \begin{bmatrix} \frac{\partial^2 \log L_0(\Psi)}{\partial d_i^2} & \frac{\partial^2 \log L_0(\Psi)}{\partial d_i \partial V_{r_1,r_2}}0 \\ \frac{\partial^2 \log L_0(\Psi)}{\partial d_i \partial V_{r_1,r_2}} & \frac{\partial^2 \log L_0(\Psi)}{\partial V_{r_1,r_2}\partial V_{r_3,r_4}} \end{bmatrix}$$

3.4.6 Modification of the EM algorithm by Using Jeffrey's Prior

Assuming the dimension of V is $t \times k$

Since

$$\mathcal{I}_c(\Psi) = \begin{bmatrix} \frac{n}{2d_i^2} & 0 \\ 0^T & diag\left(\frac{n}{d_{r_1}}\right) \end{bmatrix}$$

Jeffrey's Prior is

$$\begin{aligned}
J(\Psi) &= \det\left(\mathcal{I}_c(\Psi)\right)^{\frac{1}{2}} \\
&= \frac{n^{t(k+1)}}{2^t}\prod_{i=1}^{t} d_i^{k+2}
\end{aligned}$$

and

$$\log J(\Psi) \propto \sum_{i=1}^{t}(k+2)\log d_i$$

Therefore, we can modify the EM factor model by applying Jeffrey's Prior,

Notice the updating formula for V is the same since the Jeffrey's Prior doesn't has V term. The updating formula for D is modified as

$$\frac{\partial E[L + \log J(\Psi)]}{D_{i,i}^{-1}} = -\frac{N}{2}D_{i,i} - \frac{N}{2}C_{i,i} + \frac{N}{2}(C\delta^T V^T)_{i,i} + (k+2)D_{i,i}$$

Therefore by setting the right side equals to zero,

$$D_{i,i} = \frac{1}{1 - \frac{2(k+2)}{N}} (C - V\delta C)_{i,i} \tag{28}$$

Comparing the new updating formula for D, it can be seen that it is actually amplify D when the sample size is small or the factor numbers are big through the inverse of $1 - \frac{2(k+2)}{N}$.

It is a reasonable choice for estimating the covariance under small sample situations, because the smallest eigenvalue corresponds to the underestimated risk.

Under a simulation using 40 instruments and 10 factors, when the sample size is close to its dimension, the Jeffrey's Prior modified EM algorithm gives a significantly higher value for the minimum eigenvalues.

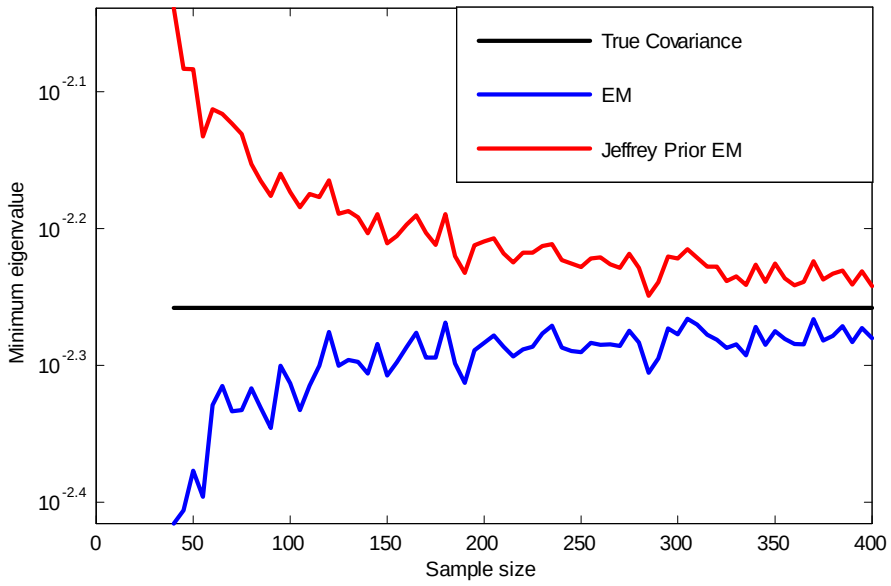


Figure 24. The Minimum Eigenvalue for Estimated Covariance.

3.5 The EH Model Approach for Covariance Estimation

3.5.1 Extra information in factor structure

Splitting the factor loading matrix V as $\begin{bmatrix} V_1 & ... & V_n \end{bmatrix}$, and write the factor formula as

$$\begin{aligned} y &= Vx + \varepsilon \\ &= \begin{bmatrix} V_1 & ... & V_k \end{bmatrix} x + \varepsilon \\ &= V_1 x_1 + ...V_k x_k + \varepsilon \end{aligned} \tag{29}$$

if we denote y_t as the return vector on t-th day, and we denote the market value for instrument i as $a_{t,i}$, and the capital flow in or out of i-th instrument as $\Delta a_{t,i}$, then

$$\Delta a_t = a_t \circ y_t \tag{30}$$

where $\circ$ is element-wise product, and therefore

$$\begin{aligned} \Delta a_t &= a_t \circ (V_1 x_1 + ...V_k x_k) + a_t \circ \varepsilon \\ &= (V_1 \circ a_t)x_1 + ... + (V_k \circ a_t)x_k + a_t \circ \varepsilon \\ &\triangleq f_1 x_1 + ... + f_k x_k + a_t \circ \varepsilon \qquad (31) \\ &\triangleq \Delta f_1 + ... + \Delta f_k + a_t \circ \varepsilon \qquad (32) \end{aligned}$$

Δf_i has the unit as capital (\$) times return (%), and reflects how much capital flow in and out of the market acting through the i-th factor.

The 32 formula explains the relationship between factor model and the capital flow: the capital flow on instruments is a results of capital flow through factors. This makes sense especially for nowadays quantitative investing participants market: the information that lead to pricing behavior is usually touching a basket of instruments which has features in common. For example, in low interest rate environment, mutual fund portfolio manager would seek to long yield generating stocks in order to obtain a stable and low risk cash flow. Another point for the capital flow formula is that, it provides a perspective to understand the forming of factors : each column of factor loadings is measuring one common feature for different instrument. This perspective would help bring more information into factor structure.

In Rubin's hidden factor model, the factor information is assumed to be always completely unknown, which is not the case. More often, we can assert reasonably certain facts about factor loadings, even without having a completely idea of it. For example, in the previous example of market with yield investing behavior, it is reasonable to assume one factor loading would have the yield for each instrument as the coefficient that,

$$V_{yield} = c_{yield}[v_1, ...v_d]^T$$

where v_i is the annual yield for i-th instrument and c_{yield} is an unknown parameter. In this way, we split the factor into two parts : we claim we understand the relationship between the i-th and j-th instruments and we can get the ratio of $\frac{v_i}{v_j}$ correct; we dont' know the magnitude of the whole factor c_{yield}. Here c_{yield} acts like the capital performance variable, which reflects how much bet would be put on this yield investment behavior. Besides this example, the CAPM or Fama-French model provides empirical way to build factor loadings using the market size. It is also reasonable to bring sector information into factor construction. In total, these particular factor construction method relates to seperate the information known and information unknown, and combine them through a constant vector and an unkown performance scalar variable.

There are several reasons that we want to bring these information into factor structure. First of all, it can further help reduce the number of factors. Although the factor model has reduce the parameters from sample covariance $C_{d\times d}$ to $\{V_{d\times k},\ D_d\}$. However, still these are a lot of parameters, which would have a high estimation variances. Second, when we try to estimate that amount of parameters, we have to use long history of data, in that case, we may lose the time efficiency and the new rise factors. By saying time efficiency, we mean that the market is not time-invariant, and when market start to change, we want to estimate how and to which direction it changes into by using a relative recent time window. By predictively construct factor loading in our way, we require a much less samples to have a good estimate and can better catch the new factors.

To expand the factor structure, first we can write factor loading matrix as

$$V \triangleq \begin{bmatrix} V_e & V_h \end{bmatrix} \tag{33}$$

where V_e is empirical factor loading matrix where we know partial information, and V_h is the hidden factor loading matrix where we don't know anything. V_h is the factor loading

matrix in original EM factor algorithm.

Next, we assume the empirical factor loading matrix V_e has the structure

$$V_e = Uc \tag{34}$$

where c is an unknown diagonal matrix $\begin{bmatrix} c_1 & & \\ & ... & \\ & & c_p \end{bmatrix}$, and U is a constant matrix.

In this way we split V into two parts: the empirical factors Uc, and the hidden factors $\widetilde{V}$. For the empirical factor, U is the constant matrix, which reflects the relative relationship inside each factor, while c is the parameter to be estimated, which reflected the magnitude of return associated with this factor. As a result, we have our factor structure as

$$\begin{aligned} y &= Vx + \varepsilon \\ V &= [Uc, \widetilde{V}] \end{aligned} \tag{35}$$

where the dimensions are

$$\begin{aligned} y &: d \times 1 \\ x &: (p+k) \times 1 \\ \varepsilon &: d \times 1 \\ U &: d \times p \\ c &: p \times p \\ \widetilde{V} &: d \times k \\ V &: d \times (p+k) \end{aligned}$$

In next section, we discuss how to implement EM algorithm on this framework. We would call our model as EH Factor Model(Empirical and Hidden factor model).

3.5.2 The EM algorithm for the EH Factor model

When using EM algorithm to solve for the update equations fro EH model, the challenge is taking the derivative of $E[L]$. Unlike statistical factor model, we can not directly apply block matrix calculus rules to $\begin{bmatrix} Uc & \widetilde{V} \end{bmatrix}$. Instead, we need to calculate the derivative of $E[L]$

w.r.t arbitrary element of V, and put the elements together to form $\frac{\partial E[L]}{\partial V}$. In other words, denoting X_{rs} as the (r,s) position element of X matrix, the definition of matrix calculus is directly applied in order to calculate the matrix derivative, that $\frac{\partial E[L]}{\partial V} = M$ where the $M_{rs} = \frac{\partial E[L]}{\partial V rs}$. Also notice that $Tr\left\{(\frac{\partial \begin{bmatrix} Uc & \widetilde{V} \end{bmatrix}}{\partial \widetilde{V}_{rs}})^T D^{-1} V\right\} = Tr\left\{V^T D^{-1} (\frac{\partial \begin{bmatrix} Uc & \widetilde{V} \end{bmatrix}}{\partial \widetilde{V}_{rs}})\right\}$, therefore

$$\begin{aligned} \frac{\partial E[L]}{\partial \widetilde{V}_{rs}} &= \sum_{i=1}^{n} E[x_i|y_i]^T \left(\frac{\partial \begin{bmatrix} Uc & \widetilde{V} \end{bmatrix}}{\partial \widetilde{V}_{rs}}\right)^T D^{-1} y_i \\ &\quad - \sum_{i=1}^{n} Tr\left\{\left(\frac{\partial \begin{bmatrix} Uc & \widetilde{V} \end{bmatrix}}{\partial \widetilde{V}_{rs}}\right)^T D^{-1} V E[x_i x_i^T|y_i]\right\} \end{aligned} \tag{36}$$

This is obtained by using $\frac{\partial Tr(AXB)}{\partial X_{rs}} = Tr(A\frac{\partial X}{\partial X_{rs}}B)$. The trick is that since $\frac{\partial X}{\partial X_{rs}}$ is a matrix that the (i,j) entry is $\frac{\partial X_{i,j}}{\partial Xr,s}$, therefore $\frac{\partial X}{\partial X_{rs}} = E_{rs}$ where E_{rs} is a matrix that is all 0 except on the (r,s) position where it is 1. Further E_{rs} can be written as $e_r e_s^T$ where e_r is a vector has 0 for all positions except has 1 on the r-th position. Therefore $\frac{\partial Tr(AXB)}{\partial X_{rs}} = Tr(AE_{rs}B) = Tr(Ae_r e_s^T B) = Tr(e_s^T BAe_r) = (BA)_{rs}$

Similarly, to calculate $Tr\left\{(\frac{\partial \begin{bmatrix} Uc & \widetilde{V} \end{bmatrix}}{\partial \widetilde{V}_{rs}})^T D^{-1} V\right\}$ it only requires to realize that $\frac{\partial \begin{bmatrix} Uc & \widetilde{V} \end{bmatrix}}{\partial \widetilde{V}_{rs}} = \begin{bmatrix} 0_{d\times p} & E_{rs} \end{bmatrix}$ where it contains a zero block submatrix because Uc is not a function of $\widetilde{V}$. Applying the tricks introduced above, it can be obtained that

$$\begin{aligned} \frac{\partial E[L]}{\partial \widetilde{V}_{rs}} &= \sum_{i=1}^{n} E[x_i|y_i]^T \begin{bmatrix} 0^T \\ E_{sr} \end{bmatrix} D^{-1} y_i \\ &\quad - \sum_{i=1}^{n} Tr\left\{\begin{bmatrix} 0^T \\ E_{sr} \end{bmatrix} D^{-1} V E[x_i x_i^T|y_i]\right\} \end{aligned} \tag{37}$$

Consider the first part on the right side of 37.

$$\begin{aligned}
\sum_{i=1}^{n} E[x_i|y_i]^T \begin{bmatrix} 0^T \\ E_{sr} \end{bmatrix} D^{-1} y_i &= \sum_{i=1}^{n} y_i^T D^{-1} \begin{bmatrix} 0 & E_{rs} \end{bmatrix} E[x_i|y_i] \\
&= \sum_{i=1}^{n} y_i^T D^{-1} \begin{bmatrix} 0 & E_{rs} \end{bmatrix} \begin{bmatrix} \alpha_i \\ A_i \end{bmatrix} \\
&= \sum_{i=1}^{n} y_i^T D^{-1} E_{rs} A_i \\
&= \sum_{i=1}^{n} (A_i y_i^T D^{-1})_{sr} \qquad (38)
\end{aligned}$$

where $E[x_i|y_i]$ is splited into two parts $E[x_i|y_i] \triangleq \begin{bmatrix} \alpha_i \\ A_i \end{bmatrix}$,with the shape that α_i is $p \times 1$ and A_i is $k \times 1$

Apply the trace trick on the second part of 37, similar approach leads to

$$\begin{aligned}
-\sum_{i=1}^{n} Tr \left\{ \begin{bmatrix} 0^T \\ E_{sr} \end{bmatrix} D^{-1} V E[x_i x_i^T|y_i] \right\} &= -\sum_{i=1}^{n} Tr \left\{ D^{-1} V E[x_i x_i^T|y_i] \begin{bmatrix} 0^T \\ E_{sr} \end{bmatrix} \right\} \\
&= -\sum_{i=1}^{n} Tr \left\{ D^{-1} V \begin{bmatrix} (K_i)_{12} \\ (K_i)_{22} \end{bmatrix} E_{sr} \right\} \qquad (39) \\
&= -\sum_{i=1}^{n} \left(D^{-1} V \begin{bmatrix} (K_i)_{12} \\ (K_i)_{22} \end{bmatrix} \right)_{rs} \qquad (40)
\end{aligned}$$

where $E[x_i x_i^T|y_i] \triangleq \begin{bmatrix} (K_i)_{11} & (K_i)_{12} \\ (K_i)_{21} & (K_i)_{22} \end{bmatrix}$. The dimension for the four subblock is: $(K_i)_{11}$ is $p \times p$, $(K_i)_{12}$ is $p \times k$, $(K_i)_{21}$ is $k \times p$, $(K_i)_{22}$ is $k \times k$.

Therefore,

$$\frac{\partial E[L]}{\partial \widetilde{V}_{rs}} = \sum_{i=1}^{n} \left(D^{-1} y_i A_i^T - D^{-1} V \begin{bmatrix} (K_i)_{12} \\ (K_i)_{22} \end{bmatrix} \right)_{rs} \qquad (41)$$

and

$$\frac{\partial E[L]}{\partial \widetilde{V}} = \sum_{i=1}^{n} \left(D^{-1} y_i A_i^T - D^{-1} V \begin{bmatrix} (K_i)_{12} \\ (K_i)_{22} \end{bmatrix} \right) \qquad (42)$$

Set the derivative to zero, cancel D^{-1} on both sides and plug back $V = \begin{bmatrix} Uc & \widetilde{V} \end{bmatrix}$, the first update equation is obatined as

$$Uc \sum_{i=1}^{n} \left((K_i)_{12} + \widetilde{V} \sum_{i=1}^{n} (K_i)_{22} \right) = \sum_{i=1}^{n} y_i A_i^T \tag{43}$$

and

$$\widetilde{V} = \left(\sum_{i=1}^{n} y_i A_i^T - Uc \sum_{i=1}^{n} (K_i)_{12} \right) \left(\sum_{i=1}^{n} (K_i)_{22} \right)^{-1} \tag{44}$$

Next we take derivative of $E[L]$ with respect to c. Since c is diagonal matrix, we can use c_j to denote the (j, j) position of c. Note that Uc can be write as $Uc = \begin{bmatrix} u_1 c_1 & ... & u_p c_p \end{bmatrix}$ where u_i is the i-th column of U. The same argument in early steps gives

$$\frac{\partial \begin{bmatrix} Uc & \widetilde{V} \end{bmatrix}}{\partial c_j} = \begin{bmatrix} \widetilde{U}_j & 0 \end{bmatrix} \tag{45}$$

where $\widetilde{U}_j = \begin{bmatrix} 0 & .. & 0 & u_j & 0 & .. & 0 \end{bmatrix}$

Following a similar approach as $\frac{\partial E[L]}{\partial V_{rs}}$, it leads to

$$\frac{\partial E[L]}{\partial c_j} = (\alpha_i)_j u_j^T D^{-1} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} u_j^T D^{-1} V \begin{bmatrix} (K_i)_{11} \\ (K_i)_{21} \end{bmatrix}_{(:,j)} \tag{46}$$

where $(\alpha_i)_j$ is j-th element of α_i. and $\begin{bmatrix} (K_i)_{11} \\ (K_i)_{21} \end{bmatrix}_{(:,j)}$ denotes the j-th column of $\begin{bmatrix} (K_i)_{11} \\ (K_i)_{21} \end{bmatrix}$.

Set the above equation to zero, we get

$$u_j^T D^{-1} \sum_{i=1}^{n} (\alpha_i)_j y_i = u_j^T D^{-1} \left\{ Uc \sum_{i=1}^{n} [(K_i)_{11}]_{(:,j)} + \widetilde{V} \sum_{i=1}^{n} [(K_i)_{21}]_{(:,j)} \right\} \tag{47}$$

Define six notations for expression simplicity:

$$
\begin{aligned}
M &\triangleq \{\sum_{i=1}^{n}[(K_i)_{11}] - \sum_{i=1}^{n}(K_i)_{12}(\sum_{i=1}^{n}(K_i)_{22})^{-1}\sum_{i=1}^{n}[(K_i)_{21}]\} \\
M_j &\triangleq \{\sum_{i=1}^{n}[(K_i)_{11}] - \sum_{i=1}^{n}(K_i)_{12}(\sum_{i=1}^{n}(K_i)_{22})^{-1}\sum_{i=1}^{n}[(K_i)_{21}]\}_{(:,j)} \\
\widetilde{u} &\triangleq U^T D^{-1} U \\
\widetilde{u}_j &\triangleq u_j^T D^{-1} U \\
m_j &\triangleq u_j^T D^{-1}\left\{\sum_{i=1}^{n}(\alpha_i)_j y_i - \sum_{i=1}^{n} y_i A_i^T (\sum_{i=1}^{n}(K_i)_{22})^{-1}\sum_{i=1}^{n}[(K_i)_{21}]_{(:,j)}\right\} \\
m &\triangleq \begin{bmatrix} m_1 & ... & m_n \end{bmatrix}^T
\end{aligned}
$$

and substitue 44 into 47, the result can be put into a clean form that

$$\widetilde{u}_j c M_j = m_j \tag{48}$$

Where the dimensions are: $\widetilde{u}_j$ is $1 \times p$, M_j is $p \times 1$, and m_j is 1×1. To use the diagonal structure of c, write $\widetilde{u}_j c M_j$ as the linear combination with c_i as coefficients, that is

$$
\begin{aligned}
\widetilde{u}_j c M_j &= \sum_{q=1}^{p}(\widetilde{u}_j)_q c_q (M_j)_q \\
&= \begin{bmatrix}(\widetilde{u}_j)_1(M_j)_1 & ... & (\widetilde{u}_j)_p(M_j)_p\end{bmatrix}\begin{bmatrix} c_1 \\ ... \\ c_p \end{bmatrix} \\
&= (\widetilde{u}_j \circ M_j)\widetilde{c}
\end{aligned} \tag{49}
$$

where $\circ$ is element-wise product, and $\widetilde{c}$ is the take the diagonal part of c and put it as a vector, that $\widetilde{c} = vec(c)$.

Since $j = 1, .., p$, there are p equations in total which forms a full rank linear equation system, and $\widetilde{c}$ can be solved as

$$\widetilde{c} = (\widetilde{u} \circ M)^{-1} m \tag{50}$$

Finally, reform c from $\widetilde{c}$, and substitute 50 back to 44, we get the update formula for $\widetilde{V}$.

Then we solve for the update equation for D matrix, which is easier and similar to the corresponding step in the statsitical factor model. The trick is to use the formula

$\frac{\partial \lg\det X}{\partial X} = X^T$, and we get

$$\frac{\partial E[L]}{\partial D^{-1}} = \frac{1}{2}nD - \frac{1}{2}nC + Vn\delta C - \frac{1}{2}V(n\Delta + \delta nC\delta^T)V^T \tag{51}$$

Set the gradient to zero, and immediately it can be seen that

$$D = diag(C - 2V\delta C + V(\Delta + \delta C\delta^T)V^T) \tag{52}$$

where $diag(.)$ operator forms a diagonal matrix based on the diagonal element of the input.

At last, notice

$$\sum_{i=1}^{n}[(K_i)_{11}] = \sum_{i=1}^{n}(\Delta + \delta y_i y_i^T \delta^T)_{(1:p,1:p)} \tag{53}$$

and

$$\begin{aligned} \sum_{i=1}^{n}(\alpha_i)_j y_i &= \sum_{i=1}^{n} y_i e_j^T \delta y_i \\ &= nC\delta^T e_j \end{aligned} \tag{54}$$

and

$$\begin{aligned} \sum_{i=1}^{n} y_i A_i^T &= \sum_{i=1}^{n} y_i \left(\delta y_i\right)^T_{(p+1:p+k)} \\ &= (nC\delta^T)_{(:,p+1:p+k)} \end{aligned} \tag{55}$$

, combine 44, 50 and 52 up and plug back the sufficient statistics, the clean version of the updating equations are

$$
\begin{aligned}
K_{11} &= n(\Delta + \delta C \delta^T)_{(1:p,1:p)} \\
K_{12} &= n(\Delta + \delta C \delta^T)_{(1:p,p+1:p+k)} \\
K_{21} &= n(\Delta + \delta C \delta^T)_{(p+1:p+k,1:p)} \\
K_{22} &= n(\Delta + \delta C \delta^T)_{(p+1:p+k,p+1:p+k)} \\
M &= K_{11} - K_{12} K_{22}^{-1} K_{21} \\
\widetilde{u} &= U^T D^{-1} U \\
C^\delta &= C\delta^T \\
m &= diag\left\{ nU^T D^{-1} \left(C^\delta_{(:,1:p)} - \left(C^\delta\right)_{(:,p+1:p+k)} \left(K_{22}\right)^{-1} \left(K_{11}\right) \right) \right\} \\
c &= diag\left\{ \left(\widetilde{u} \circ M\right)^{-1} m \right\} \\
\widetilde{V} &= \left(C^\delta_{(:,p+1:p+k)} - UcK_{12} \right) \left(K_{22}\right)^{-1} \\
V &= \left[Uc \quad \widetilde{V} \right] \\
D &= diag\left(C - 2V\delta C + V\left(\Delta + \delta C \delta^T\right) V^T \right)
\end{aligned} \tag{56}
$$

Note that K_{11} is not the same as $(K_i)_{11}$ but a new symbol. The *diag* operator above represents both the transformation from vector to matrix and from matrix to vector according to the context. The above formula is the updating equations for the EM algorithm for the EH model.

3.5.3 An Alternative Approach by Two Step MLE

A nature thinking other than EH algorithm is a two stage MLE. In this way, it treat the non-empirical factor part as residual and proceed PCA or the hidden factor model on the residual part. That is

$$
\begin{aligned}
y &= \left[Uc \quad \widetilde{V} \right] x + \varepsilon \\
&= Ucx_1 + (\widetilde{V} x_2 + \varepsilon) \\
&\triangleq U\widetilde{x} + r
\end{aligned}
$$

where $\widetilde{x}$ is the factor follow $N(0, cc^T)$, and r contains residual information.

There are two alternative approaches for this estimation.

Approach 1:

Step 1:

Denote U by it's columns $\begin{bmatrix} u_1 & ... & u_k \end{bmatrix}$,

$$\begin{aligned} \Sigma &= Ucc^TU^T + \widetilde{V}\widetilde{V}^T + D \\ &= \Sigma_{i=1}^k u_i c_i^2 u_i^T + \widetilde{V}\widetilde{V}^T + D \\ vec(\Sigma) &= \Sigma_{i=1}^k vec(u_i c_i^2 u_i^T) + vec(\widetilde{V}\widetilde{V}^T + D) \\ &= \Sigma_{i=1}^k (u_i \otimes u_i) c_i^2 + residuals \\ &= \begin{bmatrix} u_1 \otimes u_1 & ... & u_k \otimes u_k \end{bmatrix} \begin{bmatrix} c_1^2 \\ ... \\ c_k^2 \end{bmatrix} + residuals \end{aligned}$$

Here it is assumed that the residuals is in the complement space of $\begin{bmatrix} u_1 \otimes u_1 & ... & u_k \otimes u_k \end{bmatrix}$. Therefore the c parameter is obtained through

$$\begin{bmatrix} c_1^2 \\ ... \\ c_k^2 \end{bmatrix} = \begin{bmatrix} u_1 \otimes u_1 & ... & u_k \otimes u_k \end{bmatrix}^+ vec(\Sigma) \tag{57}$$

In this step what we are doing is actually to maximize the function $L = L(c|data)$. Notice we don't follow the more common way of regressing which projects the observation y onto the $R(U)$ space, and find the points c_x so that $||Uc_x - y||_2$ is minimized. Because in this way, c is obtained by taking the diagonal part of sample covariance of c_x, that is $c = diag(\frac{c_x c_x^T}{N})$, during which step all information except diagonal is discarded. Instead, we use the above vectorization so that the projection is still onto a transformed space of $R(U)$ but the parameter here is only c.

Step 2:

In the second step, $\widetilde{V}$ and D is obtained by applying EM algorithm on the residuals, that is

$$\begin{aligned} \widetilde{V}, D &= \arg\max L(\widetilde{V}, D|c, data) \\ &= \arg\max L(\widetilde{V}, D|\Sigma - Ucc^TU^T) \end{aligned}$$

Obviously, there is no guarantee that the two step maximum likelihood would give the same estimation of the parameters, that

$$\begin{aligned} \arg\max L(\widetilde{V}, D|\Sigma - Uc_0) &= \arg\max L(\widetilde{V}, D, c|\Sigma) \\ \text{where } c_0 &= \arg\max L(c|\Sigma) \end{aligned}$$

and fundamentally the two step maximum likelihood does not necessary to achieve the global maximum likelihood.

$$\max L(\widetilde{V}, D, c|data) = \max L(\widetilde{V}, D|c, data) \text{ where } c = \arg\max L(c|data)$$

In the next simulation, we would include simulation that shows clear evidence that the EH model out performs the two step maximum likelihood in most situations.

Approach 2:

In Approach 1, the main drawback is in 57, the projection matrix is a fat matrix so that c_i suffers numerical problem. The merit is that by taking kronecker product, the only variables is the diagonal part of c matrix. As alternative approach, we compared using a direct regression method, that

$$\begin{aligned} (cx) &= U^+ y \\ c &= diag((cx)(cx)^T) \end{aligned}$$

The drawback for this method is that $(cx)(cx)^T$ is not necessarily diagonal matrix so that in theory the c is just an approximation.

In out test, we find that Approach 2 is better in the sense of less estimation errors, as shown in Figure . In our later test we would apply the second approach.

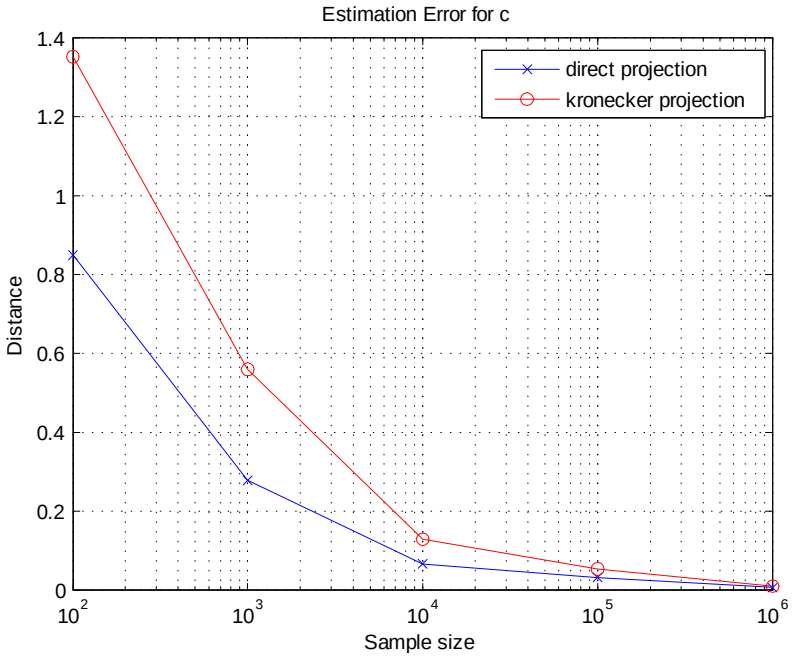


Figure 25. Compare of Two Different Projection Methods.

3.6 Simulation and Performance of the EH Model

3.6.1 Compare of the EH Model and the Statistical Factor Model

In our experiment, we use $V_{20\times 10}$ and $D_{20\times 20}$ as the true market. With each $N = 20, ..., 200$, we run the factor model for 100 times and each time with 1000 iterations.

First we constrain the size of $\bar{V}$ to be 20×10. Then we varies the number of empirical factors p from 0 to 10, and the number of hidden factors k would follow $k = 20 - p$. Note that when $p = 0$, it is the original hidden factor model.

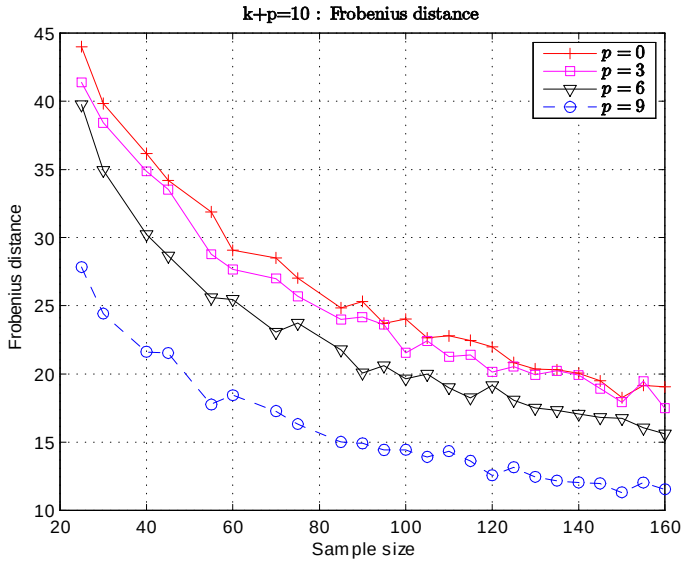


Figure 26.

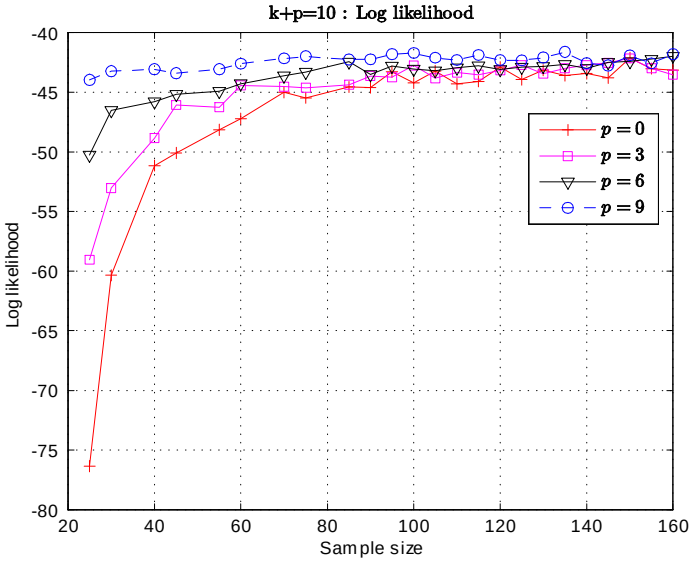


Figure 27.

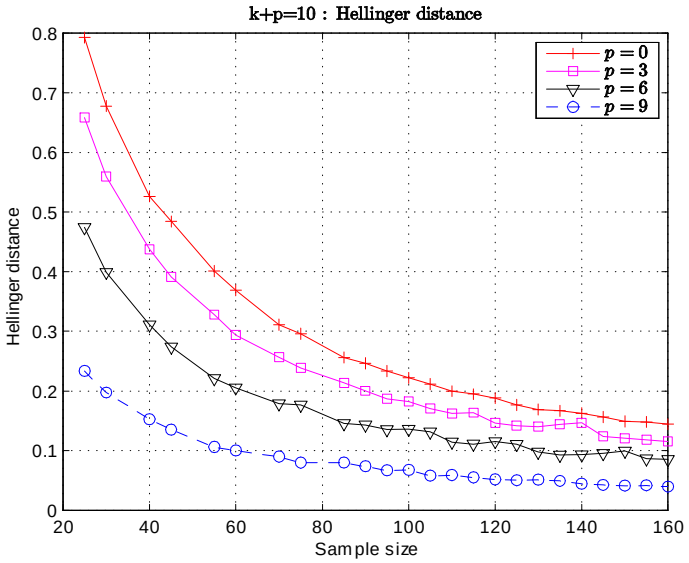


Figure 28.

As shown in Figure 2, Figure 3, and Figure 4, in all different measures, when more prior structure information($p = 9$) is provided through empirical factor loading U, the distance to true covariance is smaller than using only hidden factors($p = 0$). These effects become more significant as the sample size decrease from $N = 200$, that sample size is ten times of dimension d, to $N = 20$, where sample size is on the same size of d. Also it is noticeable that is Hellinger distance is always a better measure in our experiment, that it is smoother than the Frobenius norm as N increases, and is better than log likelihood since at larger N the log likelihood can not discriminate different p value performance. In this simulation, it proves how significant improvement that partial information can brings to the estimation, especially when the sample size is limited in practical situation.

3.6.2 Imperfection Ratio Test

One problem for empirical factor is that it is not guaranteed to be correct. Therefore we test the situations when the empirical factors has a built in noise, measured by an "imperfection ratio" α, that is

$$U_1 = U + \alpha E \tag{58}$$

where E is a random matrix with $N(0,1)$ entries and the same size as U, and instead of $V = [Uc, \widetilde{V}]$, we use $V_1 = [U_1 c, \widetilde{V}]$ and apply the EH model to estimate $(c, \widetilde{V})$. We calculate the difference between log likelihood from the EH model, $L(V_{EH}, D_{EH})$, and the log likelihood of the complete hidden factor model, $L(V_{Rubin}, D_{Rubin})$. We test with $\alpha \in [0, 0.5]$ and $N \in [20, 200]$. We fix the total factor $T = 10$, and the empirical factor $p \in [0, 10]$ and the hidden factor $k = T - p$. The result shows that a large region of the log likelihood difference is positive. The first interesting point is that the region where EH model does better do not depend on the number of empirical factors, as the zero contour level stays similar at different p level as shown in Figure 5. When the sample size $N \leq 50 \approx 2d$ (the green and yellow region in Figure 5, the EH model always beats Rubin model. Similarly, when $\alpha \in [0, 0.1]$, the EH model always win. And once again, these two regions stays unchanged no matter what p value is. This shows that EH model's advantage is at the region when the data is highly insufficient or the empirical factor has a high quality. Despite these two regions where EH dominates, the EH model has a region where it beats the complete hidden factor model with small sample size and proper imperfection ratio, and vice versa. The difference regarding different p level is the sharpness that how good the EH model or the Rubin model is at their winning region. Figure 6 draws the Hellinger distance difference $H(V_{EH}, D_{EH}) - H(V_{Rubin}, D_{Rubin})$. The Hellinger distance tells similar results : the advantage of the EH model lies at the region where imperfect ratio is low and the sample size is small.

,

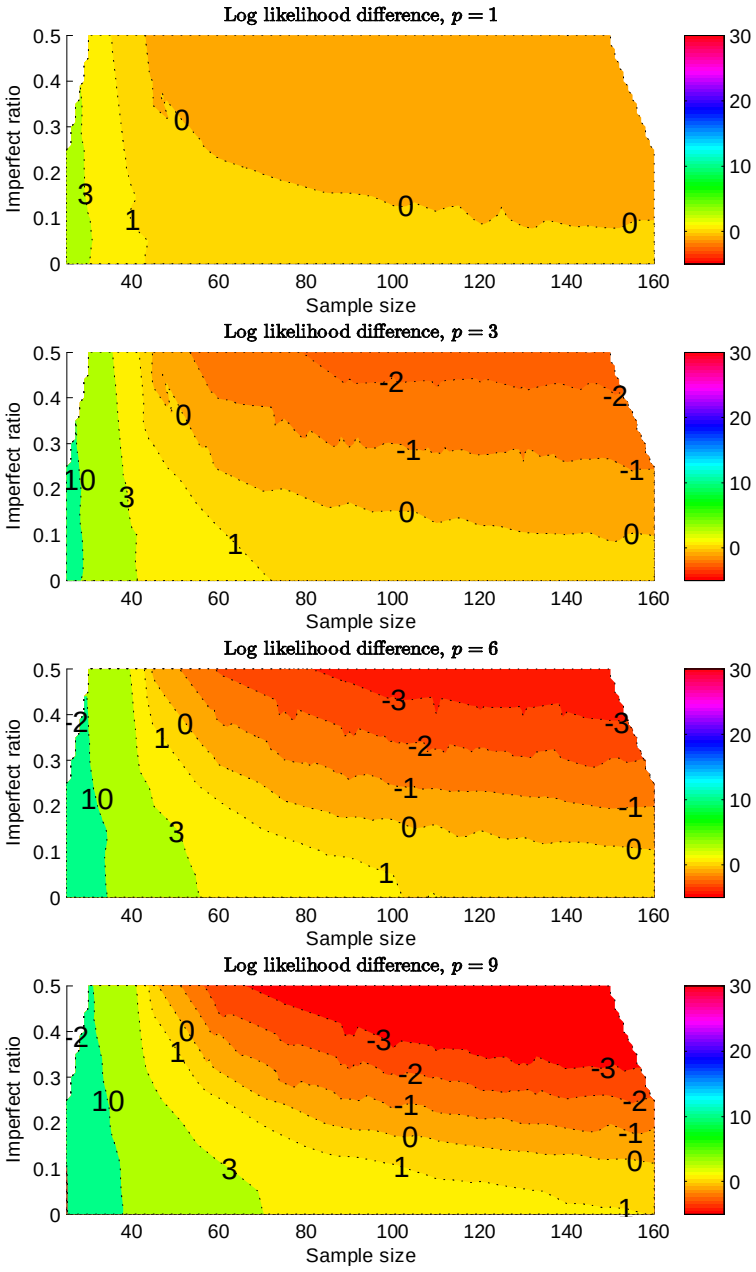

Log likelihood difference, $p = 1$
Log likelihood difference, $p = 3$
Log likelihood difference, $p = 6$
Log likelihood difference, $p = 9$
Imperfect ratio
Sample size
0.5
0.4
0.3
0.2
0.1
0
40
60
80
100
120
140
160
30
20
10
0

Figure 29.

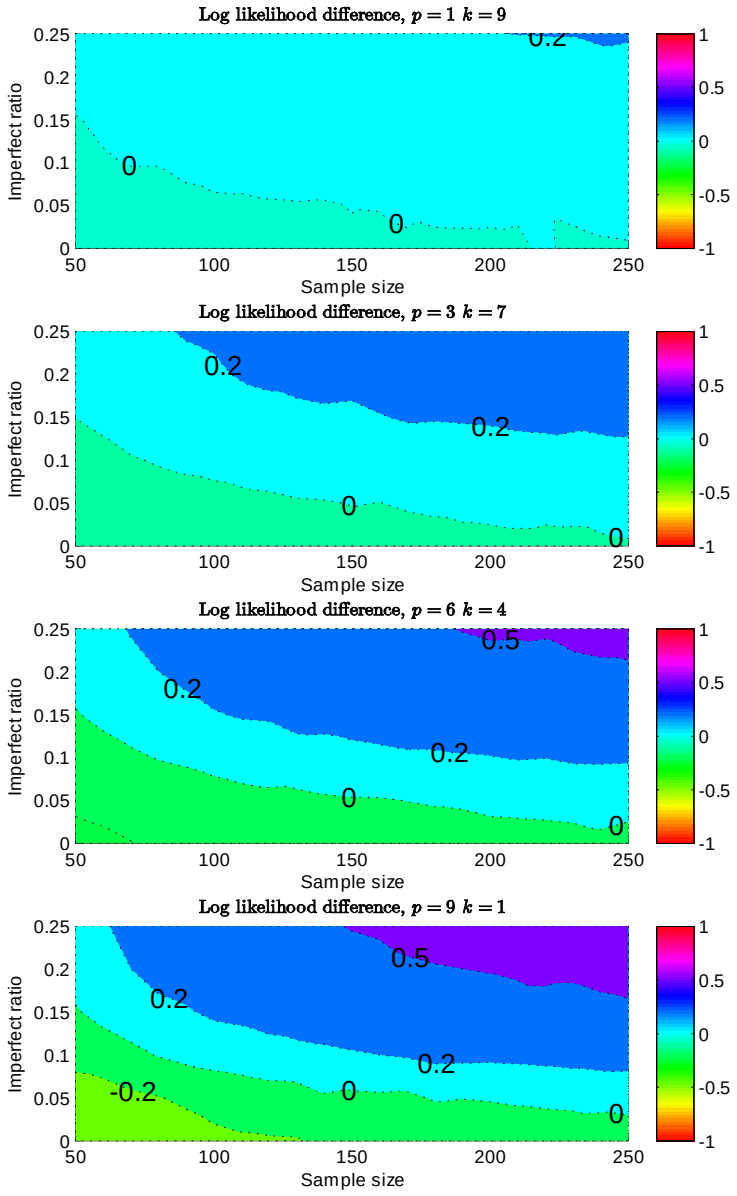


Figure 30.

More often, people would want to know that what would be the case if add empirical factors while keep the original hidden factor number unchanged. We did the experiment

by fixing $k = 10$, and $p \in [0, 10]$. Shown in Figure 6. the hidden factor number curve stays at similar level with increasing p number. In this case, bringing empirical factors have more benefits than the loss : the lower left part where the loglikelihood increases much more rapidly from 1 to 30 as p increases, compared with the right corner where the value decreases from 0 to -3. The Hellinger distance difference graph shows similar conclusion, compare Figure 8 with Figure 6, the curvature at the upper right corner becomes more smooth, while the benefits in the lower left corner stays similar.

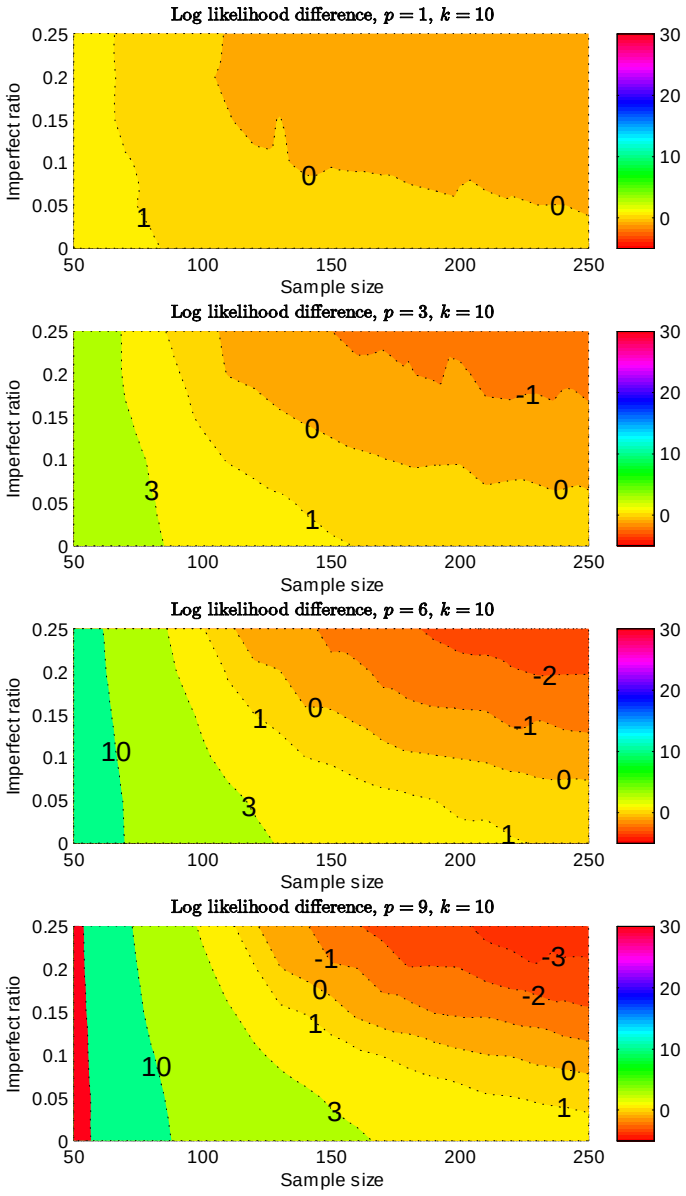
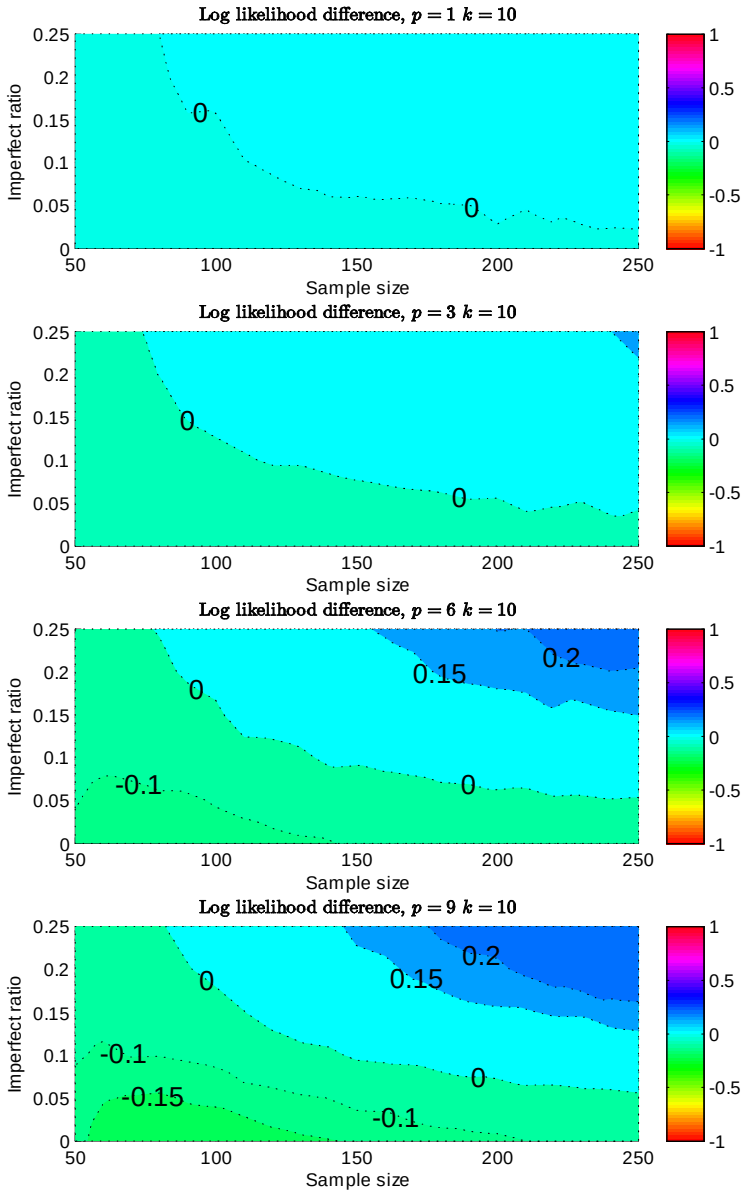


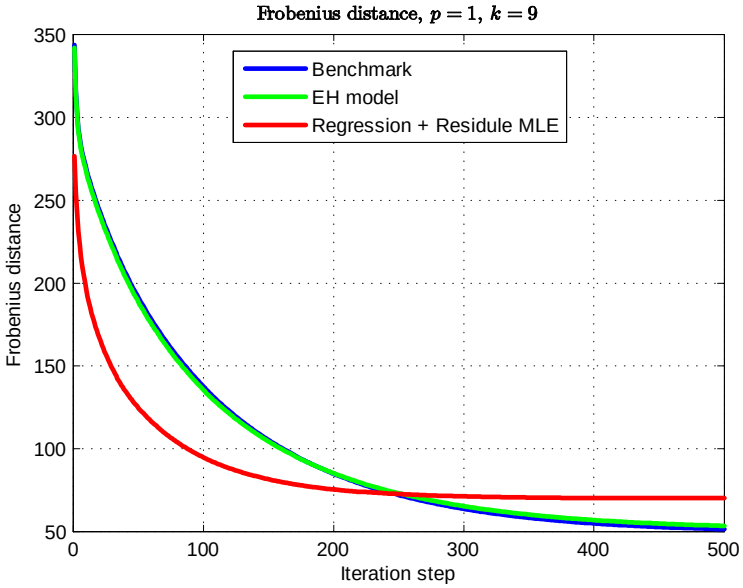
Figure 31.

Figure 32.


Log likelihood difference, $p = 1$ $k = 10$
Imperfect ratio
Sample size
0
0
Log likelihood difference, $p = 3$ $k = 10$
Imperfect ratio
Sample size
0
0
Log likelihood difference, $p = 6$ $k = 10$
Imperfect ratio
Sample size
0
0.15
0.2
-0.1
0
Log likelihood difference, $p = 9$ $k = 10$
Imperfect ratio
Sample size
0
0.15
0.2
-0.1
-0.15
0
-0.1

3.7 Compare of the EH Model and Two Step MLE

To compare the performance of the EH model, which does one step likelihood maximization, with the two step MLE, which did first regression on the empirical factors and the second step on the residuals.


Frobenius distance, $p = 1$, $k = 9$
Benchmark
EH model
Regression + Residule MLE
Frobenius distance
Iteration step
350
300
250
200
150
100
50
0
100
200
300
400
500

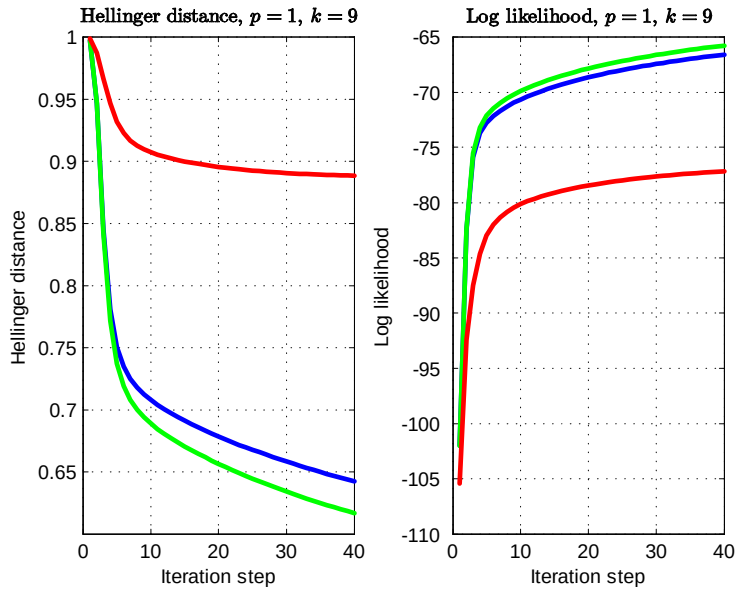
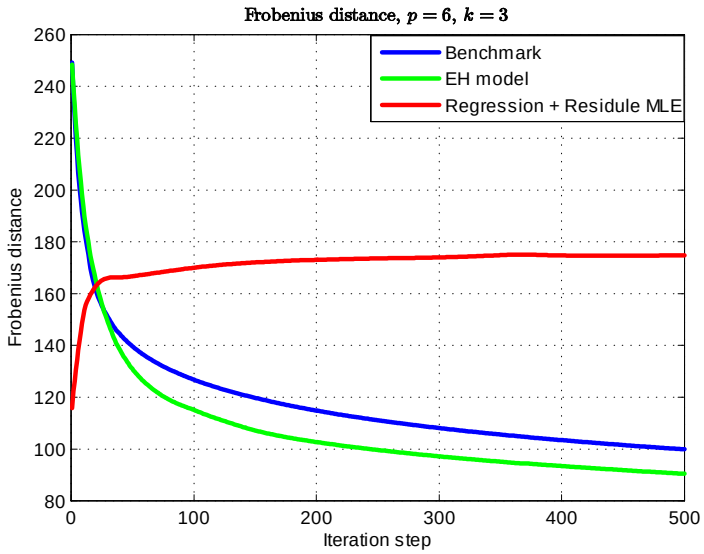

Hellinger distance, $p = 1$, $k = 9$
Hellinger distance
Iteration step
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0
10
20
30
40
Log likelihood, $p = 1$, $k = 9$
Log likelihood
Iteration step
-65
-70
-75
-80
-85
-90
-95
-100
-105
-110
0
10
20
30
40

Figure 33


Frobenius distance, $p = 6$, $k = 3$
Benchmark
EH model
Regression + Residule MLE
Frobenius distance
Iteration step
260
240
220
200
180
160
140
120
100
80
0
100
200
300
400
500

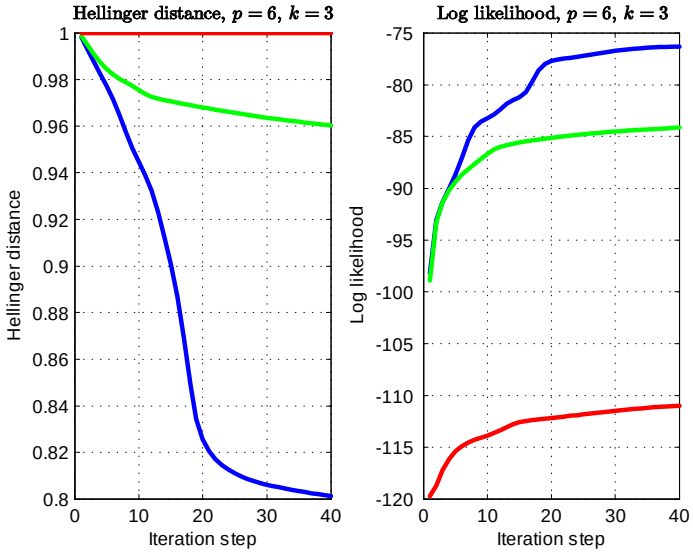

Hellinger distance, $p = 6$, $k = 3$
Hellinger distance
Iteration step
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0.8
0
10
20
30
40
Log likelihood, $p = 6$, $k = 3$
Log likelihood
Iteration step
-75
-80
-85
-90
-95
-100
-105
-110
-115
-120
0
10
20
30
40

Figure 34.

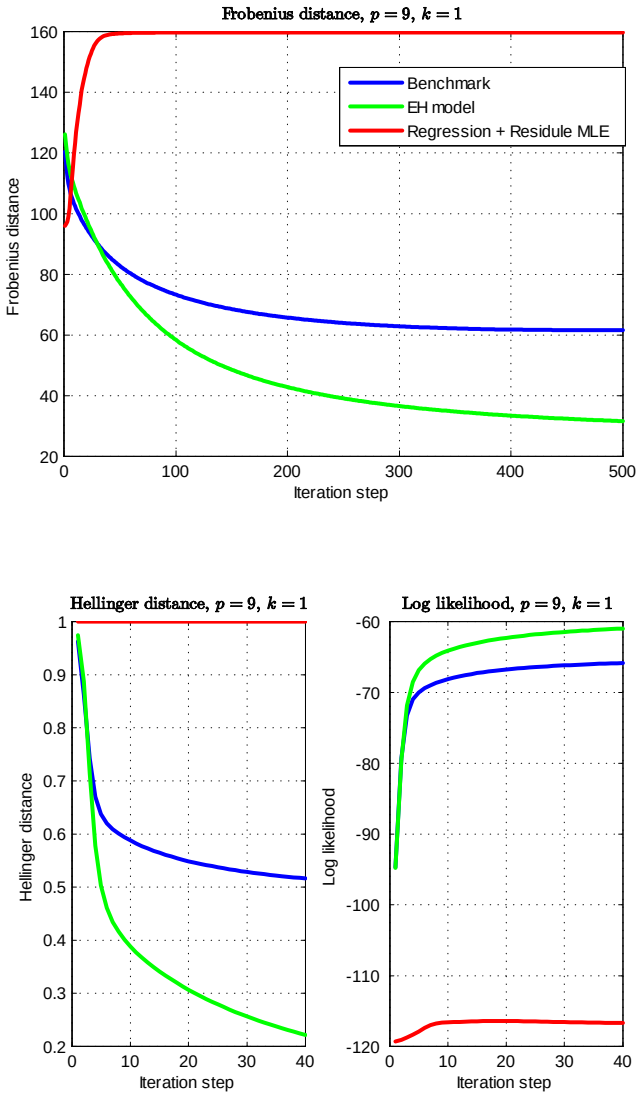


Figure 35.

As can be seen in Figure 33. to Figure 35., as the number of empirical factors increase, the estimation error becomes dominating in the two step MLE method.

We illustrate the difference using the 25% and 75% quantile boxplot. We run simulation with empirical factor p from 1 to 19 while keeping the total number for both empirical

factors and hidden factors as 20, and again on each p level the simulation runs for 100 times. Each time the mean value of estimated covariance is calculated and deducted from mean value of the true covariance, and we record the statistics $mean(vec(\overline{\Sigma} - \Sigma)$. We use this statistics to show the estimation variance and bias by boxplotting on the 100 samples for each p. In this way it gives an idea about the trade-off for the estimated covariance. We repeat this experiment for four times under different the number of samples n and the imperfect ratio α, corresponding to four scenarios as (1) large sample size, accurate empirical factors($n = 400, \alpha = 0$); (2) small sample size, accurate empirical factors ($n = 40, \alpha = 0$); (3) large sample size, noisy empirical factors($n = 40, \alpha = 0.3$). (3) small sample size, noisy empirical factors($n = 40, \alpha = 0.3$).

For case (1), as shown in Figure 36., when there is a perfect situation that enough samples are observable and the empirical information is perfectly accurate. As shown in Figure 3., the bias and variance for both the EH model and the statistical factor model are similar, while for the hybrid model, the variance and bias is higher compared with the previous two models.

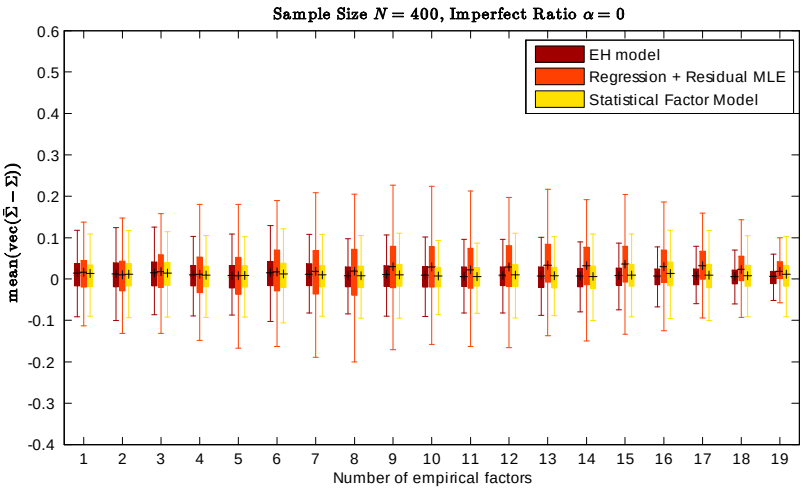


Figure 36.

For case (2), comparing Figure 37. and Figure 36., it can be seen that when there is not enough sample size, the variance of the hybrid method is 10% to 20% smaller compared with the other two models, which is a result of less parameters, but its bias almost double the other two models. This biasedness comes from the first step regression, since the underlying

assumption for this hybrid method is that the residual is perpendicular to the column space of fundamental factors, but in reality the residual after regression contains V_h, which is not necessarily perpendicular to empirical factor exposure V_e. On the other hand, both statistical factor model and the EH model are unbiased. Further, comparing the EH factor model and the statistical factor model, the EH factor model has a both smaller variance and smaller bias due to the reducing of parameters.

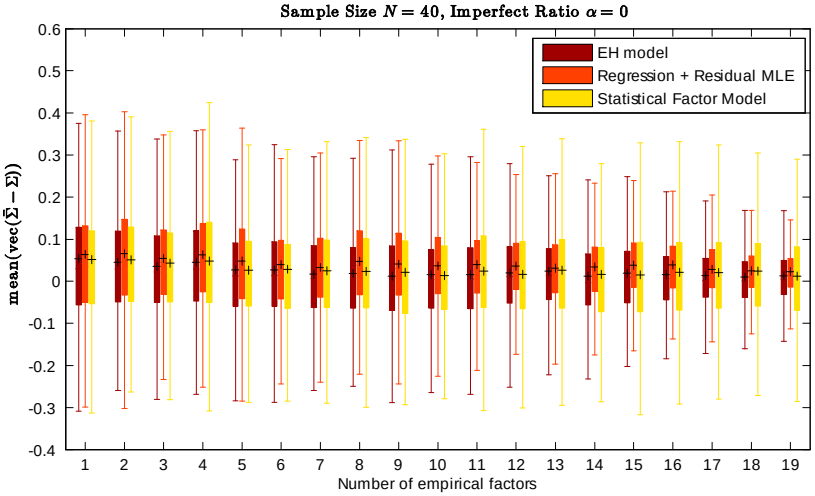


Figure 37.

For case (3), comparing Figure 38. and Figure 36., when adding noise to the empirical factors, the variance of the hybrid method has increased from 0.0694 to 0.0924, and the bias has increased from 0.0005 to 0.0011. The bias for the EH model stays at a similar level while its variance slightly increases from 0.0448 to 0.0642. This shows that the EH model is more robust to the inaccuracy in the empirical information.

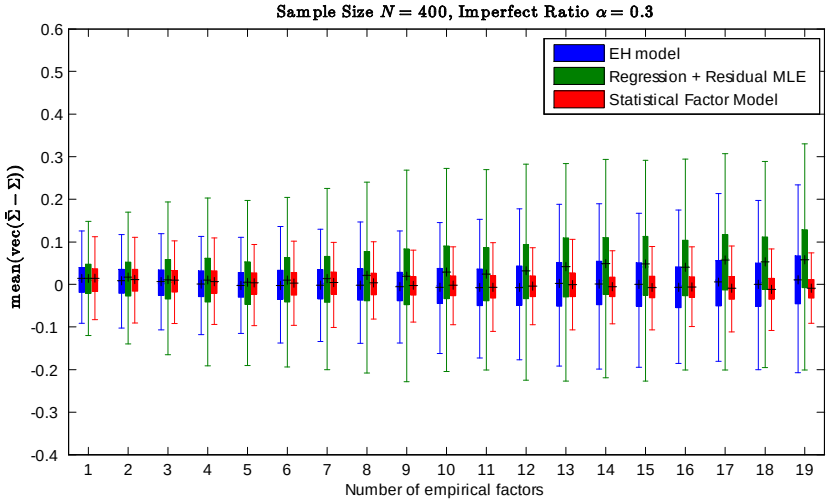


Figure 38.

For case (4), which is the worst situation that both the sample size is small and the empirical factor is noisy, the variance for three models are similar, while the bias for the EH model is significantly smaller(0.0002), compared with the hybrid method(0.0018) and the statistical factor model(0.0013).

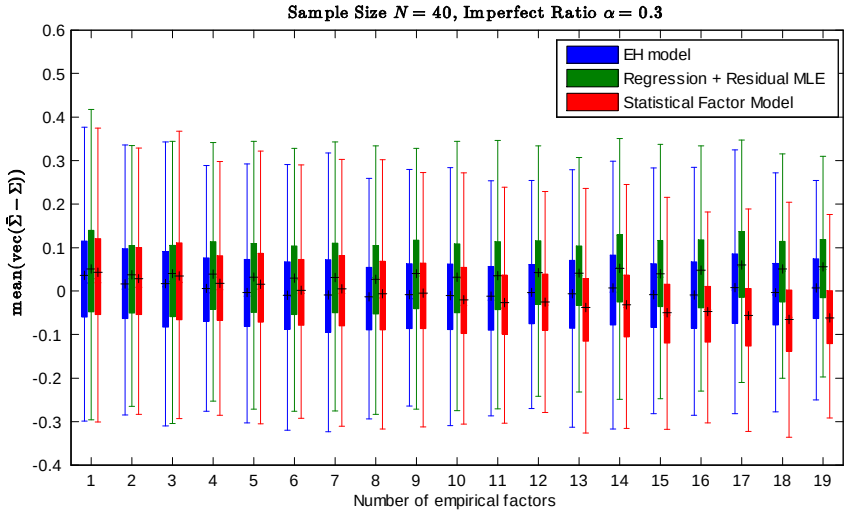


Figure 39.

	N=400, α=0		N=40, α=0		N=400, α=0.3		N=40, α=0.3	
	Var	Bias	Var	Bias	Var	Bias	Var	Bias
EH Factor Model	0.0448	0.0001	0.1219	0.0007	0.0642	0.0000	0.1251	0.0002
2MLE	0.0694	0.0005	0.1106	0.0018	0.0924	0.0011	0.1255	0.0018
Statistical Factor	0.0473	0.0001	0.1422	0.0009	0.0430	0.0001	0.1275	0.0013

Table 6. The variance and bias of three different models.

4 Conclusion

In this thesis we study the problem of modeling cross-sectional volatility structure of U.S. stock market. We identify a particular volatility spike phenomenon, that the cross-sectional volatility is significantly stronger at every 5 minute. An empirical spike study is conducted on individual stock level by applying Lee-Mykland jump detector. By constructing spike ratio statistics, the evolving paths of this spike phenomenon is studied during the 1993 to 2009 time period. In the second part, we model the volatility structure using factor structures. Particular, we propose two approaches to build a better model for small sample size problem, which is a common issue for high dimension models. We first study the fisher information matrix and derive Jeffrey's prior for the factor model. We extend the EM algorithm method for factor parameter estimation by applying the Jeffrey's prior, which turns to be more robust in the sense that risk would not be underestimated under small sample size. Next we extend the statistical factor model to be able to process empirical information. A new factor model called EH model (Empirical-Factor-and-Hidden-Factor model) is therefore introduced, which uses both fundamental information and statistical factor framework through structuring the exposure matrix (factor loading matrix) into a combination of an empirical information block and a hidden information block. In the simulation experiment, this approach shows a better estimation performance in the sense of less bias and smaller variance compared with statistical factor model and the hybrid method. It has the most advantages over the other two models when (1). the sample size is small compared with the covariance dimension; or (2). the empirical information is relatively accurate.

As a conclusion, we point out that the trade-off between the low dimension approach and high dimension approach is a variance-bias tradeoff. Particular, it is most helpful to use a simple model to detect signals, that under the help of large samples and few parameters the model can have a low estimatinon variance. When trying to estimation the volatility structure in a finer level, high dimension parametrization is preferred, but for which cases the variance on estimating each parameter maybe challenging, that special linear algebra structures would help under this circumstance.

References

[1] Abhyankar, A., D. Ghosh, et al. (1997). "Bid ask Spreads, Trading Volume and Volatility: Intra day Evidence from the London Stock Exchange." *Journal of Business Finance & Accounting* 24(3): 343-362.

[2] Andersen, T. and T. Bollerslev (1997). "Intraday periodicity and volatility persistence in financial markets." *Journal of empirical finance* 4(2-3): 115-158.

[3] Andersen, T., T. Bollerslev, et al. (2000). "Intraday and interday volatility in the Japanese stock market." *Journal of International Financial Markets, Institutions and Money* 10(2): 107-130.

[4] Briner,B, and Connor,G. "How much structure is best? A comparison of market model, factor model and unstructured equity covariance matrices." *Journal of Risk*. 10.4 3-30(2008).

[5] Cai, C., R. Hudson, et al. (2004). "Intra Day Bid Ask Spreads, Trading Volume and Volatility: Recent Empirical Evidence from the London Stock Exchange." *Journal of Business Finance & Accounting* 31(5 6): 647-676.

[6] Chan,L.K.,Karceski,J, and Lakonishok.J, "On portfolio optimization: forecasting covariances and choosing the risk model." *The Review of Financial Studies*, 12, 937-974(1999).

[7] Chen.N, Roll.R, and Ross.S. "Economic forces and the stock market." *Journal of Business*, 59,383-403(1986).

[8] Chan, K. and W. Fong (2000). "Trade size, order imbalance, and the volatility-volume relation* 1." *Journal of Financial Economics* 57(2): 247-273.

[9] Cheung, Y. (1995). "Intraday Returns and the day-end effect: Evidence from the Hong Kong equity market." *Journal of Business Finance & Accounting* 22(7): 1023-1034.

[10] Connor,G, "Three types of factor models: A comparison of their explanatory power." *Financial Analysts Journal*, vol 51: pp42-46, May/June(1995).

[11] Chilson, J., R. Ng, et al. (2006). "Parallel computation of high-dimensional robust correlation and covariance matrices." *Algorithmica* 45(3): 403-431.

[12] Darrat, A., S. Rahman, et al. (2003). "Intraday trading volume and return volatility of the DJIA stocks: A note." *Journal of Banking & Finance* 27(10): 2035-2043.

[13] Dempster, A., N. Laird, et al. (1977). "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society.* Series B (Methodological) 39(1): 1-38.

[14] Ding, D. and S. Lau (2001). "An analysis of transactions data for the stock exchange of Singapore: patterns, absolute price change, trade size and number of transactions." *Journal of Business Finance & Accounting* 28(1 2): 151-174.

[15] Duhigg, C. (2009). "Stock traders find speed pays, in milliseconds." *The New York Times*, 2009, 23.

[16] Eaton,M.L, "Multivariate statistics: a vector space approach." *John Wiley and Sons.* pp. 116–117. ISBN 0-471-02776-6(1983).

[17] Ellul, A. (2001). "Trading behaviour, price discovery and volatility in competing market microstructure."

[18] Ellul, A., H. Shin, et al. (2002). "Toward deep and liquid markets: Lessons from the open and close at the London stock exchange." Working Paper (London School of Economics, Financial Markets Group).

[19] Foster, F. and S. Viswanathan (1990). "A theory of the interday variations in volume, variance, and trading costs in securities markets." *Review of Financial Studies* 3(4): 593.

[20] Fama,E.F, and French,K.R, "Common risk factors in the returns on stocks and bonds." *Journal of Financial Economics*, 33, 3C56(1993).

[21] Foster, F. and S. Viswanathan (1990). "A theory of the interday variations in volume, variance, and trading costs in securities markets." *Review of Financial Studies* 3(4): 593.

[22] French, K. (1980). "Stock returns and the weekend effect* 1." *Journal of Financial Economics* 8(1): 55-69.

[23] Garvey, R. and F. Wu (2009). "Intraday time and order execution quality dimensions." *Journal of Financial Markets* 12(2): 203-228.

[24] Gencay, R., et al. (2001). An introduction to high-frequency finance, *Access Online via Elsevier*, 2001.

[25] Gencay, R., F. Selçuk, et al. (2001). "Scaling properties of foreign exchange volatility." *Physica A: Statistical Mechanics and its Applications* 289(1-2): 249-266.

[26] Granger, C. and C. Sin (2000). "Modelling the absolute returns of different stock indices: exploring the forecastability of an alternative measure of risk." *Journal of Forecasting* 19(4): 277-298.

[27] Gerety, M. and J. Mulherin (1992). "Trading halts and market activity: an analysis of volume at the open and the close." *Journal of Finance* 47(5): 1765-1784.

[28] Gibbons, M. and P. Hess (1981). "Day of the week effects and asset returns." *Journal of Business* 54(4): 579-596.

[29] Granger, C. and Z. Ding (1995). "Some properties of absolute return: An alternative measure of risk." *Annales d'Economie et de Statistique*: 67-91.

[30] Hamao, Y. and J. Hasbrouck (1995). "Securities trading in the absence of dealers: Trades and quotes on the Tokyo Stock Exchange." *Review of Financial Studies* 8(3): 849.

[31] Ho, R. and Y. Cheung (1994). "Seasonal pattern in volatility in Asian stock markets." *Applied Financial Economics* 4(1): 61-67.

[32] Ito, T. and Y. Hashimoto (2006). "Intraday seasonality in activities of the foreign exchange markets: Evidence from the electronic broking system." *Journal of the Japanese and International Economies* 20(4): 637-664.

[33] Jagannathan, R, and Ma, T, "Risk reduction in large portfolios: why imposing the wrong constraints helps." *Journal of Finance*, 58,1651-1684(2003).

[34] Jain, P. and G. Joh (1988). "The dependence between hourly prices and trading volume." *Journal of Financial and Quantitative Analysis* 23(03): 269-283.

[35] Kalev, P. and L. Pham (2009). "Intraweek and intraday trade patterns and dynamics." *Pacific-Basin Finance Journal* 17(5): 547-564.

[36] Keasey, K. and H. Short (1999). "Corporate governance: From accountability to enterprise." *Accounting and business research* 29(4): 337-352.

[37] Keim, D. and R. Stambaugh (1984). "A further investigation of the weekend effect in stock returns." *Journal of Finance* 39(3): 819-835.

[38] Kleidon, A. and I. Werner (1993). Round-the-clock trading: Evidence from UK cross-listed securities, NBER.

[39] Laloux, L. et al. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 3(03), 391-397 (2000).

[40] Lee, S. and P. Mykland (2008). "Jumps in financial markets: A new nonparametric test and jump dynamics." *Review of Financial studies* 21(6): 2535.

[41] Lin, C. (2006). "Trading Behavior in Taiwan FX Market: Trading Pattern and the Impact of Macroeconomic News." *National Cheng Kung University Institute of International Business*, 2006

[42] Maronna, R. (1976). "Robust M-estimators of multivariate location and scatter." *The annals of statistics*: 51-67.

[43] Menchero, J., Morozov, A., and Shepard, P., "The Barra global equity model (GEM2)", Research notes, *MSCI Barra*(2008).

[44] Müller, U., M. Dacorogna, et al. (1990). "Statistical study of foreign exchange rates, empirical evidence of a price change scaling law, and intraday analysis* 1." *Journal of Banking & Finance* 14(6): 1189-1208.

[45] Rosenberg, B.A, "Extra-market components of covariance in security returns", *Journal of Financial and Quantitative Analysis*, vol.9, no.2 (March 1974):263-73(1974).

[46] Rubin, D. and D. Thayer (1982). "EM algorithms for ML factor analysis." *Psychometrika* 47(1): 69-76.

[47] Taylor, S. (2000). "Stock index and price dynamics in the UK and the US: new evidence from a trading rule and statistical analysis." *The European Journal of Finance* 6(1): 39-69.

[48] Taylor, S. (2008). Modelling financial time series, World Scientific Pub Co Inc.

[49] McInish, T. and R. Wood (1990). "An analysis of transactions data for the Toronto Stock Exchange:: Return patterns and end-of-the-day effect." *Journal of Banking & Finance* 14(2-3): 441-458.

[50] Wang, E., R. Hudson, et al. (2000). "Tick size and the compass rose: further insights." *Economics Letters* 68(2): 119-125.

[51] Wilcox, R. (2005). Introduction to robust estimation and hypothesis testing, Academic Press.

[52] Wood, R., T. McInish, et al. (1985). "An investigation of transactions data for NYSE stocks." *Journal of Finance* 40(3): 723-739.

[53] Woodbury, M.A, "Inverting modified matrices." *Memorandum report* 42: 106(1950).

[54] Yadav, P. and P. Pope (1992). "Intraweek and intraday seasonalities in stock market risk premia: Cash and futures." *Journal of Banking & Finance* 16(1): 233-270.

[55] Zhao, J.-H., Philip, L., & Jiang, Q. (2008). ML estimation for factor analysis: EM or non-EM? *Statistics and Computing*, 18(2), 109-123.

A Appendix A: Framework of EM Algorithm

EM algorithm is a method that find the maximum likelihood estimation in an iterative way. Usually EM algorithm is used to estimate the models with hidden variables. Also, if an explicit model can be formulated more easily with introducing additional data, EM algorithm can be used. For example, if the original model is to estimate the parameter Ψ which maximizes

$$\log L(\Psi; y) = \log p(y; \Psi)$$

, then if assuming observing additional data x, the model has a more explicit structure or an easier estimation, that

$$\begin{aligned} p(y;\Psi) &= \sum_x p(y|x;\Psi)p(x;\Psi) \\ &= \sum_x L(y|x;\Psi)p\left(x;\Psi\right) \\ &= E_{x|y}\left[L(y|x;\Psi)\right] \end{aligned}$$

that is to express probability $p(y;\Psi)$ as as the marginal probability function of $p(y,x;\Psi)$

The way EM algorithm works is to iteratively increase $E_{x|y}[\log L(y|x;\Psi)]$ by defining the Q function $Q(\Psi^{k+1};\Psi^k)$ as

$$Q(\Psi^{k+1};\Psi^k) = E_{x|y,\psi^k}\left[\log L(\Psi^{k+1},y|x)\right] \tag{59}$$

And the updating formula for find Ψ^{k+1} is by solving the zero equation of 59.

B Appendix B: Convergence of EM algorithm

Dempster, Laird and Rubin (1977) showed the function $L(\Psi;y)$ is non-decreasing.

Theorem 1. ***In each EM iteration step, the log likelihood is nondecreasing, that***

$$L(\Psi^{k+1};y) \geqslant L(\Psi^k;y)$$

proof:

Since

$$\log L(\Psi^{k+1};y) = \log p(y;\Psi^{k+1})$$

and

$$p(x|y;\Psi)p(y;\Psi) = p(x,y;\Psi)$$

, therefore

$$\begin{aligned} \log L(\Psi^{k+1};y) &= \log p(y;\Psi^{k+1}) \\ &= \log\left\{p(y,x;\Psi^{k+1})/p(x|y;\Psi^{k+1})\right\} \end{aligned}$$

Define the complete data as $z = [x,y]^T$.

$$\begin{aligned}\log L(\Psi^{k+1};y) &= \log\left\{p(z;\Psi^{k+1})/p(x|y;\Psi^{k+1})\right\}\\ &= \log L(\Psi^{k+1};z)-\log p(x|y;\Psi^{k+1})\end{aligned}$$

Take expectation with respect to x and using Ψ^k as the parameter, that is to use $p(x;\Psi^k)$ when taking the integration,

$$\begin{aligned}\log L(\Psi^{k+1};y) &= E_{x|y,\Psi^k}[\log L(\Psi^{k+1};y)]-E_{x|y,\Psi^k}\left[\log p(x|y;\Psi^{k+1})\right]\\ &= Q(\Psi^{k+1};\Psi^k)-E_{x|y,\Psi^k}\left[\log p(x|y;\Psi^{k+1})\right]\end{aligned}$$

Similarly,

$$\log L(\Psi^k;y)=Q(\Psi^k;\Psi^k)-E_{x|y,\Psi^k}\left[\log p(x|y;\Psi^k)\right]$$

and therefore

$$\begin{aligned}\log L(\Psi^{k+1};y)-\log L(\Psi^k;y) &= \left(Q(\Psi^{k+1};\Psi^k)-Q(\Psi^k;\Psi^k)\right)\\ &\quad+\left(E_{x|y,\Psi^k}\left[\log p(x|y;\Psi^{k+1})\right]-E_{x|y,\Psi^k}\left[\log p(x|y;\Psi^k)\right]\right)\end{aligned}$$

First notice that

$$Q(\Psi^{k+1};\Psi^k)-Q(\Psi^k;\Psi^k)\geqslant 0$$

since Ψ^{k+1} is found by maximize $Q(\Psi;\Psi^k)$.

Second

$$\begin{aligned}E_{x|y,\Psi^k}\left[\log p(x|y;\Psi^k)\right]-E_{x|y,\Psi^k}\log p(x|y;\Psi^{k+1}) &= E_{x|y,\Psi^k}\left[\log\left\{p(x|y;\Psi^{k+1})/p(x|y;\Psi^k)\right\}\right]\\ &\geqslant \log E_{x|y,\Psi^k}\left[p(x|y;\Psi^{k+1})/p(x|y;\Psi^k)\right]\\ &= \log\int_x\left\{p(x|y;\Psi^{k+1})/p(x|y;\Psi^k)\right\}p(x|y;\Psi^k)dx\\ &= \log\int_x p(x|y;\Psi^{k+1})dx\\ &= \log 1\\ &= 0\end{aligned}$$

Proof complete.

C Appendix C: EM Algorithm for Exponential Family

For exponential family, EM algorithm has a simpler expression.

Theorem 2. *In case y follows exponential family, that $p(y;\Psi) = h(y)g(\Psi)\exp\left\{\Psi^T T(y)\right\}$, the updating equation for Ψ^{k+1} satisfies that*

$$E_{x|y,\psi^k}\left[T(y)\right] = E_{\Psi^{k+1}}[T(y)]$$

proof:

$$Q(\Psi^{k+1};\Psi^k) = E_{x|y,\psi^k}\left[\log L(\Psi^{k+1};y|x)\right]$$

if $y|x$ follows exponential family, that

$$\begin{aligned} p(y|x,\Psi) &= h(y)g(\Psi)\exp\left\{\Psi^T T(y)\right\} \\ \log p(y|x,\Psi) &= \log h(y) + \log g(\Psi) + \Psi^T T(y) \\ E_{x|y,\psi^k}\left[\log L(\Psi^{k+1},y|x)\right] &= \log h(y) + \log g(\Psi^{k+1}) + (\Psi^{k+1})^T E_{x|y,\psi^k}\left[T(y)\right] \end{aligned}$$

Notice by the property of exponential family,

$$\frac{\partial \log g(\Psi)}{\partial \Psi} = -E[T(y)]$$

Therefore, take differentiation on both side with respect to Ψ^{k+1}, and set it to zero

$$\begin{aligned} 0 &= -E_{\Psi^{k+1}}[T(y)] + E_{x|y,\psi^k}\left[T(y)\right] \\ E_{x|y,\psi^k}\left[T(y)\right] &= E_{\Psi^{k+1}}[T(y)] \end{aligned}$$

where Ψ^{k+1} is in $T(y)$ function.